智能问答系统技术

Subjects: Computer Science, Artificial Intelligence
Contributor: Ming Gao, Mengshi Li, Tianyao Ji, Nanfang Wang, Guowu Lin, Qinghua Wu

Intelligent question answering system is an innovative information service system which 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.

Keywords: intelligent question-answering system ; information retrieval

1. Introduction

Intelligent question answering system is an innovative information service system which 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 get 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 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; Secondly, 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. [14] 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 improved BERTserini algorithm proposed in this paper has three main advantages.

(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 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 time for providing answers.

2. Background of the technology

2.1. FAQ

Frequently Asked Questions (FAQ) are a collection of common questions and answers designed to help users quickly find answers to their questions. The key is to establish a rich and accurate database of predefined questions, which consists of questions and their corresponding answers. They are manually curated from target documents. FAQs provide answers corresponding to user questions by matching them with the most similar questions.

2.2. BM25 algorithm

The Best Match 25 (BM25) algorithm, initially proposed by Stephen Robertson and his team in 1994, is applied in 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: first, the query statement involves tokenization to generate morphemes; then, it calculates the relevance score between each morpheme and the search results.

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

$$Score(D,Q) = \sum_i^n W_i \cdot R(q_i, D)$$

In this context, represents a query statement, represents a morpheme obtained from . For Chinese, the segmented results obtained from tokenizing query can be considered as morpheme . represents a search result document. represents the weight of morpheme , and represents the relevance score between morpheme and document . There are multiple calculation methods for weight parameter , with Inverse Document Frequency (IDF) being one of the commonly used approaches. The calculation process for IDF is as follows:

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

In the equation, represents the total number of documents in the index, and represents the number of documents that contain .

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

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

where and are adjustment factors, represents the frequency of morpheme appearing in document , denotes the length of document , and represents the average length of all documents.

2.3. Anserini

Anserini [18] 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.

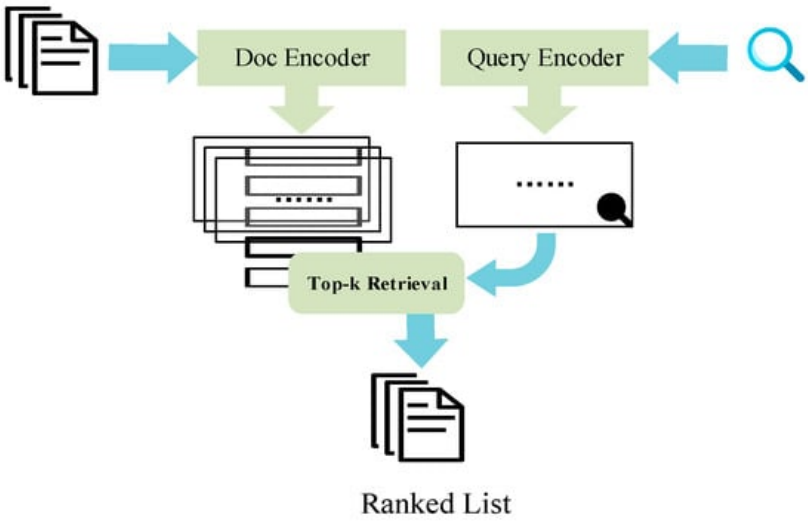

Figure 1. The flowchart of the Anserini algorithm.

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.

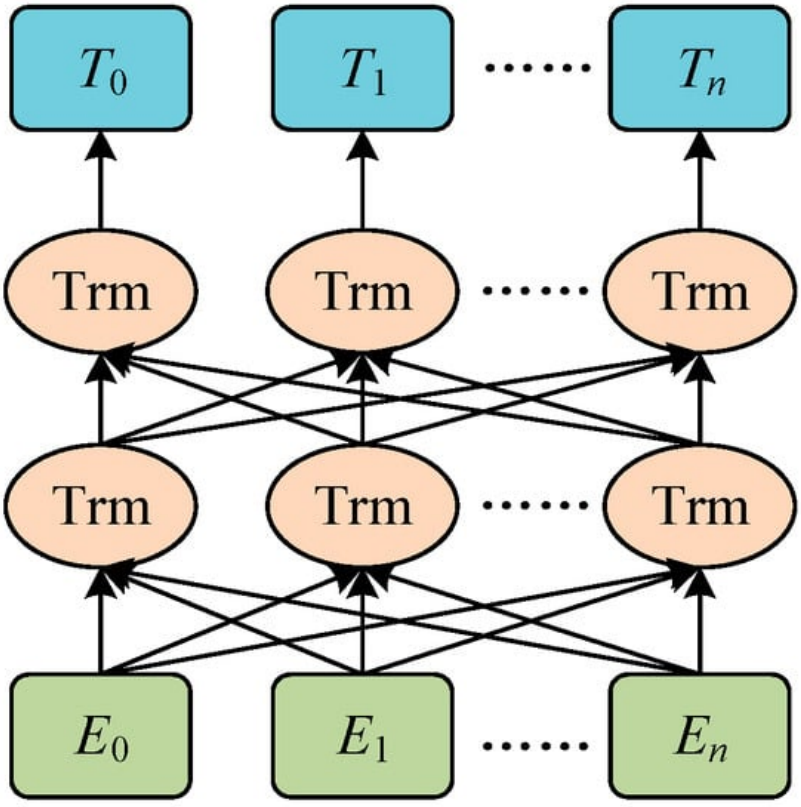

Figure 2. Architecture of BERT.

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**.

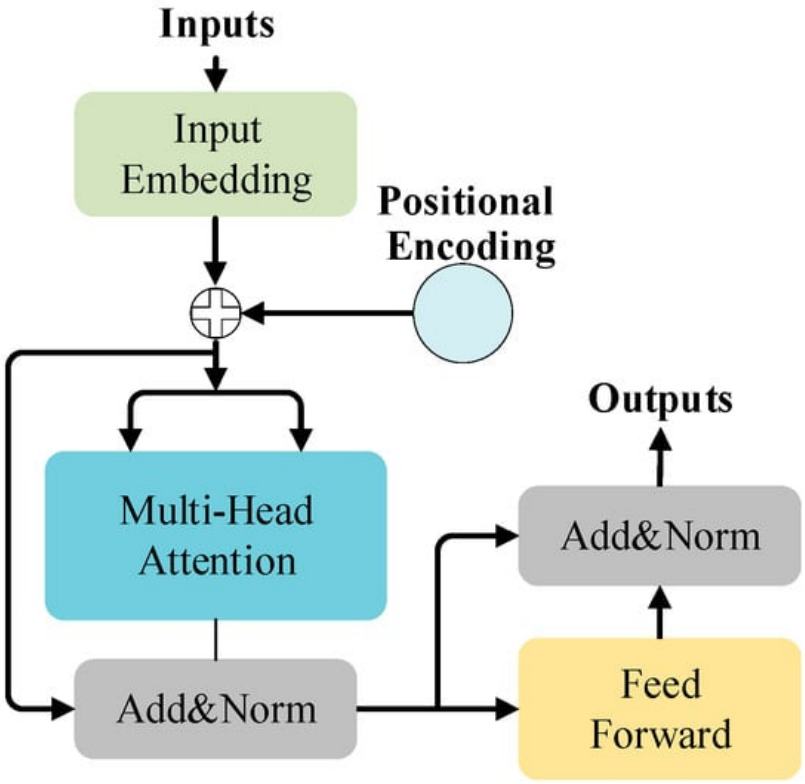

Figure 3. Transformer Encoder Principle.

In recent years, several Chinese BERT models have been proposed in the Chinese language domain. Among these, the chinese-BERT-wwm-ext model [19] released by the HIT·iFLYTEK Language Cognitive Computing Lab (HFL) has gained significant attention and serves as a representative example. This model was pre-trained on a corpus of Chinese encyclopedias, news articles, and question-and-answer texts, which contains a total of 5.4 billion words. The model uses the whole-word masking (wwm) strategy.

2.5. BERTserini algorithm

The architecture of BERTserini algorithm [14] is depicted in Figure 4. The algorithm employs the Anserini information extraction algorithm in conjunction with a pre-trained 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. 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.

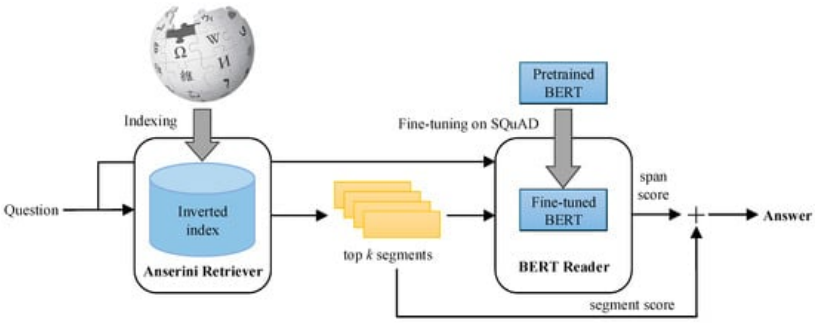

Figure 4. Architecture of BERTserini.