processes

Article

Key Technologies of Intelligent Question-Answering System for Power System Rules and Regulations Based on Improved BERTserini Algorithm

Ming Gao [1,2], Mengshi Li [1], Tianyao Ji [1,*], Nanfang Wang [2], Guowu Lin [2] and Qinghua Wu [1]

[1] College of Electric Power Engineering, South China University of Technology, Guangzhou 510640, China; 202011002583@mail.scut.edu.cn (M.G.); mengshili@scut.edu.cn (M.L.); wuqh@scut.edu.cn (Q.W.)
[2] School of Electrical Engineering, Guangzhou City University of Technology, Guangzhou 510800, China; wnf2184649296@163.com (N.W.); 16676676716@163.com (G.L.)
* Correspondence: tyji@scut.edu.cn

Abstract: With the continuous breakthrough of natural language processing, the application of intelligent question-answering technology in electric power systems has attracted wide attention. However, at present, the traditional question-answering system has poor performance and is difficult to apply in engineering practice. This paper proposes an improved BERTserini algorithm for the intelligent answering of electric power regulations based on a BERT model. The proposed algorithm is implemented in two stages. The first stage is the text-segmentation stage, where a multi-document long text preprocessing technique is utilized that accommodates the rules and regulations text, and then Anserini is used to extract paragraphs with high relevance to the given question. The second stage is the answer-generation and source-retrieval stage, where a two-step fine-tuning based on the Chinese BERT model is applied to generate precise answers based on given questions, while the information regarding documents, chapters, and page numbers of these answers are also output simultaneously. The algorithm proposed in this paper eliminates the necessity for the manual organization of professional question–answer pairs, thereby effectively reducing the manual labor cost compared to traditional question-answering systems. Additionally, this algorithm exhibits a higher degree of exact match rate and a faster response time for providing answers.

Keywords: intelligent question-answering system; improved BERTserini algorithm; rules and regulations; information retrieval

check for updates

Citation: Gao, M.; Li, M.; Ji, T.; Wang, N.; Lin, G.; Wu, Q. Key Technologies of Intelligent Question-Answering System for Power System Rules and Regulations Based on Improved BERTserini Algorithm. *Processes* **2024**, *12*, 58. https://doi.org/10.3390/pr12010058

Academic Editor: Peng Li

Received: 22 November 2023
Revised: 20 December 2023
Accepted: 22 December 2023
Published: 26 December 2023

1. Introduction

The intelligent question-answering system is an innovative information service system that integrates natural language processing, information retrieval, semantic analysis and artificial intelligence. The system mainly consists of three core parts, which are question analysis, information retrieval and answer extraction. Through these three parts, the system can provide users with accurate, fast and convenient answering services.

The representative systems of the intelligent question-answering system include:

(1) Rule-based algorithms (1960s–1980s). The question-answering system based on this pattern mainly relies on writing a lot of rules and logic to implement the dialogue. ELIZA [1], developed by Joseph Weizenbaum in the 1960s, was the first chatbot designed to simulate a conversation between a psychotherapist and a patient. PARRY [2] is a question-and-answer system developed in the 1970s that simulates psychopaths. The emergence of ELIZA and PARRY provided diverse design ideas and application scenarios for subsequent intelligent question-answering systems, thereby promoting the diversification and complexity of dialogue systems. However, the main problem of this model is its lack of flexibility and extensibility. It relies too much on rules or templates set by humans, and

consumes a lot of time and manpower. When the questions become complicated, it is difficult to obtain satisfactory answers through simple rules set by the model.

(2) Statistics-based algorithms (1990s–2000s). The question-answering system based on this model adopts the method of statistical learning to learn patterns and rules from a large number of dialogue data. Common algorithms include Vector Space Model [3] and Conditional Random Fields [4]. ALICE (Artificial Linguistic Internet Computer Entity) [5] is an open-source natural language processing project. The system in question is an open-domain question-answering platform capable of addressing queries across a multitude of subjects and domains. Jabberwacky [6] is an early intelligent chatbot employing machine learning and conversational models to enhance its responses continually. These systems are designed to train models that can learn the relationships between questions and answers present in the corpus. Therefore, these models can carry out more natural and smooth dialogue. However, the ability of context understanding and generalization ability is weak, so it is difficult to adapt to model sharing and transfer learning in various professional fields. Moreover, considering statistical models are trained on a large corpus, this kind of model may suffer from data bias when dealing with domain-specific problems and fail to provide accurate answers.

(3) Algorithms based on hybrid technology (2010s–early 2020s). The question-answering system, grounded on this model, can amalgamate diverse techniques encompassing rules, statistics, and machine learning. It leverages multiple input modalities, including speech, image, and text, to interoperate seamlessly. The overarching objective is to facilitate users in accomplishing specific tasks or goals within designated domains, such as booking, traveling, shopping, or ordering food. This synergistic integration of multifarious technologies and input modes fosters a more sophisticated and intelligent dialogue system. Typical question-answering systems based on the hybrid technology model include Apple's Siri [7], Microsoft's Cortana [8], Amazon's Alexa [9], Facebook's M [10], and Google's Google Assistant [11]. These systems are centered around artificial intelligence and natural language processing technology, aiming to furnish users with personalized and convenient information and services to cater to diverse needs.

The system built based on this pattern has stronger context understanding and personalized customization, but there are two shortcomings: first, the quality of dialogue in such a system is not stable; second, the generalization ability of the model is limited. It is difficult to realize model sharing, transfer learning and answer generation in professional fields. The training of this model requires excessive investment in computing and data resources, and its training and deployment speed is slow.

(4) Algorithms based on pre-trained language (2020s). The model is based on pre-trained language models such as BERT [12], GPT (Generative Pre-trained Transformer) [13], etc. These models are pre-trained on large-scale data and they learn rich language representation and context understanding skills to generate more natural, fluid, and accurate responses. In addition, through the supervised training on domain-specific question-answering datasets, the question-answering system can answer questions in specialized professional fields. Ref. [14] primarily investigates a question-answering system for agriculture. This system utilizes artificial intelligence technology and relevant datasets to provide farmers with information on topics such as weather, market prices, plant protection, and government plans. Ref. [15] proposed a TD-BERT model based on BERT. This model leverages the powerful semantic representation capabilities of BERT and integrates target information to enhance the accuracy of sentiment classification. Ref. [16] proposed a BERTserini algorithm which improves the exact match rate of the question-answering system. In comparison to the original BERT algorithm, the proposed method surpasses its processing byte limit and can provide accurate answers for multi-document long texts.

Although systems built on the BERTserini algorithm perform well on public datasets, there are some problems in the application in professional fields such as electrical power engineering. Considering the low exact match rate and poor answer quality, engineering applications of these models are challenging. The problems are mainly caused by the following aspects.

(1) Lack of model expertise: Language models such as BERT or GPT are usually pre-trained from large amounts of generic corpus collected on the Internet. However, the digital realm offers limited professional resources pertaining to industries like electrical power engineering. As a result, the model has insufficient knowledge reserve when dealing with professional question, which affects the quality of the answers; (2) Differences in document format: There are significant differences between the format of documentation in the electrical power engineering field and that of public datasets. The documents in the electrical power engineering field often exhibit unique formatting, characterized by an abundance of hierarchical headings. It is easy to misinterpret the title as the main content and mistakenly use it as the answer to the question, leading to inaccurate results; (3) Different scenario requirements: Traditional answering systems do not need to pay attention to the source of answers in the original document. However, a system designed for professional use must provide specific source information for its answers. If such information is not provided, there may arise doubts regarding the accuracy of the response. This further diminishes the utility of the application in particular domains.

This paper proposes an improved BERTserini algorithm to construct an intelligent question-answering system in the field of electrical power engineering. The proposed algorithm is divided into two stages:

The first stage is text segmentation. During this phase, the text is segmented and preprocessed. Firstly, a multi-document long text preprocessing method that supports rules and regulations text is proposed. This approach can accurately segment rules and regulations text and generate an index file of answer location information. By doing so, the system can better comprehend the structure of the regulation text, enabling it to locate the answer to the user's question more accurately. Secondly, through the FAQ [17] pre-module, high-frequency questions are intercepted for question pre-processing. This module matches and classifies user-raised questions based on a pre-defined list of common questions, intercepting and addressing high-frequency issues. This reduces the repetition of processing the same or similar problems and enhances the system's response efficiency. Finally, Anserini [18] is employed to extract several paragraphs highly relevant to user problems from multi-document long text. Anserini is an information-retrieval tool based on a vector space model that represents a user question as a vector and each paragraph in a multi-document long text as a vector. By calculating the similarity between the user problem vector and each paragraph vector, several paragraphs with high relevance to the user problem can be selected. These paragraphs serve as candidate answers for the system to further analyze and generate the final answer.

The second stage is the answer-generation and source retrieval stage. During this phase, the Chinese Bert model undergoes fine-tuning [19], which comprises two steps involving key parameter adjustments. This process enhances the model's comprehension of the relationship between the question and the answer, thereby improving the accuracy and reliability of the generated response. Subsequently, based on the input question, the Bert model extracts several candidate answers from the N paragraphs with the highest similarity to the question, as determined by Anserini. The user can then filter through these multiple relevant paragraphs to identify the answer that best aligns with their query. Finally, the candidate answers are weighted, and the highest-rated answer is outputted along with the chapter and position information of the answer in the original document. This approach facilitates users in quickly locating the most accurate answer while providing pertinent contextual information.

The improved BERTserini algorithm proposed in this paper has three main contributions.

(1) The proposed algorithm implements multi-document long text preprocessing technology tailored for rules and regulations text. Through optimization, the algorithm segments rules and regulations into distinct paragraphs based on its inherent structure and supports answer output with reference to chapters and locations within the document.

The effectiveness of this pretreatment technology is reflected in the following three aspects: first, through accurate segmentation, paragraphs that may include questions can be extracted more accurately, thus improving the accuracy of answer generation. Secondly, the original Bert model exhibits a limitation that it outputs the heading of rules and regulations text as the answer frequently. To address this issue, an improved BERTserini algorithm has been proposed. Finally, the algorithm is able to accurately give the location information of answers in the original document chapter. The algorithm enhances the comprehensiveness and accuracy of reading comprehension, generating answers to questions about knowledge and information contained in professional documents related to the field of electric power. Consequently, this leads to a marked improvement in answer quality and user experience for the question-answering system.

(2) The proposed algorithm optimizes the training of the corpus in the field of electrical power engineering and fine-tunes the parameters of the large language model. This method eliminates the necessity for the manual organization of professional question–answer pairs, knowledge base engineering, and manual template establishment in BERT reading comprehension, thereby effectively reducing labor costs. This enhancement significantly enhances the accuracy and efficiency of the question-answering system.

(3) The proposed algorithm has been developed for the purpose of enhancing question-answering systems in engineering applications. This algorithm exhibits a higher degree of exact match rate of questions and a faster response for providing answers.

The remaining sections of this article are organized as follows. Section 2 provides an introduction to the background technology of intelligent question-answering systems. Section 3 describes the procedural steps of an improved BERTserini algorithm. Section 4 presents the experimental results of the proposed algorithm and its implementation in engineering applications. Finally, Section 5 draws conclusions.

2. Background of the Technology

2.1. FAQ

Frequently Asked Questions (FAQs) are a collection of frequently asked questions and answers designed to help users quickly find answers to their questions [17]. The key is to build a rich and accurate database of preset questions, which consists of questions and the corresponding answers. They are manually collated from the target documents. The FAQ provides an answer that corresponds to the user's question by matching it with the most similar question.

2.2. BM25 Algorithm

The Best Match 25 (BM25) algorithm [18,19] was initially proposed by Stephen Robertson and his team in 1994 and applied to the field of information retrieval. It is commonly used to calculate the relevance score between documents and queries. The main logic of BM25 is as follows: Firstly, the query statement involves word segmentation to generate morphemes. Then, the relevance score between each morpheme and the search result is calculated.

Finally, by weighting summing the relevance scores of the morpheme with the search results, the relevance score between the retrieval query and the search result documents is obtained. The formula for calculating BM25 algorithm is as follows:

$$Score(D, Q) = \sum_{i}^{n} W_i \cdot R(q_i, D) \tag{1}$$

In this context, Q represents a query statement, q_i represents a morpheme obtained from Q. For Chinese, the segmented results obtained from tokenizing query Q can be considered as morpheme q_i. D represents a search result document. W_i represents the weight of morpheme q_i, and $R(q_i, D)$ represents the relevance score between morpheme q_i and document D. There are multiple calculation methods for weight parameter W_i, with

Inverse Document Frequency (IDF) being one of the commonly used approaches. The calculation process for IDF is as follows:

$$IDF(q_i) = \log(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}) \tag{2}$$

In the equation, N represents the total number of documents in the index, and $n(q_i)$ represents the number of documents that contain q_i.

Finally, the relevance scoring formula for the BM25 algorithm can be summarized as follows:

$$Score(D, Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)} \tag{3}$$

where k_1 and b are adjustment factors, $f(q_i, D)$ represents the frequency of morpheme q_i appearing in document D, $|D|$ denotes the length of document D, and $avgdl$ represents the average length of all documents.

2.3. Anserini

Anserini [20] is an open-source information retrieval toolkit that supports various text-based information retrieval research and applications. The goal of Anserini is to provide an easy-to-use and high-performance toolkit that supports tasks such as full-text search, approximate search, ranking, and evaluation on large-scale text datasets. It enables the conversion of text datasets into searchable index files for efficient retrieval and querying. Anserini incorporates a variety of commonly used text retrieval algorithms, including the BM25 algorithm. With Anserini, it becomes effortless to construct a BM25-based text retrieval system and perform efficient search and ranking on large-scale text collections. The flowchart of the algorithm is illustrated in Figure 1.

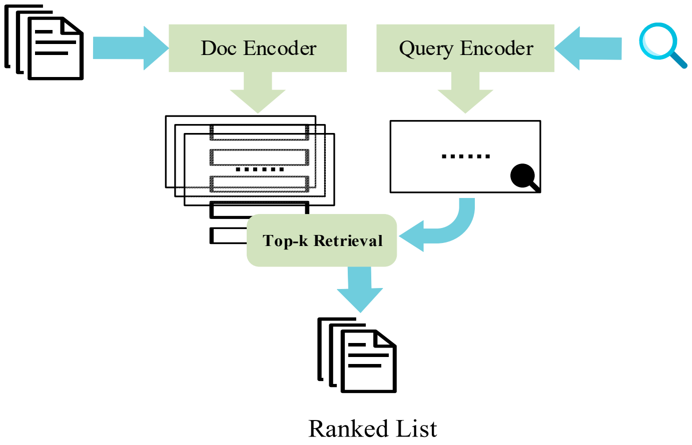


Figure 1. The flowchart of the Anserini algorithm.

2.4. BERT Model

Bidirectional Encoder Representations from Transformers (BERT) [12] is a pre-trained language model proposed by Google in 2018. The model structure is shown in Figure 2. In the model, E_i represents the encoding of words in the input sentence, which is composed of the sum of three word embedding features. The three word embedding features are Token Embedding, Position Embedding, and Segment Embedding. The integration of these three words embedding features allows the model to have a more comprehensive understanding of the text's semantics, contextual relationships, and sequence information, thus enhancing the BERT model's representational power. The transformer structure in the figure is represented as Trm. The T_i represents the word vector that corresponds to the trained word E_i.

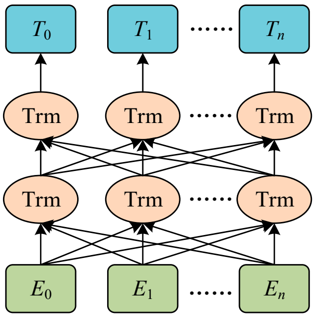


Figure 2. Architecture of BERT.

BERT exclusively employs the encoder component of the Transformer architecture. The encoder is primarily comprised of three key modules: Positional Encoding, Multi-Head Attention, and Feed-Forward Network. Input embeddings are utilized to represent the input data. Addition and normalization operations are denoted by "Add&norm". The fundamental principle of the encoder is illustrated in Figure 3.

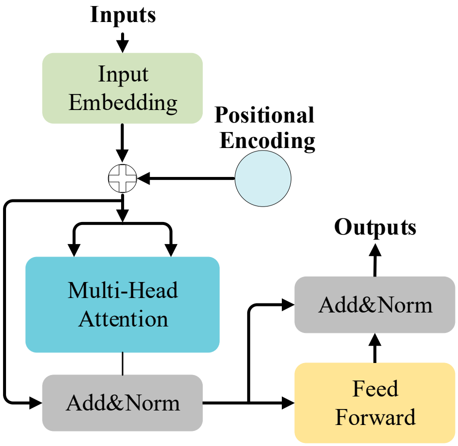


Figure 3. Transformer encoder principle.

In recent years, several Chinese BERT models have been proposed in the Chinese language domain. Among these, the chinese-BERT-wwm-ext model [21] released by the HIT·iFLYTEK Language Cognitive Computing Lab (HFL) has gained significant attention and serves as a representative example. This model, based on the original Google BERT model, underwent further pretraining using a total vocabulary of 5.4 billion words, including Chinese encyclopedia, news, and question-answer datasets. The model adopts the Whole Word Masking (wwm) strategy, which is an improvement tailored to Chinese language characteristics. In Chinese processing, as words are composed of characters, and a word may consist of one or more characters, it becomes necessary to mask the entire word rather than just a single character. The wwm strategy is designed to better understand and capture the semantics of Chinese vocabulary. In summary, this model is an improved Chinese version of BERT that, through whole-word masking, exhibits enhanced performance in Chinese language understanding.

2.5. BERTserini Algorithm

The architecture of BERTserini algorithm [16] is depicted in Figure 4. The algorithm employs the Anserini information extraction algorithm in conjunction with a pretraining BERT model. In this algorithm, the Anserini retriever is responsible for selecting text paragraphs containing the answer, which are then passed to the BERT reader to determine the answer scope. From Figure 4, it can be observed that BERTserini is an intelligent question-answering system that combines the BERT language model with the Anserini information retrieval system. It synergistically harnesses the powerful language understanding capabilities of BERT and the efficient retrieval functionalities of Anserini. This algorithm exhibits significant advantages over traditional algorithms. It demonstrates fast execution speed similar to traditional algorithms while also possessing the characteristics of end-to-end matching, resulting in more precise answer results. Furthermore, it supports extracting answers to questions from multiple documents. This algorithm is primarily applied to open-domain question-answering tasks, where the system needs to find answers to questions from a large amount of unstructured text.

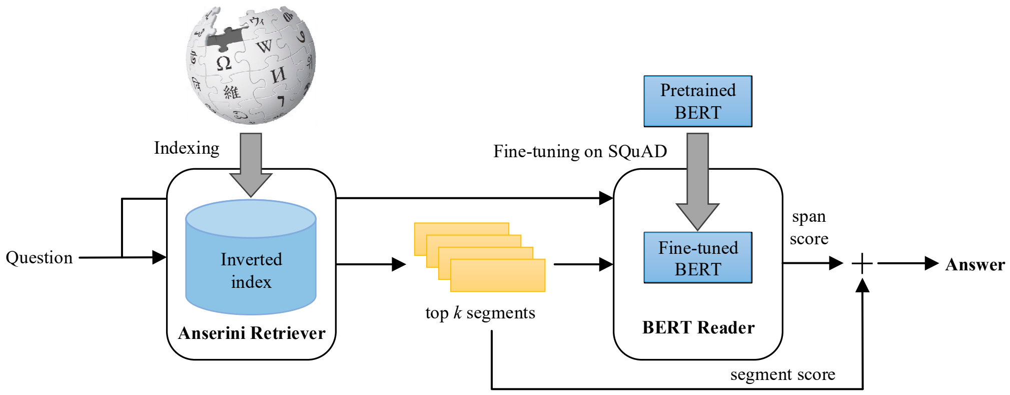


Figure 4. Architecture of BERTserini.

3. Improved BERTserini Algorithm

3.1. Algorithm Description

The improved BERTserini algorithm presented in this paper can be divided into two stages, and the flowchart is illustrated in Figure 5.

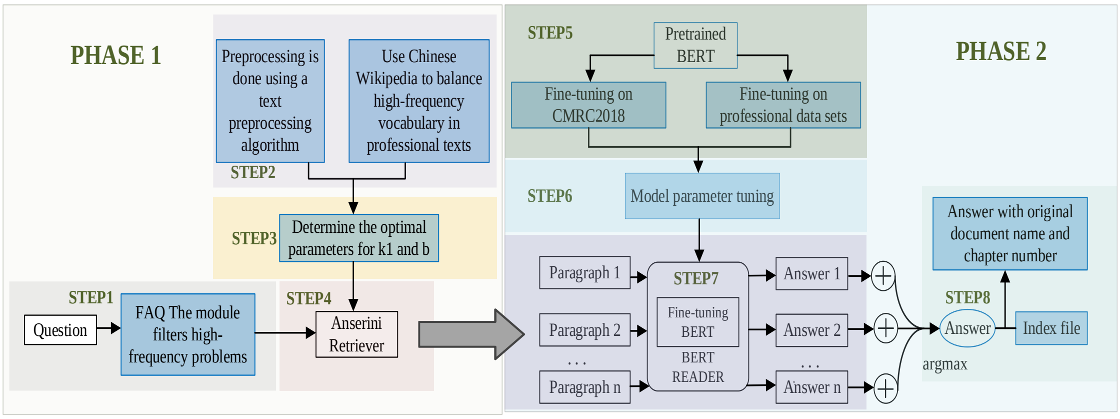

PHASE 1
Preprocessing is done using a text preprocessing algorithm
Use Chinese Wikipedia to balance high-frequency vocabulary in professional texts
STEP2
STEP3
Determine the optimal parameters for k1 and b
STEP1
Question
FAQ The module filters high-frequency problems
STEP4
Anserini Retriever
STEP5
Pretrained BERT
Fine-tuning on CMRC2018
Fine-tuning on professional data sets
STEP6
Model parameter tuning
Paragraph 1
Paragraph 2
...
Paragraph n
STEP7
Fine-tuning BERT
BERT READER
Answer 1
Answer 2
...
Answer n
PHASE 2
Answer with original document name and chapter number
STEP8
Answer
Index file
argmax

Figure 5. Flowchart of the proposed algorithm.

(1) Phase 1: Text Segmentation Stage

The first stage is text segmentation stage, which comprises two key components: (1) Question preprocessing: The FAQ module is utilized to intercept high-frequency questions in advance, thereby achieving question preprocessing. If the FAQ module cannot provide an answer that corresponds to the user's query, then the query is transferred to the subsequent stage of paragraph extraction. Anserini retrieval technology is utilized for paragraph extraction, enabling the rapid extraction of highly relevant paragraphs which are pertinent to user queries within multi-document long text. (2) Document preprocessing: Due to the high degree of keyword overlap in power regulation documents. The paper proposes a multi-document long text preprocessing method supporting regulation texts, which can accurately segment the regulation texts and support the retrieval and tracing of the answer chapters' sources.

STEP 1: The FAQ module filters out high-frequency problems.

The FAQ module is designed to pre-process questions by intercepting and filtering out high-frequency problems. To achieve this, the module requires a default question library that contains a comprehensive collection of manually curated questions and their corresponding answer pairs from the target document. By matching the most similar question to the user's inquiry, the FAQ module can efficiently provide an accurate answer based on the corresponding answer to the question.

The FAQ module employs ElasticSearch, an open-source distributed search and analysis engine, to match user queries in a predefined question library. ElasticSearch is built upon the implementation of Lucene, an open-source full-text search engine library released by the Apache Foundation, and incorporates Lucene's BM25 text similarity algorithm. This algorithm calculates similarity by evaluating word overlap between the user query's text and the default question library, as shown in (3).

The FAQ module will directly return the preset answer to the matched question if the BM25 score returned by ElasticSearch exceeds the predetermined threshold. In cases where the return score falls below this threshold, instead of returning an answer, the question is referred to subsequent steps.

STEP 2: Text preprocessing and document index generation.

This step involves two tasks.

The first task is due to the high overlap of professional terminology in similar regulatory texts. If Anserini is directly used to retrieve and calculate the paragraphs in professional documents, it may result in an issue where certain professional terms have lower weights W_i in Equation (2). The main reason for this is that if we assume q_i is a power industry term initially used as a retrieval keyword, its occurrence in multiple professional documents results in a larger value for $n(q_i)$ in the calculation of Equation (2). This value becomes essentially close to the total document count N, leading to a decrease in the calculated result of $IDF(q_i)$. The issue arising from this is that when retrieving the professional term q_i, the original expectation was to find paragraphs or documents strongly related to it. However, due to the decrease in $IDF(q_i)$, the probability of finding paragraphs or documents strongly associated with this professional term is actually reduced. Conversely, in this situation, some non-specialized terms may have relatively larger IDF values. This situation is exactly opposite to the intended calculation goal of the IDF algorithm. For keywords that possess strong discriminative power for document categories, the expectation is that documents containing such keywords should be relatively scarce in the corpus. Consequently, the IDF value for these keywords should be larger.

For example, "generator" is a professional term and keyword in power regulatory texts. However, due to its high frequency across multiple professional documents, the IDF value calculated according to Equation (2) may not be high. On the other hand, non-specialized terms like "tool" may have a higher IDF value because of their infrequent occurrence in professional documents. As a result, after inputting the retrieval query, Anserini calculates and retrieves documents that are not strongly related to the professional term, contrary to the intended outcome. Therefore, in the process of constructing the index file, besides incorporating regulatory texts, Chinese Wikipedia textual data has been included. This action increases the value of N, consequently enlarging the gap between and $n(q_i)$. This adjustment elevates the calculated IDF value for professional terms according to Equation (2), thereby mitigating the adverse effects caused by the high frequency of certain professional terms.

The second task Involves proposing a multi-document lengthy text preprocessing algorithm that supports regulatory texts. This algorithm accurately segments regulatory texts, retains information about the sections to which paragraphs belong, and generates an index file. The specific method is as follows:

Convert documents in .pdf or .docx format to plain text in .txt format.

Remove irrelevant information such as header/footer and page number.

Use regular expressions to extract the title number from the text (for example: Section 3.3.1), and match the title number to the text.

Use rules to filter out paragraphs in the text such as tables and pictures that are not suitable for machine reading comprehension.

Use Anserini to divide the text title number into words and index the corresponding text.

STEP 3: Determine the two parameters k1 and b.

The k1 and b parameters utilized in the Anserini module are empirically selected to determine the optimal parameters for this study. A specific methodology is employed, starting from 0.1 within their respective value ranges and incrementing by 0.05 to systematically explore all possible combinations of k1 and b values. The selection of the best k1 and b values is based on the accuracy assessment of the second stage Bert reading comprehension module.

STEP 4: Extract paragraphs and generate paragraph scores.

Based on the user's question, Anserini extracts relevant paragraphs from the preprocessed document by filtering out those that are not related to the query. It then matches the question with the paragraphs in the index and selects the top N paragraphs with the highest relevance to the question. This paragraph is evaluated using the BM25 algorithm, as specified in Equations (1)–(3), and is denoted by $\mathrm{S}_{\mathrm{anserini}}$.

(2) Phase 2: answer generation and source retrieval stage

The second stage is the answer-generation and source-retrieval stage. After undergoing two steps of fine-tuning and key parameter tuning, the model is capable of extracting accurate answers from N paragraphs based on the given question. Additionally, the model can output the chapter information of the answer in the original document according to the index file.

STEP 5: Select the appropriate Chinese Bert model and fine-tune it.

In this research, the Chinese-Bert-WWM-EXT Base model is chosen as the foundational framework. The initial step involves fine-tuning the model using the Chinese Open domain Question answering dataset (CMRC2018). Subsequently, a second round of fine-tuning is conducted by employing the training exam questions related to rules and regulations as specialized datasets.

STEP 6: Algorithm parameter tuning.

Based on the structural and characteristic features of regulatory documents, the following five crucial parameters of the improved BERTserini algorithm have been optimized:

paragraph_threshold. The paragraph threshold is employed to exclude paragraphs with Anserini scores below this specified limit, thereby conserving computational resources.

phrase_threshold. The answer threshold serves as a filter, excluding responses with a Bert reader score below the specified limit.

remove_title. Removes the paragraph title. If this item is True (y = True, n = False), paragraph headings are not taken into account when the Bert reader performs reading comprehension.

max_answer_length. The maximum answer length. The maximum length of an answer is allowed to be extracted when the Bert reader performs a reading comprehension task.

mu. Score weight is implemented to evaluate both the answer and paragraph using the Bert reader and Anserini extractor, subsequently calculating the final score value of the answer.

STEP 7: Extract the answers and give a reading comprehension score.

Bert is used to extract the exact answers to the question from the N paragraphs extracted by Anserini. The sum of the probability of starting and ending positions (logits) for each answer predicted by the model is used as the score of the answer generated by the Bert reading comprehension module. It can be expressed by the following equation:

$$S_{bert} = max(start\ logit) + max(end\ logit) \tag{4}$$

STEP 8: The candidate answers are scored by a comprehensive weighted score, rank the answers by score, output the answer with the highest score, and give the original document name and specific chapter information for the answer.

Use the following equation to calculate the overall weighted score of the answer:

$$S = (1 - \mu) \times S_{anserini} + \mu \times S_{bert} \quad (5)$$

The final score of the answer is calculated by the above formula. $S_{anserini}$ represents the BM25 score returned by the Anserini extracter, and S_{bert} represents the answer score returned by Bert. The answers are sorted by the calculated answer score, and the final output is the answer with the highest score. According to the index file, the original document name and chapter information are output together.

3.2. Main Innovations

(1) Multi-document long text preprocessing method which can process rules and regulations text and support answer provenance retrieval.

In this paper, a multi-document long text preprocessing method is proposed that facilitates answer provenance retrieval and can effectively process the rules and regulations text, which provides a technical path for the construction of intelligent question-answering system in specific professional fields. The innovation point of this method is reflected in STEP 2. This method divides the rules and regulations into chapters. The original document name of each paragraph and its chapter number information can be preserved. To address the issue of excessive frequency of certain proper nouns, the method incorporates text data from Chinese Wikipedia and performs balance processing. By incorporating a larger corpus, the frequency of a specific proper noun in the text can be effectively diminished, thereby mitigating its influence on the model. This innovative preprocessing method can improve the calculation effect of the subsequent reading comprehension module. The answer can be provided in the original document, including chapter and location information.

(2) Determination of optimal parameters of Anserini and improved BERTserini algorithm.

① Determination of the optimal parameters of Anserini. In STEP 3, the optimal parameters of Anserini are determined. All possible combinations of k1 and b are experimentally tried one by one. And the best value is selected according to the answer performance of the subsequent reading comprehension module questions. The determination of the optimal parameters of Anserini improves the performance of the intelligent question-answering system and the exact match of answers (EM).

② Determination of the optimal parameters of the improved BERTserini algorithm: In STEP 6, the optimal parameters of the improved BERTserini algorithm are determined. According to the structure and characteristics of regulation documents, five important parameters are optimized. Thus, the algorithm can determine the reasonable threshold of generating candidate answers when the Bert reading comprehension module performs the reading comprehension task. And the answer generation does not take into account the paragraph title and the optimal overall rating weight and other details that constitute high-quality questions and answers.

(3) Fine-tuning of multi-data sets for Bert reading comprehension model.

This step is illustrated in STEP 5. The Bert model is pre-trained using the CMRC2018 data, and a two-step fine-tuning was carried out using the existing rules and regulations exam questions. By making full use of data sets in different fields, the accuracy and generalization ability of the model are improved. This method achieves better results in question-answering system. At the same time, this method also reduces the time and labor cost required for the manual editing of question–answer pairs in traditional model training. It also significantly improves Bert's reading comprehension of rules and regulations.

(4) Clever use of FAQ.

The clever use of the FAQ is reflected in STEP1. In this paper, the existing rules and regulations are used to train and test questions, which constitutes the questions and answers pairs required by the pre-FAQ module to intercept some high-frequency questions. In this way, a low-cost FAQ module is constructed, which improves the answering efficiency of high-frequency questions, and also improves the exact match rate (EM) of the intelligent question-answering system.

4. Results Analysis of the Experiment

4.1. Data Description

4.1.1. Document Description

For the present study, a total of 30 documents including regulations, provisions, and operation manuals related to the theme of power safety are selected, such as a company power grid work regulations. The total size of the documents is 30.1 MB, and the intelligent system is required to preprocess all the content within the documents, perform machine reading comprehension, and efficiently answer questions.

4.1.2. Fine-Tuning Dataset Description

In this study, four datasets are experimented for fine-tuning the Bert model, which include Chinese Machine Reading Comprehension 2018 (CMRC2018) [22], Delta Reading Comprehension Dataset (DRCD) [23], Safety Procedure Test Item data set (SPTI), and a dataset generated through data augmentation based on documentations of a power grid company. The first two datasets are open-source. Among them, the CMRC2018 dataset contains a large amount of Chinese text. After fine-tuning, it can be adapted to specific domains or application scenarios, thereby improving performance. DRCD is also a Chinese machine reading comprehension dataset, primarily used to train and evaluate models in understanding Chinese texts and answering related questions. The text in the DRCD dataset is sourced from various authentic corpora, including Chinese Wikipedia, to ensure a simulation of real-world scenarios. Based on end-to-end manual evaluation, the results indicate that the model trained using CMRC2018 data performs the best in this study. Therefore, it has been selected as the fine-tuning training dataset. The dataset follows the format of the SQuAD dataset [24]. It consists of a total of 10,142 training samples, 3219 validation samples, and 1002 testing samples. The overall size of the dataset is 32.26 MB. The SPTI consists of 1020 training and examination questions related to electrical safety regulations.

4.1.3. BERT Model Description

In this study, the Chinese-BERT-wwm-ext model [25] released by the HFL is used for training.

4.1.4. Parameter Tuning Explanation for Improved BERTserini Algorithm

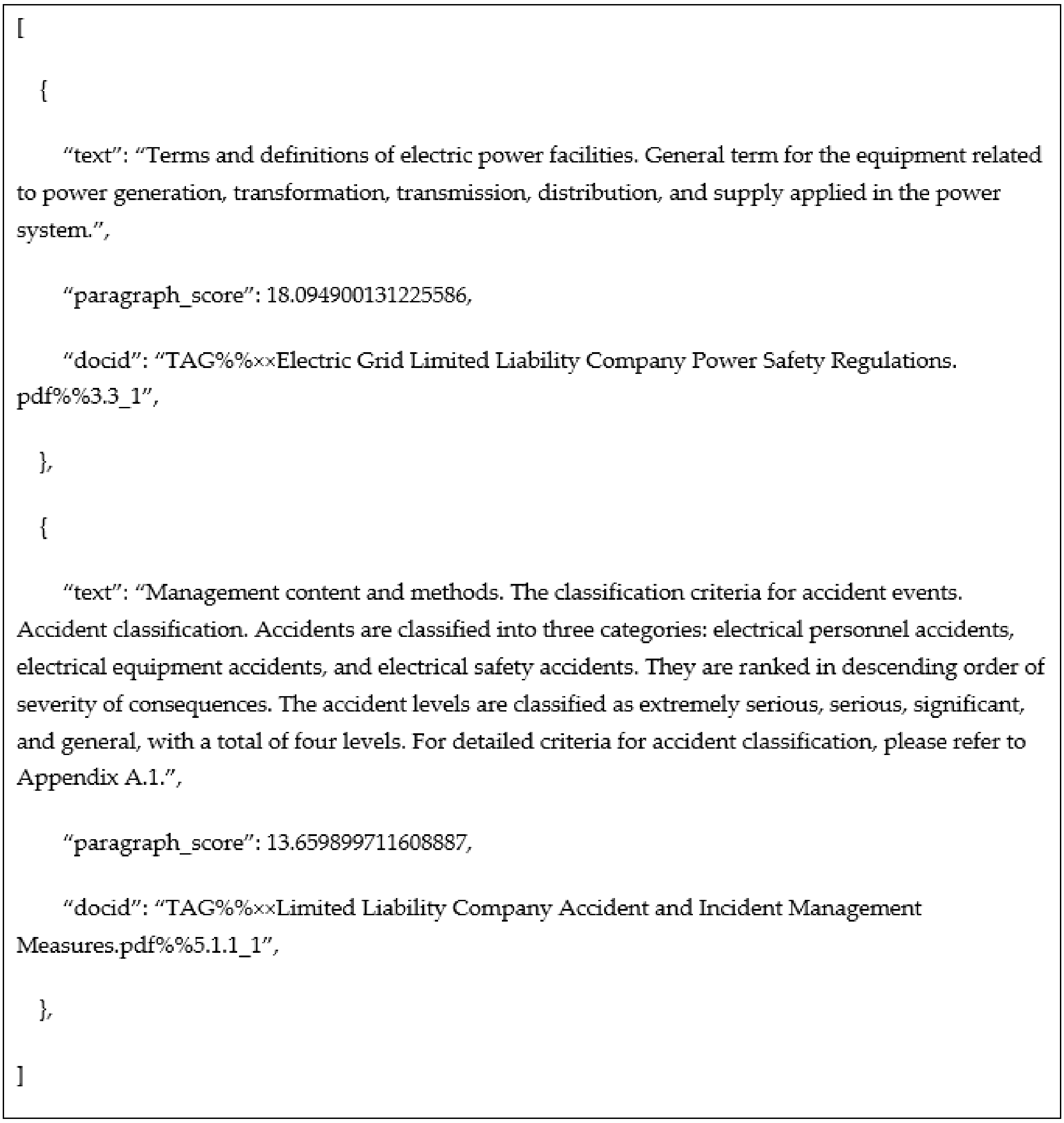
The parameter settings in this study are as follows. paragraph_threshold = 10, phrase_threshold = 0, remove_title = n (n = False, y = True), if remove_title = y, the paragraph titles will not be considered by the BERT reader algorithm during reading comprehension. max_answer_length = 50, mu = 0.6.

The parameter in the BM25 algorithm used in the Anserini module has a value range of (0–1), and the parameter has a value range of (0–3).

4.2. Document Preprocessing Performance

In accordance with the document pre-processing algorithm proposed, the document format output by Anserini is illustrated in Figure 6. Within this context, "text" denotes the output paragraphs obtained from Anserini, "paragraph_score" represents the specific score assigned to each paragraph, this is the S_{anserini} mentioned in STEP 4 in Section 3.

Finally, “docid” indicates the name of the document along with the corresponding section information where the paragraph is situated.



```
[

  {

    “text”: “Terms and definitions of electric power facilities. General term for the equipment related
to power generation, transformation, transmission, distribution, and supply applied in the power
system.”,

    “paragraph_score”: 18.094900131225586,

    “docid”: “TAG%%××Electric Grid Limited Liability Company Power Safety Regulations.
pdf%%3.3_1”,

  },

  {

    “text”: “Management content and methods. The classification criteria for accident events.
Accident classification. Accidents are classified into three categories: electrical personnel accidents,
electrical equipment accidents, and electrical safety accidents. They are ranked in descending order of
severity of consequences. The accident levels are classified as extremely serious, serious, significant,
and general, with a total of four levels. For detailed criteria for accident classification, please refer to
Appendix A.1.”,

    “paragraph_score”: 13.659899711608887,

    “docid”: “TAG%%××Limited Liability Company Accident and Incident Management
Measures.pdf%%5.1.1_1”,

  },

]
```

Figure 6. Document preprocessing of the Anserini module.

4.3. Question-Answering Performance

The comparison of the question-answering performance before and after the improvement of the BERTserini algorithm is presented in Table 1. It can be observed that the original BERTserini algorithm exhibits inaccuracies in extracting the start and end positions of answers when addressing power regulations and standards questions, and even results in incomplete sentences. Compared to the original BERTserini algorithm, the improved BERTserini algorithm proposed in this paper can accurately locate the paragraph containing the correct answer and perform precise answer extraction. Additionally, it removes specific details like paragraph headings during the answering process, adapting to the structural characteristics of professional domain regulatory texts. The answers to certain questions are more accurate and concise than manually generated standard answers.

Table 1. Comparison of question-answering performance before and after the improvement of the BERTserini algorithm.

Question	Standard Answer	Original BERTserini Algorithm			Improved BERTserini Algorithm		
		Answer	Whether Exact Match	Trace the Source and Results of the Answers	Answer	Whether Exact Match	Trace the Source and Results of the Answers
When should the audited units send the relevant documents and information to the professional audit teams?	The audited unit should send the relevant documents and information 5 working days before the audit.	The audited unit should send the relevant documents and information 5 working days before the audit.	Yes	No	5 working days before the audit.	Yes	Yes (× × Power Grid Co., LTD. Safety Production Risk Management System Audit Business Guide 5.1.5)
What is the key inspection content of quarterly safety production?	At the end of each quarter, the safety supervision department at all levels shall determine the key contents of supervision in the next quarter according to the annual work plan, seasonal characteristics, and key work arrangements.	Quarterly safety production focus supervision within.	No	No	At the end of each quarter, the safety supervision department at all levels shall determine the key contents of supervision in the next quarter according to the annual work plan, seasonal characteristics, and key work arrangements.	Yes	Yes (× × Power Grid Co., LTD. Safety Production Risk Management System Audit Business Guide 5.2.2)
What does correction mean?	Measures taken to eliminate nonconformities that have been identified.	Correction means the elimination of discrepancies that have been found.	No	No	Measures taken to eliminate nonconformities that have been identified.	Yes	Yes (× × Power Grid Co., LTD. Safety Zone Representative Management Service Guide 4.5)
What is the safety supervision department directly under the company responsible for?	To organize the preparation and issuance of the annual safety measure plan of the unit, supervise and evaluate the implementation of the safety measure plan of the unit and the power supply units at the county.	The safety supervision department of each unit directly under the company is responsible for organizing and compiling.	No	No	To organize the preparation and issuance of the annual safety measure plan of the unit, supervise and evaluate the implementation of the safety measure plan of the unit and the power supply units at the county.	Yes	Yes (× × Power Grid Co., LTD. Safety zone Representative Management Service Guide 5.1.2)
What is an electric utility?	A general term for equipment related to generation, transformation, transmission, distribution and supply used in power systems.	Utility applied to electricity.	No	No	A general term for equipment related to generation, transformation, transmission, distribution and supply used in power systems.	Yes	Yes (× × Power Grid Co., LTD. Power Safety Working Regulations 3.3)

4.4. Comparison of Different Algorithms

This paper uses the Exact Match rate (EM), Recall rate (R), and $F1$ score to measure the question-answering performances of different algorithms. Among them, EM represents the percentage of questions in the question-answering system where the answers provided are an exact match with the standard answers.

The specific calculation formula is as follows:

$$EM = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} I(\hat{y}_i = y_i) \tag{6}$$

where $n_{samples}$ represents the total number of samples. $I(x)$ is an indicator function that takes the value of 1 when $\hat{y}_i$ is identical to y_i, and 0 otherwise. As can be seen from the formula, a higher EM value indicates a higher exact match.

Recall rate is to determine the proportion between the number of questions accurately answered by the question-answering system and the total number of questions.

The calculation formula is as follows:

$$R = \frac{TP}{TP + FN} \tag{7}$$

where TP signifies the accurate count of questions answered correctly by the question-answering system. Conversely, FN denotes the incorrect count of questions that were responded to inaccurately by the system.

The calculation formula for the F1 score is as follows:

$$F1 = \frac{2(EM \times R)}{EM + R} \tag{8}$$

In this study, Algorithm 1, as shown in Table 2, used the original BERTserini algorithm, adopting the algorithmic steps from reference [17] to construct the intelligent question-answering system. In Algorithm 2, an additional pre-processing algorithm is incorporated based on Algorithm 1. Fine-tuning is conducted using the CMRC2018 dataset, and parameter optimization for the BERTserini algorithm is performed. Algorithm 3 is an extension of Algorithm 2, incorporating the SPTI dataset for fine-tuning. Algorithm 4 is an improved version of the Bertserini algorithm, which is based on Algorithm 3. In addition to incorporating the SPTI dataset for fine-tuning, a pre-processing FAQ module based on short-text similarity calculation is added to filter out frequently asked questions. This module enables more efficient and effective preprocessing of the questions.

Table 2. Comparison of different algorithms.

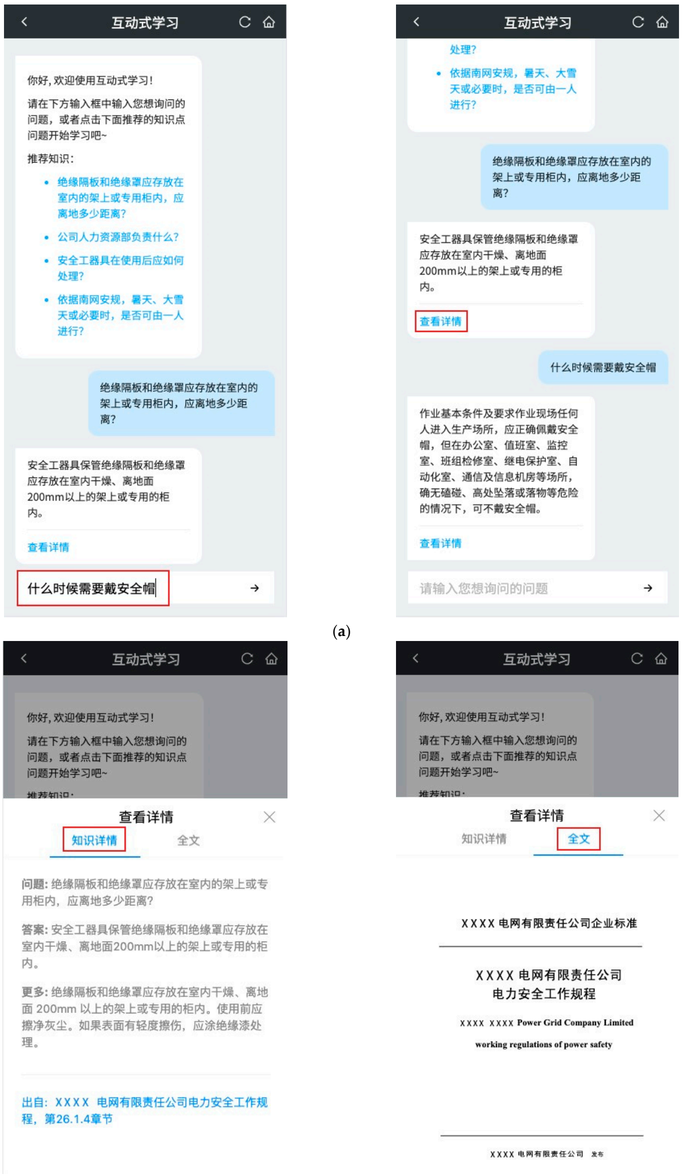
Algorithm	Content	EM		R		F1	
		Value	Percentage of Improvement	Value	Percentage of Improvement	Value	Percentage of Improvement
Algorithm 1	Original BERTserini	0.261	——	0.453	——	0.331	——
Algorithm 2	Document preprocessing + Original BERTserini +Fine-Tuning (CMRC2018) + Parameter tuning	0.502	48%	0.783	42.1%	0.615	46.1%
Algorithm 3	Document preprocessing + Original BERTserini + Fine-Tuning (CMRC2018) + Parameter tuning + Fine-Tuning (SPTI)	0.702	62.8%	0.919	50.7%	0.796	58.4%
Algorithm 4	Document preprocessing + Original BERTserini + Fine-Tuning (CMRC2018) + Parameter tuning + Fine-Tuning (SPTI) + FAQ	0.856	69.5%	0.976	53.6%	0.912	63.7%

Note: In the table "Percentage of improvement" is calculated based on the values of Algorithm 1 as a reference point. It represents the relative increase in evaluation value (EM, R, and F1) achieved by Algorithm 2, Algorithm 3, and Algorithm 4 compared to Algorithm 1.

As shown in Table 2, the EM value for Algorithm 1 is only 0.261, indicating poor performance. After adopting Algorithm 3, the EM value reaches 0.702, representing an improvement of 62.8%. After adopting the proposed Algorithm 4, the EM value reaches 0.856, demonstrating the best performance. In comparison to Algorithm 1, the proposed algorithm achieved an improvement of 69.5% in terms of EM value, 53.6% in terms of R value, and 63.7% in terms of F1 value. These results demonstrate a practical level of engineering advancement.

4.5. Engineering Application

An intelligent question-answering system for power regulations and standards is constructed based on the proposed improved BERTserini algorithm and experimental data presented in this paper, the UI interface of intelligent question-answering system based on improved BERTserini algorithm, as shown in Figure 7. The English explanation of UI interface in intelligent question-answering system based on improved BERTserini algorithm is shown in Figure 8.



互动式学习

你好, 欢迎使用互动式学习!

请在下方输入框中输入您想询问的问题，或者点击下面推荐的知识点问题开始学习吧~

推荐知识:

- 绝缘隔板和绝缘罩应存放在室内的架上或专用柜内，应离地多少距离?
- 公司人力资源部负责什么?
- 安全工器具在使用后应如何处理?
- 依据南网安规，暑天、大雪天或必要时，是否可由一人进行?

绝缘隔板和绝缘罩应存放在室内的架上或专用柜内，应离地多少距离?

安全工器具保管绝缘隔板和绝缘罩应存放在室内干燥、离地面200mm以上的架上或专用的柜内。

查看详情

什么时候需要戴安全帽

互动式学习

处理?

- 依据南网安规，暑天、大雪天或必要时，是否可由一人进行?

绝缘隔板和绝缘罩应存放在室内的架上或专用柜内，应离地多少距离?

安全工器具保管绝缘隔板和绝缘罩应存放在室内干燥、离地面200mm以上的架上或专用的柜内。

查看详情

什么时候需要戴安全帽

作业基本条件及要求作业现场任何人进入生产场所，应正确佩戴安全帽，但在办公室、值班室、监控室、班组检修室、继电保护室、自动化室、通信及信息机房等场所，确无磕碰、高处坠落或落物等危险的情况下，可不戴安全帽。

查看详情

请输入您想询问的问题

(a)

互动式学习

你好, 欢迎使用互动式学习!

请在下方输入框中输入您想询问的问题，或者点击下面推荐的知识点问题开始学习吧~

推荐知识:

查看详情

知识详情 全文

问题: 绝缘隔板和绝缘罩应存放在室内的架上或专用柜内，应离地多少距离?

答案: 安全工器具保管绝缘隔板和绝缘罩应存放在室内干燥、离地面200mm以上的架上或专用的柜内。

更多: 绝缘隔板和绝缘罩应存放在室内干燥、离地面 200mm 以上的架上或专用的柜内。使用前应擦净灰尘。如果表面有轻度擦伤，应涂绝缘漆处理。

出自: XXXX 电网有限责任公司电力安全工作规程，第26.1.4章节

(b)

互动式学习

你好, 欢迎使用互动式学习!

请在下方输入框中输入您想询问的问题，或者点击下面推荐的知识点问题开始学习吧~

推荐知识:

查看详情

知识详情 全文

XXXX 电网有限责任公司企业标准

XXXX 电网有限责任公司
电力安全工作规程

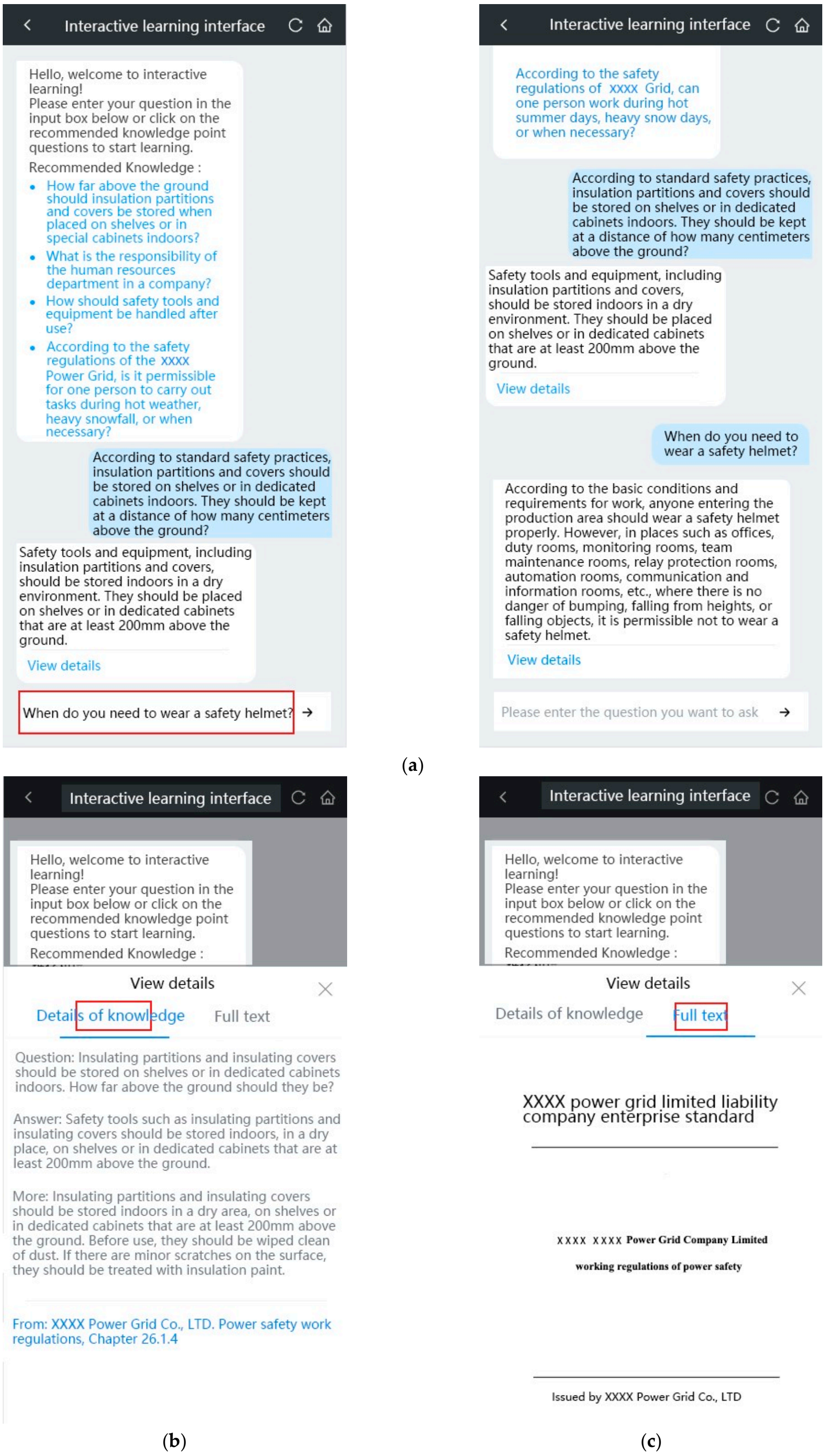
XXXX XXXX Power Grid Company Limited

working regulations of power safety

XXXX 电网有限责任公司 发布

(c)

Figure 7. UI interface of intelligent question-answering system based on improved BERTserini algorithm. (**a**) Multi-turn interactive question-answering interface. (**b**) Knowledge details page. (**c**) Full-text source page.



Interactive learning interface

Hello, welcome to interactive learning!
Please enter your question in the input box below or click on the recommended knowledge point questions to start learning.

Recommended Knowledge :

- How far above the ground should insulation partitions and covers be stored when placed on shelves or in special cabinets indoors?
- What is the responsibility of the human resources department in a company?
- How should safety tools and equipment be handled after use?
- According to the safety regulations of the XXXX Power Grid, is it permissible for one person to carry out tasks during hot weather, heavy snowfall, or when necessary?

According to standard safety practices, insulation partitions and covers should be stored on shelves or in dedicated cabinets indoors. They should be kept at a distance of how many centimeters above the ground?

Safety tools and equipment, including insulation partitions and covers, should be stored indoors in a dry environment. They should be placed on shelves or in dedicated cabinets that are at least 200mm above the ground.

View details

When do you need to wear a safety helmet?

Interactive learning interface

According to the safety regulations of XXXX Grid, can one person work during hot summer days, heavy snow days, or when necessary?

According to standard safety practices, insulation partitions and covers should be stored on shelves or in dedicated cabinets indoors. They should be kept at a distance of how many centimeters above the ground?

Safety tools and equipment, including insulation partitions and covers, should be stored indoors in a dry environment. They should be placed on shelves or in dedicated cabinets that are at least 200mm above the ground.

View details

When do you need to wear a safety helmet?

According to the basic conditions and requirements for work, anyone entering the production area should wear a safety helmet properly. However, in places such as offices, duty rooms, monitoring rooms, team maintenance rooms, relay protection rooms, automation rooms, communication and information rooms, etc., where there is no danger of bumping, falling from heights, or falling objects, it is permissible not to wear a safety helmet.

View details

Please enter the question you want to ask

(**a**)

Interactive learning interface

Hello, welcome to interactive learning!
Please enter your question in the input box below or click on the recommended knowledge point questions to start learning.
Recommended Knowledge :

View details

Details of knowledge | Full text

Question: Insulating partitions and insulating covers should be stored on shelves or in dedicated cabinets indoors. How far above the ground should they be?

Answer: Safety tools such as insulating partitions and insulating covers should be stored indoors, in a dry place, on shelves or in dedicated cabinets that are at least 200mm above the ground.

More: Insulating partitions and insulating covers should be stored indoors in a dry area, on shelves or in dedicated cabinets that are at least 200mm above the ground. Before use, they should be wiped clean of dust. If there are minor scratches on the surface, they should be treated with insulation paint.

From: XXXX Power Grid Co., LTD. Power safety work regulations, Chapter 26.1.4

Interactive learning interface

Hello, welcome to interactive learning!
Please enter your question in the input box below or click on the recommended knowledge point questions to start learning.
Recommended Knowledge :

View details

Details of knowledge | Full text

XXXX power grid limited liability company enterprise standard

XXXX XXXX Power Grid Company Limited

working regulations of power safety

Issued by XXXX Power Grid Co., LTD

(**b**) (**c**)

Figure 8. English explanation of UI interface in intelligent question-answering system based on improved BERTserini algorithm. (**a**) Multi-turn interactive question-answering interface. (**b**) Knowledge details page. (**c**) Full-text source page.

The system provides users with a multi-turn interactive question-answering interface on the topic of power safety, as illustrated in Figures 7a and 8a. Users can ask questions by either voice input or manual input. After sending the question, they will receive the system's response within 400 ms. Clicking on the "view details" link below the answer will cause the system to pop up a window displaying the source of the answer, including the name of the original document and the chapter number, as shown in Figures 7b and 8b. Clicking on the "full text" link allows users to view the content of the original document where the answer is located, as shown in Figures 7c and 8c.

5. Conclusions

The improved BERTserini algorithm proposed in this paper is designed for intelligent question-and-answer processing of power regulation documents. In comparison to the original BERTserini algorithm, this approach offers the following advantages:

(1) The improved BERTserini algorithm supports multi-document long text preprocessing for rules and regulations. This algorithm is capable of answering documents containing 30+ rules and regulations with a length of 30M+ bytes. This addresses the issue in the original BERTserini algorithm where document titles of regulatory documents were erroneously output as answers. Furthermore, it accurately provides the document name and chapter/page number information for answers that the original BERTserini algorithm could not identify. These enhancements significantly enhance the quality of answers and user experience in the question-answering system.

(2) The improved BERTserini algorithm proposed in this paper underwent two rounds of fine-tuning using the CMRC2018 and the specialized dataset SPTI. Algorithm parameters were also optimized. The intelligent question-answering system built upon it demonstrates a more precise answer generation capability compared to the original BERTserini algorithm when addressing domain-specific questions.

(3) The improved BERTserini algorithm proposed in this paper significantly enhances the exact match rate for intelligent question-answering in the domain of regulatory texts. Experimental data indicate that, compared to the original BERTserini algorithm, the exact match rate has increased by 69.5%, the R-value has improved by 53.6%, and the F1-value has risen by 63.7%. The algorithm maintains an average question–answer response time of within 400 milliseconds, meeting the requirements for engineering applications.

The improvements made to the BERTserini algorithm proposed in this paper are versatile, with the expectation that they can be widely applied in the research and construction of intelligent question-answering systems for regulatory texts across various industries. The limitations of this study lie in the current engineering practices, which are currently confined to the power industry. There is a lack of engineering cases for the construction of intelligent question-answering systems in industries such as petroleum, steel, transportation, and others where regulatory knowledge is prevalent. The generalizability of the algorithmic process across multiple domains needs further validation.

The next research direction involves applying this algorithm to construct intelligent question-answering systems for regulatory texts in other industry sectors. Additionally, by incorporating algorithmic iterations and leveraging advancements in technology, particularly with large language models, there is a continuous effort to optimize and enhance the effectiveness of the question-answering system.

Author Contributions: Writing—original draft preparation, M.G.; validation, M.G. and T.J.; formal analysis, M.L.; investigation, Q.W.; writing—review and editing, N.W. and G.L.; visualization, T.J.; supervision, M.G.; M.G. and M.L. conceived the idea and provided resources. N.W. and G.L. wrote the manuscript. T.J. and M.G. provided guidelines. M.G., M.L., T.J., N.W., G.L. and Q.W. designed the study and participated in the experiment. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: "Chinese Machine Reading Comprehension 2018 (CMRC2018)" at https://ymcui.com/cmrc2018 (accessed on 4 January 2018). "Delta Reading Comprehension Dataset (DRCD)" at https://github.com/DRCKnowledgeTeam/DRCD (accessed on 30 July 2018).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shum, H.Y.; He, X.; Li, D. From Eliza to XiaoIce: Challenges and Opportunities with Social Chatbots. *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 10–26. [CrossRef]
2. Xavier, A. Intelligence artificielle et psychiatrie: Noces d'or entre Eliza et Parry. *Linform. Psychiatr.* **2017**, *93*, 51–56.
3. Hb, B.G.; Kumar, M.A.; Kp, S. Vector Space Model as Cognitive Space for Text Classification. *arXiv* **2017**, arXiv:1708.06068. [CrossRef]
4. Ma, P.; Jiang, B.; Lu, Z.; Li, N.; Jiang, Z. Cybersecurity Named Entity Recognition Using Bidirectional Long Short-Term Memory with Conditional Random Fields. *Tsinghua Sci. Technol.* **2021**, *26*, 259–265. [CrossRef]
5. Mittal, A.; Agrawal, A.; Chouksey, A.; Shriwas, R. A Comparative Study of Chatbots and Humans. *Int. J. Adv. Res. Comput. Commun. Eng.* **2016**, *5*, 1055–1057.
6. Jabberwacky Website. Available online: http://www.jabberwacky.com (accessed on 30 August 2023).
7. Bohouta, G.; Kpuska, V.Z. Next-Generation of Virtual Personal Assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). In Proceedings of the IEEE CCWC 2018, the 8th IEEE Annual Computing and Communication Workshop and Conference, Las Vegas, NV, USA, 8–10 January 2018.
8. Poushneh, A. Impact of auditory sense on trust and brand affect through auditory social interaction and control. *J. Retail. Consum. Serv.* **2021**, *58*, 102281. [CrossRef]
9. Lei, X.; Tu, G.H.; Liu, A.X.; Ali, K.; Li, C.Y.; Xie, T. The Insecurity of Home Digital Voice Assistants—Amazon Alexa as a Case Study. *arXiv* **2017**, arXiv:1712.03327. [CrossRef]
10. Straga, D. Facebook Messenger as a Tool for Building Relationships with Customers: Bachelor Thesis. Ph.D. Thesis, Univerza v Ljubljani, Fakulteta za Družbene Vede, Ljubljana, Slovenia, 2017.
11. Berdasco, A.; López, G.; Diaz, I.; Quesada, L.; Guerrero, L.A. User Experience Comparison of Intelligent Personal Assistants: Alexa, Google Assistant, Siri and Cortana. *Proceedings* **2019**, *31*, 51. [CrossRef]
12. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
13. Dehouche, N. Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3): "The best time to act was yesterday. The next best time is now". *Ethics Sci. Environ. Polit.* **2021**, *21*, 17–23. [CrossRef]
14. Jain, N.; Jain, P.; Kayal, P.; Sahit, J.; Pachpande, S.; Choudhari, J. AgriBot: Agriculture-specific question answer system. *Preprint* **2019**. [CrossRef]
15. Gao, Z.; Feng, A.; Song, X.; Wu, X. Target-dependent sentiment classification with BERT. *IEEE Access* **2019**, *7*, 154290–154299. [CrossRef]
16. Yang, W.; Xie, Y.; Lin, A.; Li, X.; Tan, L.; Xiong, K.; Li, M.; Lin, J. End-to-End Open-Domain Question Answering with Bertserini. *arXiv* **2019**, arXiv:1902.01718.
17. Abo Khamis, M.; Ngo, H.Q.; Rudra, A. FAQ: Questions Asked Frequently. In Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, San Francisco, CA, USA, 26 June–1 July 2016. [CrossRef]
18. Kadhim, A.I. Term Weighting for Feature Extraction on Twitter: A Comparison between BM25 and TF-IDF. In Proceedings of the 2019 International Conference on Advanced Science and Engineering (ICOASE), Zakho, Iraq, 2–4 April 2019; pp. 124–128. [CrossRef]
19. Robertson, S.E.; Walker, S.; Jones, S.; Hancock-Beaulieu, M.M.; Gatford, M. Okapi at TREC-3. *Nist Spec. Publ. SP* **1995**, *109*, 109.
20. Yang, P.; Fang, H.; Lin, J. Anserini: Reproducible Ranking Baselines Using Lucene. *J. Data Inf. Qual. (JDIQ)* **2018**, *10*, 1–20. [CrossRef]
21. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Yang, Z. Pre-Training with Whole Word Masking for Chinese Bert. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3504–3514. [CrossRef]
22. Cui, Y.; Liu, T.; Che, W.; Xiao, L.; Chen, Z.; Ma, W.; Wang, S.; Hu, G. A span-extraction dataset for Chinese machine reading comprehension. *arXiv* **2018**, arXiv:1810.07366.
23. Shao, C.C.; Liu, T.; Lai, Y.; Tseng, Y.; Tsai, S. DRCD: A Chinese machine reading comprehension dataset. *arXiv* **2018**, arXiv:1806.00920.
24. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. Squad: 100,000+ questions for machine comprehension of text. *arXiv* **2016**, arXiv:1606.05250.
25. Bert Website. Available online: https://github.com/google-research/bert (accessed on 11 March 2020).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.