Quest Journals
Journal of Software Engineering and Simulation
Volume 11 ~ Issue 1 (2025) pp: 30-40
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

An Intelligent Cost Review Scheme for Power Engineering Projects Based on Natural Language Processing

Yin Lv
North China Electric Power University

ABSTRACT: *Traditional manual cost review in power engineering projects faces inefficiencies and inaccuracies due to the complexity of data and rules. This study aims to introduce an intelligent cost review scheme that automates data extraction and rule interpretation using natural language processing techniques and completes automatic review. The proposed scheme consists of three modules: information extraction, rule interpretation, and intelligent review, utilizing techniques such as named entity recognition and relation extraction. Experimental results demonstrate the scheme's effectiveness in accurately review power engineering cost documents, reducing manual workload, and improving efficiency. This research offers an effective solution applicable to power engineering projects, advancing intelligent cost review in project management by automating information extraction and rule interpretation, and achieving the goal of automatic review.*

KEYWORDS: *Cost review, Intelligent scheme, Natural language processing, Automatic interpretation, Information extraction , Intelligent review*

R*eceived 25 Jan., 2024;* ***R****evised 31 Jan., 2025;* ***A****ccepted 13 Feb., 2025 © The author(s) 2025.*
P*ublished with open access at* ***www.questjournas.org***

I. INTRODUCTION

Power engineering cost management is a vital part of power engineering projects. It is applied to the proofreading process of project costs to ensure that cost control and management achieve expected goals. Proofreading is a process of comprehensive review of all aspects of power engineering projects, aiming to ensure the reliability and compliance of the project, and to ensure smooth progress and high-quality completion of power engineering project.

Traditional cost review in power industry mainly relies on manual proofreading. Review experts need to manually extract data, sharing information by electronic documents, and some data also need to be entered manually for calculation. However, this traditional approach has obvious limitations. First, the review efficiency is too low. Due to the large number of power engineering projects, it is inefficient to rely solely on review experts for manual cost review. Within limited time, they can only check the accuracy and validity of the basic data for different projects, and cannot provide timely feedback on changes in the review data. Secondly, the accuracy of the results cannot be guaranteed. Due to the complicated entries of proofreading data and rules, manual reviews cannot ensure the accuracy of proofreading results, which may lead to losses of projects. In addition, manual method brings high labor and time cost. The existing review program requires reviewers to have a certain understanding of reviewing rules, which results in high initial training costs for reviewers and adds additional labor and time costs.

In order to reduce the burden of manual review and improve efficiency and accuracy, traditional cost analysis software uses automated techniques to help managers better obtain, integrate and analyze information, and achieve effective control and optimization of project reviews. But traditional cost analysis software also shows its shortcomings. First of all, these software have difficulties in processing complex unstructured data . Especially for large-scale textual engineering project data, they are often difficult to accurately and efficiently extract important cost information, which affects the accuracy of their review. Secondly, although automation technology is adopted, its degree of automation is still low and it cannot completely replace manual review. It still requires some manual participation, which limits its potential to improve efficiency and accuracy. Finally, they are unable to adapt to the needs of various engineering projects, which limits its application in diverse power engineering projects.

To solve these problems, we aim to introduce an intelligent cost review scheme in power engineering, which can realize automated extraction and processing of textual data and rules, reducing the workload for

reviewers. At the same time, we aim to transform complex proofreading data and rules to computer-understandable symbolic language, and then perform automatic cost review, improving the accuracy of proofreading results.

In this paper, we propose a novel scheme to intelligently review textual power engineering cost data using natural language processing technology. Specifically, our scheme mainly includes three key modules: information extraction, rule interpretation and intelligent proofreading. Firstly, in the information extraction module, we adopt Named Entity Recognition (NER) technique to extract key cost information from the original unstructured text of the power project. Secondly, in the rule interpretation module, we utilize NER and Relationship Extraction (RE) technologies to convert the unstructured text of the review rules into computer-understandable symbolic language. Finally, in the intelligent proofreading module, we perform the intelligent proofreading task and give proofreading feedback. We perform experiments on two engineering review data set to verify the effectiveness of the proposed scheme. Our experimental results show that the proposed scheme performs well in both NER and RE tasks. Our case study show that it can conduct intelligent proofreading according to the generated symbolic language of proofreading rules.

Our contribution can be summarized as follows:

We proposed a set of universal solutions that are not only applicable to the field of power engineering, but can also be applied to the intelligent review of cost in other engineering projects, providing a feasible solution for the intelligence of engineering project management.

Our solution achieves automatic and intelligent extraction of information, can effectively handle complex data structures confirming its feasibility and effectiveness in practical applications through experiments, making important contributions to the technological development in the field of engineering project management.

We transformed the automatic interpretation process of proofreading rules into an information extraction task. By combining technologies such as NER and RE, we achieved efficient automatic interpretation of proofreading rules, providing reliable technical support for subsequent intelligent proofreading.

The following parts of this paper are organized as follows: Section 2 introduces related work, Section 3 explains our research methods, Section 4 presents the experimental results, Section 5 discusses the results and concludes the paper and proposes prospects for future work.

II. RELATED WORK

2.1 ENTITY RELATIONSHIP EXTRACTION

With the advent of the big data era, establishing a model that can quickly and efficiently extract effective information from a large amount of open domain and unstructured data has become an important issue in the current field of Natural Language Processing (NLP). As the core task of information extraction [1], entity relationship extraction aims to quickly and efficiently extract the entity pairs and their semantic relationships contained in text sentences by modeling text sentences, and then obtain the structured triple in the sentences, formatted as <entity 1, relationship, entity 2>. The acquired triple information is used in downstream natural language processing tasks such as large-scale knowledge graph construction [2], machine reading, text summarization, question and answer system [3], machine translation [4], and semantic web annotation. In recent years, with the rise of information extraction related research and the rapid development of deep learning, research on entity relationship extraction has continued to deepen, producing a large number of excellent research results.

Early entity relationship extraction was seen as two subtasks, Named Entity Recognition (NER) [5,6] and Relationship Extraction (RE) [7,8]. For these two tasks, researchers initially studied entity relationship extraction using pipeline methods. Firstly, They begin constructing entity recognition models [9,10] using artificial feature extraction and kernel function and then built models that could recognize their semantic relationships based on entity pairs [11-13] to achieve entity relationship extraction. With the rapid development of deep learning technology in recent years, some end-to-end deep learning models have gradually emerged. Occupying a dominant position, deep learning based NER related research has achieved fruitful results [14-20], and in the field of RE, deep learning models have also brought excellent outcomes [21-29], demonstrating their effectiveness on several publicly available benchmark datasets.

2.2 RULE INTERPRETATION

In recent decades, automated rule checking (ARC) methods and systems based on natural language processing (NLP) have been extensively studied [30]. However, in existing ARC systems, the rule interpretation phase still requires a lot of manual work [31], so in order to improve the efficiency and transparency of the rule interpretation phase, many automated rule interpretation methods have been proposed.

To enable a comprehensive understanding of text rules, natural language processing algorithms [32] have been used to develop automatic rule interpretation methods. NLP algorithms can be mainly divided into

two methods: handwritten rule methods and statistical methods [33]. The handwritten rule methods relies on human-defined symbolic pattern matching rules, including Backus-Naur Normal Form (BNF) notation, regular expression grammar, and the Prolog language. Statistical methods, including Support Vector Machines (SVM), Hidden Markov Models (HMM), Conditional Random Field (CRF) and Naive Bayes [33], are implemented by automatically constructing probabilistic list "rules" from training data sets (such as large annotated text bodies), similar to machine learning algorithms.

In the AEC industry, Zhang & ElGohary [34] pointed out that compared with general non-technical texts (such as news articles, general websites), regulatory texts in specific fields are more suitable for automated NLP because regulatory texts have fewer homophone conflicts and coreference resolution problems. Furthermore, domain-specific ontologies are easier to develop than ontologies that capture general knowledge across multiple domains. Domain ontology can provide a shared vocabulary by defining abstract concepts and relationships, including classification of concepts, equivalent and disjoint concepts, and enumeration of terms. Therefore, the raw data has a clear meaning, making it easier for machines to automatically process and integrate the data through the ontology [35]. Domain ontology can enhance the automatic interpretability and understandability of domain-specific texts [34]. Therefore, in the AEC industry, many studies on automatic rule interpretation methods based on NLP have adopted rule-based and ontology-based methods [32].

Zhang & El-Gohary [34,36] proposed an automatic rule interpretation method that consists of three stages: Text classification is used to identify relevant sentences in regulatory texts; Information extraction extracts words and phrases from relevant sentences; and Information transformation will extract the information and converted into Horn clause or B-Prolog representation. Zhou & El-Gohary [37] proposed a rule-based ontology-enhanced information extraction method to extract building energy requirements from energy-saving specifications and format them into B-Prolog representation. Zhou et al. [38] proposed a fully automatic rule extraction method based on a deep learning model and a set of context-free grammars (CFG) to automatically interpret regulatory rules into pseudo-decode format with high versatility, accuracy and interpretability. Xu and Cai [39] proposed an NLP framework based on ontology and rules to automatically interpret utility regulations into deontic logic (DL) clauses to achieve semantic alignment between rules and ontologies. Zheng et al. [40] proposed a knowledge-informed framework based on natural language processing to improve ARC, which enhanced the rule interpretation process by introducing semantic alignment and conflict resolution. Finally, an algorithm was developed to identify the correct SPARQL function for each rule and generate SPARQL-based Queries for interpreting complex rules that require extra implicit data to be inferred.

III. METHOD

This study proposes an intelligent cost review scheme for power project based on natural language processing, aiming to automate the proofreading process through the application of natural language processing technology. As shown in Figure (1), the ontology concepts such as the terms involved in the power original text and the verification rules are first defined. Subsequently, the original unstructured text of the power project is input into the information extraction module, and the name and attributes of the power equipment are extracted from the original unstructured text. Then input the rule text into the proofreading rule interpretation module to obtain the fact triplet transformed by the rules. Thereby a symbolic language for computer executable processing can be obtained. Finally, use the generated proofreading data to proofread the key cost information. The functions and implementation methods of each module will be introduced in detail below.

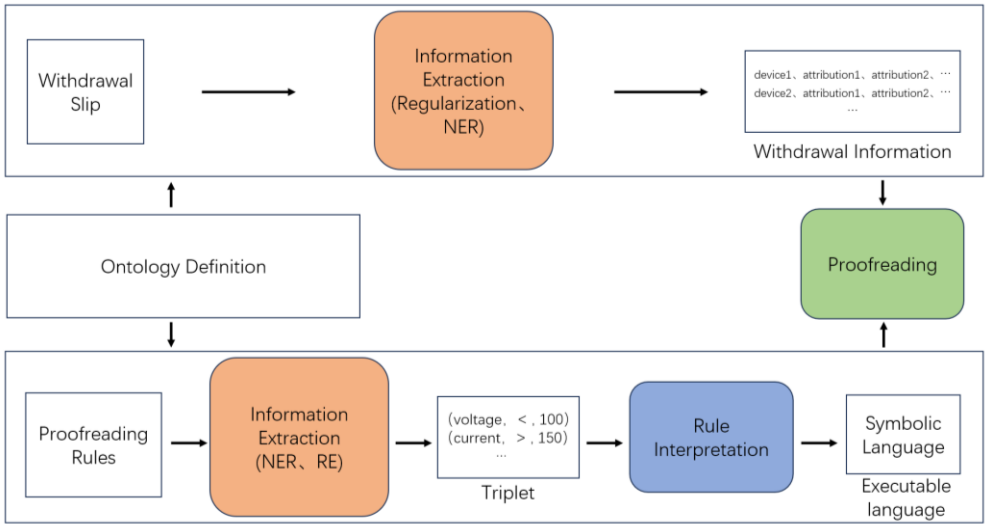

Figure (1): Framework of intelligent cost verification scheme

3.1 ONTOLOGY DEFINITION

In the scheme, we first defined the ontology of related concepts such as power equipment and power attributes, and established a standardized ontology concept system. These ontology definitions play the role of unifying and standardizing the power nouns and terms involved in text language. 8 different ontology definitions were designed after consideration as shown in Table 1, to ensure consistency and accuracy during the review process. These ontology definitions provide an important semantic basis for the subsequent review process.

Table 1: ontology definitions

Device	Device--such as generators, transformers, switchgear, etc.--refers to components and devices related to electricity, used for generating, transmitting, distributing, or consuming electrical energy.
Voltage	Voltage is an important concept in the power system, representing the driving force of charge flow in the power system. It is measured in Volts (V).
Current	Current is a measure of the flow of charges in the power system, indicating the amount of charge passing through a point per unit time. The current is measured in Amperes (A).
Capacity	Capacity usually refers to the capacitance of a capacitor, indicating the amount of charge it can store. The capacitance is measured in Faradas (F).
Type	Type refers to the type or specification of device, and describes the specific design, function, and working principle of the device. For example, transformers have different types, such as oil immersed transformers, dry-type transformers, etc.
Phase	Phase is used to describe the mutual relationship between the power source and load in the power system. It indicates the electrical connection method between the power supply and load used in the system, commonly including single-phase and three-phase systems.
Windings	Windings refers to the number of coils or coil groups in power device. In device such as transformers or motors, the number of windings can affect the electrical characteristics and operational performance of the equipment.
Value	Value usually refers to specific numerical values related to electrical parameters or attributes. For example, the voltage value can be 220V, the current value can be 10A, the capacity value can be 10F, etc.

3.2 EXTRACTION OF ORIGINAL UNSTRUCTURED TEXT INFORMATION

The original unstructured text information extraction module uses regular expressions and NER models to extract the name and attribute information of the power equipment from the original text. This is the first core module of the scheme, whose purpose is to convert the original text information expressed in natural language into structured data that can be understood and processed by computers.

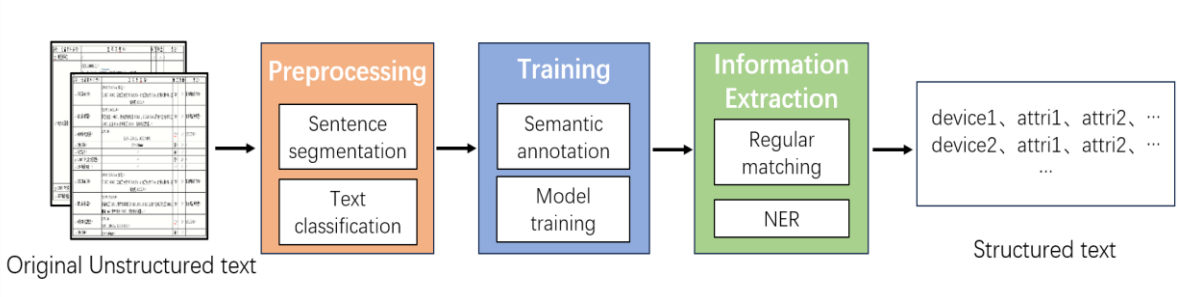

Figure (2): Intelligent cost review original unstructured text information extraction

As shown in Figure (2), we first segment the original document text into a text form described in a single sentence. Next, we trained a named entity recognition model. By semantically annotating original text, the model was able to automatically extract the attribute information of the corresponding device in the original text. The manual annotation method in BIO (Begin-inside-outside) format is used here to assign semantic labels to words and phrases in sentences. The structure diagram of the NER model is shown in Figure (3). The model adopts the structure of BERT+BiLSTM+CRF. Specifically, the model's inputs include tokenized text sequences $X_{\text{token}} = (x_{\text{token},1}, x_{\text{token},2}, \ldots, x_{\text{token},N})$, token segment embeddings $X_{\text{seg}} = (x_{\text{seg},1}, x_{\text{seg},2}, \ldots, x_{\text{seg},N})$, and position embeddings $X_{\text{pos}} = (x_{\text{pos},1}, x_{\text{pos},2}, \ldots, x_{\text{pos},N})$. These inputs are processed through the BERT module to obtain contextualized representations $H_{\text{BERT}} = (h_{\text{BERT},1}, h_{\text{BERT},2}, \ldots, h_{\text{BERT},N})$ for each token. Subsequently, the BiLSTM

module takes the contextualized representations from BERT as input, encoding the token sequence bidirectionally to generate BiLSTM output sequences $H_{\text{BiLSTM}} = (h_{\text{BiLSTM},1}, h_{\text{BiLSTM},2}, ..., h_{\text{BiLSTM},N})$. Finally, the CRF module receives the BiLSTM output sequences and utilizes conditional random fields to predict the optimal label sequence $Y = (y_1, y_2, ..., y_N)$, with the training loss function denoted as:

$$L = L_{CRF}(X, Y) + \lambda \cdot || \theta ||$$

Here, $L_{CRF}(X, Y)$ is the conditional random field loss function, representing the difference between the predicted label sequence and the true label sequence; λ is the coefficient of the regularization term, used to control the complexity of the model; $|| \theta ||_2^2$ denotes the square of the L2 norm of the model parameters.

Finally, we combine the pre-written regular expression rules to match device names and attributes, and combine the information extracted by the NER model to obtain structured data. This way, computers can process the data directly.

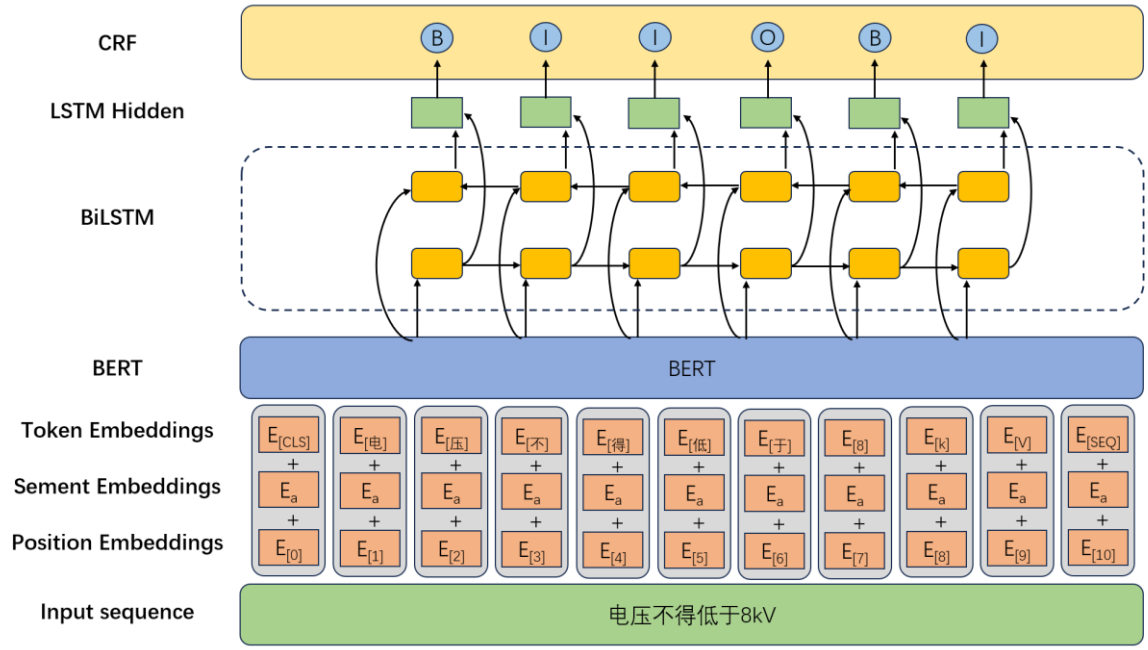

Figure (3): Structure of NER model

3.3 AUTOMATIC INTERPRETATION OF RULES

The automatic interpretation module of proofreading rules uses the Named Entity Recognition (NER) model and the Relationship Extraction (RE) model to convert proofreading rules expressed in natural language into fact triplet, and encodes these triples into computer-readable symbolic language that performs processing. This is the second core module of the scheme. Its goal is to convert the review rules expressed in natural language by domain experts into a computer-understandable form so that the computer can accurately understand and execute the rules.

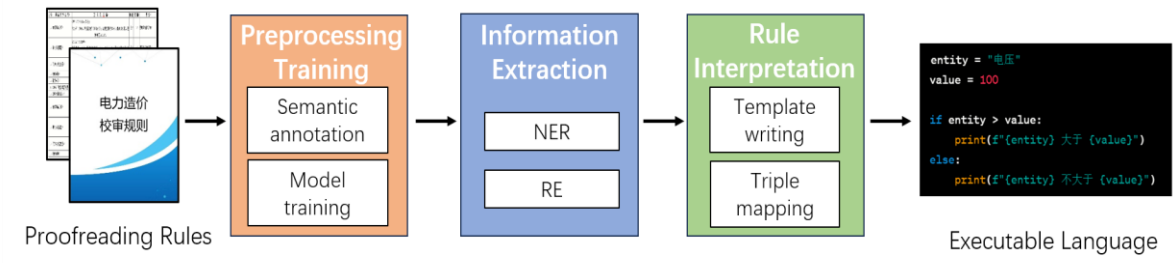

Figure (4): Intelligent cost review rule interpretation

As shown in Figure (4): First, we need to pre-train a NER model and a RE model. By semantically annotating the rule text data, the model can automatically extract the attribute information corresponding to the device in the rule text. During operation, this module first identifies various entities in the review rules

expressed in natural language. Then, the RE model is used to extract the relationships between these entities. These relationships are defined into five situations, namely greater than, less than, not greater than, not less than, and equal to.

Finally, the review rules expressed in natural language are converted into the form of fact triples. After completing the NER and RE, enter the structural data generation step and the relationship in the form of triples is converted into a symbolic language that can be processed by the computer to form conditions set and conclusions set. The process is shown in Figure 4. In this way, the proofreading rules are encoded into a computer-executable format, so that the computer can understand and proofread according to the rules.

3.4 INTELLIGENT PROOFREADING

In this scheme, the structured data of the original text and the executable symbolic language of the proofreading rules are combined to realize the intelligent proofreading function, which is used to judge the compliance of the cost information.

During the intelligent review process, the scheme performs automated verification and comparison based on the original text information and review rules. The scheme will check whether the device name and attributes in the original text comply with the specifications, and whether they meet the conditions constraints in the review rules. Based on the proofreading results, the scheme generates a proofreading report.

The review report includes the review results of the original text, that is, the determination of compliance or non-compliance. At the same time, the report also details the matching of the review rules, indicating which rules are satisfied or not satisfied. In addition, the report also provides possible suggestions for improvement. Based on the scheme's analysis of the original text and the review rules, it gives suggestions for optimizing the original text to improve its accuracy and compliance.

Through intelligent proofreading and the generation of proofreading reports, the scheme can automatically proofread original text, reducing the need for manual intervention and improving the efficiency and accuracy of proofreading. The program flow is shown in Figure (5).

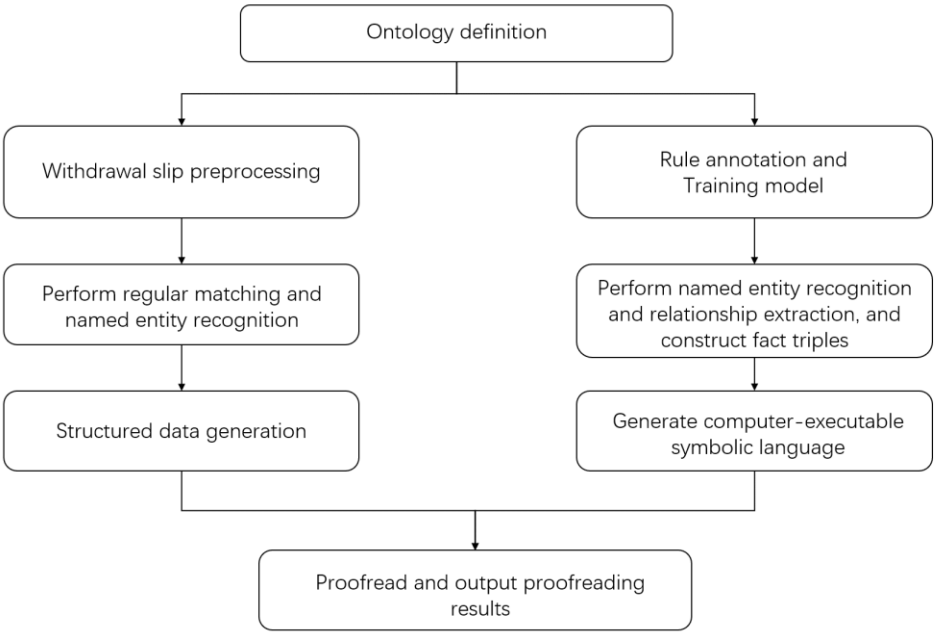

Figure (5): Intelligent cost review scheme process

IV. EXPERIMENT

In order to evaluate the performance of the intelligent scheme for power engineering cost review based on natural language processing, we conducted experiments, including the evaluations of the NER model, RE model, and the intergration. We conduct a comprehensive and in-depth evaluation of our proposed method using an experimental setup of five random initializations, and all reported numbers are the average of the results of five random initialization runs

4.1 DATA SETS AND DIVISION OF DATA SETS

In order to verify the effectiveness of our scheme, Two data sets were adopted. One is a public construction engineering review data set, which is used to verify the information extraction capability of the scheme and contains 8 defined entity types. At the same time, a data set for proofreading in power engineering

was created. A batch of unstructured original text of cost information and proofreading rules were first collect and organized. These data were then annotated by domain experts and contains a total of 8 defined entity types and 5 types of relationships, forming the final data set. Each data set were divided into a training set and a test set in a ratio of 8:2.

4.2 EXPERIMENTAL SETTINGS

We randomly initialize the experimental setup five times and report their average, using the Huggingface package to fine-tune the pre-trained chinese-bert-wwm-ext. We use NVIDIA 2080 GPU to train the model. We consider the AdamW optimizer (Loshchilov and Hutter, 2019) with the learning rate set to 3e−5 and weight decay set to 0.01. The batch size was set to 32. The number of training epochs is 3.

4.3 EXPERIMENTAL SETTINGS

4.3.1 MODEL TRAINING

We used a NER model and trained it using deep learning methods.The annotated requisition data set was used for semantic annotation to train the model to accurately extract device nouns and their corresponding attributes.

4.3.2 EVALUATION INDICATORS

To evaluate the performance of the NER model, Evaluation metrics, including precision, recall, and F1 score were adopted. The precision rate indicates the correct proportion of device nouns and attributes extracted by the model, and the recall rate indicates the proportion of device nouns and attributes correctly extracted by the model to the total correct device nouns and attributes. The F1 value is the harmonic average of the precision rate and the recall rate.

4.3.3 COMPARE TO BASELINE

HMM

HMM(Hidden Markov Model) is a classic probabilistic graphical model that is widely used in natural language processing, speech recognition,etc. Its advantage is that it can model the hidden state in sequence data and make inferences based on the observed data, so it performs well in sequence labeling tasks such as part-of-speech tagging and speech recognition. However, HMM models perform weakly when dealing with long-term dependencies and complex-structured data, making it difficult to capture deeper semantic information.

CNN

CNN(Convolutional Neural Network) is a deep learning model that has achieved great success in computer vision, and it also has important applications in natural language processing. It extracts features from input data through convolution operations and pooling operations, and is suitable for processing fixed-length sequence data, such as text classification and sentiment analysis. Compared with traditional methods, CNN can automatically learn local features of data, but its ability to process long-term dependencies in sequence data and temporal information within the sequence is relatively limited.

LSTM

LSTM(Long Short-Term Memory Network) is a variant of Recurrent Neural Network (RNN) specifically designed to solve the memory problem of long sequence data. It effectively captures long-term dependencies through a gating mechanism (forgetting gate, input gate, output gate), and can learn temporal information in sequence data. LSTM has achieved remarkable results in tasks such as machine translation, speech recognition, and text generation, and has become an important tool in the field of natural language processing.

BILSTM

BiLSTM(Bidirectional LSTM) is a variant of LSTM that introduces hidden layers in both forward and backward directions into the model, which can better capture the bidirectional information in sequence data. It performs well in sequence annotation tasks such as part-of-speech tagging and NER, and can better understand contextual information and improve the performance of the model.

LSTM+CRF

The LSTM+CRF(Conditional Random Field) model combines LSTM and CRF models. It uses the LSTM network to extract sequence features, and globally optimizes sequence labeling through the CRF model, thereby improving the performance of the sequence labeling task. This model has achieved good results in tasks such as NER and part-of-speech tagging, and can effectively handle local features and global constraints in sequence data.

By comparing these baseline models, we can have a more comprehensive understanding of their respective characteristics and applicable scenarios, so that we can choose the appropriate model to solve specific problems in practical applications.

4.3.4 EXPERIMENTAL RESULTS

The experimental results are shown in Table 2. The model we trained achieved excellent performance on data set for proofreading in power engineering. In terms of evaluation indicators, our model achieved an F1 value of 100% and an F1 value of 73% on two data sets respectively. On the construction engineering review dataset, our method outperforms LSTM+CRF F1 by nearly 5%. This shows that our model can accurately identify device names and attribute information in irregular text.

Table 2: Experimental Results of NER

	Proofreading data set in the field of power engineering			construction engineering review data set		
	P	R	F1	P	R	F1
HMM	91.95%	98.49%	95.11%	56.73%	61.41%	58.98%
CNN	97.79%	97.71%	97.43%	66.31%	66.23%	65.81%
LSTM	97.87%	97.80%	97.55%	68.44%	69.51%	67.60%
BiLSTM	98.42%	98.41%	98.32%	68.57%	69.47%	67.96%
LSTM+CRF	99.73%	99.73%	99.73%	70.33%	68.92%	68.11%
ours	99.99%	99.99%	99.99%	70.00%	75.00%	73.00%

4.4 RELATIONSHIP EXTRACTION EXPERIMENT

4.4.1 MODEL TRAINING

The experimental model comes from the above model by removing its CRF layer. The annotated review rule data set was used for semantic annotation, and to train the model to automatically extract device nouns and their corresponding attribute information in the rule text. At the same time, Five relationship were defined: greater than, less than, not greater than, not less than, and equal, which were used to jointly extract entities and relationships in rules.

4.4.2 EVALUATION INDICATORS

In order to evaluate the performance of the RE model, we still use precision, recall and F1 value as evaluation indicators.

4.4.3 EXPERIMENTAL RESULTS

The experimental results of relation extraction are shown in Table 3. The relation extraction model trained also achieved satisfactory performance at 100% F1 value on the power review rules data set. The overall F1 value was 1.2% higher compared to BiLSTM. This shows that our model can effectively extract the device nouns and their corresponding attribute in the review rules, and accurately identify the relationship between them.

Table 3: Experimental results of RE

	BiLSTM			Ours		
	P	R	F1	P	R	F1
greater than	98.89%	96.74%	97.80%	99.99%	99.99%	99.99%
less than	98.99%	99.99%	99.49%	99.99%	99.99%	99.99%
equal to	99.35%	98.59%	98.97%	99.99%	99.99%	99.99%
not greater than	99.01%	99.34%	99.17%	99.99%	99.99%	99.99%
not less than	97.42%	99.62%	98.51%	99.99%	99.99%	99.99%
Total	98.73%	98.86%	98.79%	99.99%	99.99%	99.99%

4.5 EXPERIMENTAL SETTINGS

We integrated the NER and RE models into the intelligent scheme for power project cost review to conduct overall performance evaluation. The schematic diagram of proofreading is shown in Figure (6) and Figure (7). After information extraction and rule interpretation, the structured device information and symbolic

rule information are obtained. The matching process is to find the rules of the corresponding device based on the device information, and then compare the conditions according to the order of the rules. If the conditions in the set does not met current rule, compare the next rule. After the set is met, compare whether the constraints of the conclusion set are met. If not, it does not meet the review rules, and a non-compliance report will be given indicating which rules have not been met, and provides modification suggestions to increase or decrease the corresponding value based on the rules themselves. If the conditions set of all rules is not met, then the cost information is also considered qualified.

4.5.1 DATA PREPARATION

We gathered the collected the original unstructured text of the power project data, selected a part of it, and modified the attributes and values in it so that it didn't meet the conditions of the review rules as counterexamples for testing. We follow the principle of equal numbers and set the number of positive examples and counterexamples to 1:1.

4.5.2 EXPERIMENTAL SETTINGS

We used the original unstructured text in the test set as input, then automatically reviewed it, and recorded the match between the results and the rules. At the same time, we also recorded the suggestions for improvement given.

4.5.3 EVALUATION INDICATORS

We used accuracy as an evaluation indicator to evaluate the scheme's review performance for the original unstructured text of the power project. Additionally, we evaluate the accuracy and usefulness of the scheme's suggestions for improvements.

Through the above experiments, we can evaluate the performance of the intelligent cost review scheme for power engineering on NER and RE tasks, and comprehensively evaluate the overall review and report generation capabilities of the scheme.

4.5.4 EXPERIMENTAL RESULTS

The overall scheme evaluation results show that the scheme shows high accuracy and reliability in review tasks. The scheme can automatically conduct intelligent proofreading based on the input text of the original unstructured text and the executable format of the proofreading rules, and provide corresponding proofreading results. On the test set, the scheme achieved 100% accuracy, proving the effectiveness of the scheme in improving the accuracy and compliance of cost information proofreading.

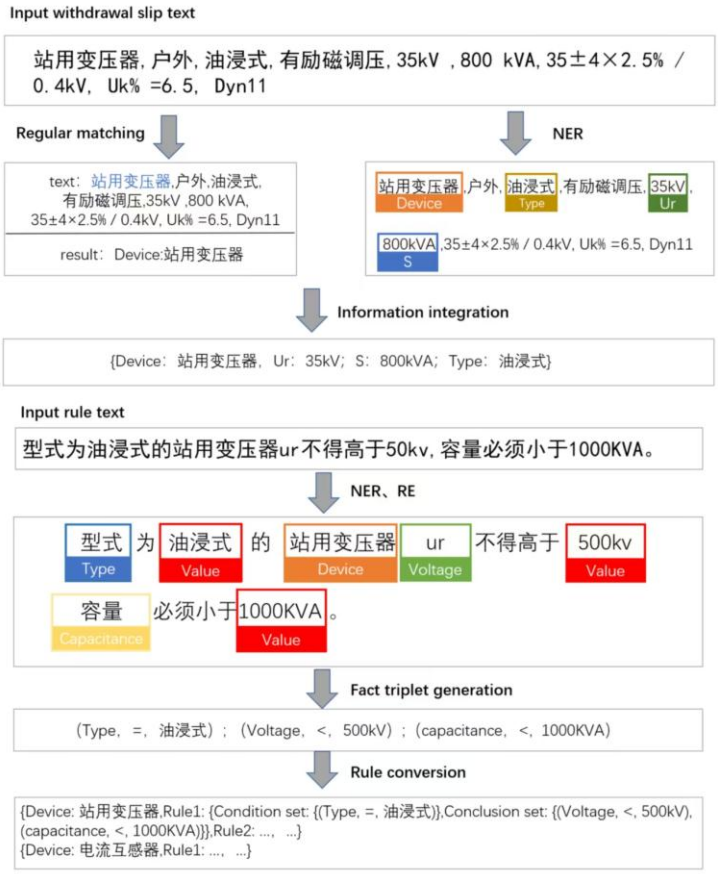

Figure (6): Example of Information Extraction and Rule Interpretation in Intelligent Cost Review Scheme

V. CONCLUSION

This research is dedicated to the development of an intelligent cost review scheme for power project based on Natural Language Processing. It uses NER and RE technology to achieve automatic extraction of proofreading information and automatic transformation of review rules. Experimental results show that our method achieves excellent performance in the NER and RE tasks of power review rules, which is of great significance for realizing automated cost review in power engineering. Through the NER model, we can accurately identify the name and attribute information of the power equipment in the original unstructured text, providing accurate basic data for the subsequent transformation of review rules. The training of the RE model enables the scheme to accurately transform the proofreading rules and generate the corresponding computer symbolic language to realize automated proofreading. Our scheme can automatically perform intelligent proofreading based on executable format of the original unstructured text and proofreading rules, and provide corresponding proofreading results, providing an efficient and accurate solution for power engineering project proofreading work.

However, in practical applications, scheme implementation still faces some challenges. The diversity and complexity of different original unstructured texts and review rules may have an impact on the performance of the scheme, so the flexibility and adaptability of the scheme need to be further improved. In addition, in order to train the model, a large amount of annotated data and computing resources are required, and the accuracy of the scheme also needs to be ensured through verification and tuning of more actual cases. In future research, we will continue to improve the performance and functionality of the scheme, by enhancing the accuracy of NER and RE models, increasing the flexibility of the scheme to adapt to different engineering projects, and optimizing the training process and validation methods of the scheme. We believe that this scheme will bring a more efficient and accurate review result to the power engineering industry and promote the smooth progress of projects.y.

REFERENCES

[1]. L B, L., Chen, Y. Z., & Yus, W. "Research on information extraction: a survey." Computer Engineering and Applications, 2003, 39(10), 1-5. (in Chinese)

[2]. Zhang, S. X. "Research on key technologies of the information extraction." Diss., Beijing University of Posts and Telecommunications, 2007. (in Chinese)

[3]. Socher, R., Huval, B., Manning, C. D., et al. "Semantic compositionality through recursive matrix-vector spaces." Proceedings of the 2012 joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea: Association for Computational Linguistics, 2012, 1201-1211.

[4]. Ebrahim, J., & Dou, D. J. "Chain based RNN for relation classification." Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado: Association for Computational Linguistics, 2015, 1244-1249.

[5]. Sundermeyer, M., Schlueter, R., & Ney, H. "LSTM neural networks for language modeling." Proc. Interspeech 2012, Portland, OR, USA: ISCA, 2012, 194-197.

[6]. Zhou, P., Shi, W., Tian, J., et al. "Attention-based bidirectional long short-term memory networks for relation classification." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany: Association for Computational Linguistics, 2016, 207-212.

[7]. LW, J., & Qi, F. "Chinese entity relation extraction based on multi-features self-attention Bi-LSTM." Journal of Chinese Information Processing, 2019, 33(10), 47-56. (in Chinese)

[8]. Zeng, D. J., Liu, K., Lai, S. W., et al. "Relation classification via convolutional deep neural network." Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014, 2335-2344.

[9]. Shen, Y. T., & Huang, X. J. "Attention-based convolutional neural network for semantic relation extraction." Proceedings of CoLING 2016, the 26th international Conference on Computational Linguistics: Technical Papers, Osaka, Japan: The COLING 2016 Organizing Committee, 2016, 2526-2536.

[10]. Wang, L. L., Cao, Z., De Melo, G., et al. "Relation classification via multi-level attention CNNs." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany: Association for Computational Linguistics, 2016, 298-1307.

[11]. Li, Q., Li, L. L., Wang, W. N., et al. "A comprehensive exploration of semantic relation extraction via pre-trained CNNs." Knowledge-based Systems, 2020, 194, 105488.

[12]. Bai, T., Guan, H. T., Wang, S., et al. "Traditional Chinese medicine entity relation extraction based on CNN with segment attention." Neural Computing and Applications, 2022, 34(4), 2739-2748.

[13]. Cao, W. D., Xu, X. L. "Entity relationship extraction based on R-BERT-CNN." Computer Applications and Software, 2023, 40(4), 222-229. (in Chinese)

[14]. Radford, A., Narasimhan, K., Salimans, T., et al. "Improving language understanding by generative pre-training." 2018-06-11.

[15]. Devlin, J., Chang, M. W., Lee, K., et al. "BERT: pre-training of deep bidirectional transformers for language understanding." Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN: Association for Computational Linguistics, 2019, 4171-4186.

[16]. Wu, S. C., & He, Y. F. "Enriching pre-trained language model with entity information for relation classification." Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China: Association for Computing Machinery, 2019, 2361-2364.

[17]. Huang, Y., Li, Z. X., Deng, W., et al. "D-BERT: incorporating dependency-based attention into BERT for relation extraction." ACM Transactions on Intelligent Technology, 2021, 6(4), 417-425.

[18]. Chen, X. L., Tang, L. Y., Hu, Y., et al. "Extracting entity and relation of landscape plant's knowledge based on ALBERT model." Journal of Geo-information Science, 2021, 23(7), 1208-1220. (in Chinese)

[19]. Xu, S. A., Sun, S. H., Zhang, Z. Y., et al. "BERT gated multi-window attention network for relation extraction." Neurocomputing, 2022, 492, 516-529.
[20]. Miwa, M., & Bansal, M. "End-to-end relation extraction using LSTMs on sequences and tree structures." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL2016), Berlin, Germany: Association for Computational Linguistics, 2016, 1105-1116.
[21]. Zeng, X. R., Zeng, D. J., He, S. Z., et al. "Extracting relational facts by an end-to-end neural model with copy mechanism." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia: Association for Computational Linguistics, 2018, 506-514.
[22]. Wei, Z. P., Su, J. L., Wang, Y., et al. "A novel cascade binary tagging framework for relational triple extraction." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), Virtual, Online, United States: Association for Computational Linguistics, 2020, 1476-1488.
[23]. Li, X. M., Luo, X. T., Dong, C. H., et al. "TDEER: an efficient translating decoding schema for joint extraction of entities and relations." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, 8055-8064.
[24]. Gao, C., Zhang, X., Li, L. Y., et al. "ERGM: a multi-stage joint entity and relation extraction with global entity match." Knowledge-based Systems, 2023, 271, 110550.
[25]. Sun, C. Z., Gong, Y. Y., Wu, Y. B., et al. "Joint type inference on entities and relations via graph convolutional networks." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy: Association for Computational Linguistics, 2019, 1361-1370.
[26]. Wang, Y. C., Yu, B. W., Zhang, Y. Y., et al. "TPLinker: single-stage joint extraction of entities and relations through token pair linking." 2020-10-26. Retrieved from https://doi.org/10.48550/arXiv.2010.13415.
[27]. Shang, Y. M., Huang, H. Y., Mao, X. L. "OneRel: joint entity and relation extraction with one module in one step." Proceedings of the 36th AAAI Conference on Artificial intelligence (AAAI 2022), Virtual, Online: Association for the Advancement of Artificial Intelligence, 2022, 36, 11285-11293.
[28]. Wang, Y. J., Sun, C. Z., Wu, Y. B., et al. "UniRE: a unified label space for entity relation extraction." Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th international Joint Conference on Natural Language Processing, Virtual, Online: Association for Computational Linguistics, 2021, 220-231.
[29]. Mintz, M., Bills, S., Snow, R., et al. "Distant supervision for relation extraction without labeled data." Proceedings of the joint Conference of the 47th Annual Meeting of the ACL and the 4th international Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore: Association for Computational Linguistics, 2009, 1003-1011.
[30]. Fenves, S. J. "Tabular decision logic for structural design." J. Struct. Div. 92(6) (1966), 473–490, https://doi.org/10.1061/JSDEAG.0001567.
[31]. Eastman, C., Lee, J. M., Jeong, Y. S., Lee, J. K. "Automatic rule-based checking of building designs." Autom. Constr. 18(8) (2009), 1011–1033, https://doi.org/10.1016/j.autcon.2009.07.002.
[32]. Fuchs, S. "Natural Language Processing for Building Code Interpretation: Systematic Literature Review Report." 2021. Retrieved from https://www.researchgate.net/profile/Stefan-Fuchs-8/publication/351354243_Natural_Language_Processing_for_Building_Code_Interpretation_Systematic_Literature_Review_Report/links/6093496e299bf1ad8d7d8a0f/Natural-Language-Processing-for-Building-Code-Interpretation-Systematic-Literature-Review-Report.pdf.
[33]. Nadkarni, P. M., Ohno-Machado, L., Chapman, W. W. "Natural language processing: an introduction." J. Am. Med. Inform. Assoc. 18(5) (2011), 544–551, https://doi.org/10.1136/amiajnl-2011-000464.
[34]. Zhang, J., El-Gohary, N. M. "Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking." J. Comput. Civ. Eng. 30(2) (2016), 04015014, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000346.
[35]. Xu, X., Cai, H. "Semantic approach to compliance checking of underground utilities." Autom. Constr. 109 (2020), 103006, https://doi.org/10.1016/j.autcon.2019.103006.
[36]. Zhang, J., El-Gohary, N. M. "Automated information transformation for automated regulatory compliance checking in construction." J. Comput. Civ. Eng. 29(4) (2015), B4015001, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000427.
[37]. Zhou, P., El-Gohary, N. "Ontology-based automated information extraction from building energy conservation codes." Autom. Constr. 74 (2017), 103–117, https://doi.org/10.1016/j.autcon.2016.09.004.
[38]. Zhou, Y. C., Zheng, Z., Lin, J. R., Lu, X. Z. "Integrating NLP and context-free grammar for complex rule interpretation towards automated compliance checking." Comput. Ind. 142 (2022), 103746, https://doi.org/10.1016/j.compind.2022.103746.
[39]. Xu, X., Cai, H. "Ontology and rule-based natural language processing approach for interpreting textual regulations on underground utility infrastructure." Adv. Eng. Inform. 48 (2021), 101288, https://doi.org/10.1016/j.aei.2021.101288.
[40]. Zheng, Z., et al. "Knowledge-informed semantic alignment and rule interpretation for automated compliance checking." Automation in Construction, vol. 142, 2022, pp. 104524, https://doi.org/10.1016/j.autcon.2022.104524.

Funding Statement: This work is supported by the Fundamental Research Funds for the Central Universities (2023MS137).