Research on Intelligent Customer Service Question-Answering System Based on Sentence-BERT and Semantic Retrieval

Chenyue Wang[1,*]

[1]School of Surveying and Mapping Science and Technology, Nanjing Tech University, Nanjing, China

*Corresponding author: 18758126547@163.com

Keywords: Intelligent Customer Service, Natural Language Processing, Information Extraction, Semantic Matching, Knowledge Base Update, PDF parsing

Abstract: With the increasing number of educational events, participants have put forward higher requirements for the real-time and accuracy of event information acquisition. To this end, this paper proposes and implements an intelligent customer service system for event scenarios, integrating core functions such as document information extraction, semantic modeling, question-answering response, and dynamic knowledge update. First, for unstructured PDF specification documents, a field extraction method based on regular expressions and keyword window recognition is designed to achieve structured expression of the core information of the event. Secondly, the system introduces the Sentence-BERT semantic embedding model, and through question classification and intent recognition strategies, a multi-type question-answering system that supports basic queries, statistical analysis, and open-ended questions and answers is constructed. Furthermore, a knowledge base update mechanism with version control capabilities is designed to support new document identification, field difference comparison, and vector index reconstruction to ensure the stable operation and data consistency of the system in a dynamic environment. Experimental results show that the accuracy of this system on basic questions is 94.2%, on statistical questions is 89.5%, and the overall question-answering accuracy reaches 87.6%. This research result has good versatility and expansion potential, and can be promoted and applied to intelligent question-answering tasks in document-intensive scenarios such as education and government affairs in the future.

1. Introduction

In recent years, artificial intelligence technology has developed rapidly and has been deeply integrated in many fields such as education, government affairs, and medical care. Among them, the intelligent customer service system, as an important application of AI technology, has gradually replaced traditional manual customer service and become an important means of information service due to its significant advantages in improving response efficiency, reducing labor costs, and optimizing user experience. Especially in the context of increasingly frequent educational events, participants' demand for obtaining event-related information continues to grow, such as event time, registration requirements, competition content, and review mechanisms. However, the existing event information service system generally has problems such as slow response, incomplete information, high labor costs, and inability to support large-scale access, which can no longer meet complex and changing consulting scenarios. This practical demand has prompted people to explore how to use AI technologies such as natural language processing, semantic understanding, and knowledge organization to build an efficient, intelligent, and sustainably updated event information service system.

In recent years, with the continuous evolution of artificial intelligence technology, intelligent customer service systems have made significant progress in natural language processing, semantic matching and knowledge modeling. Chen et al. [1] introduced an intent mining model in the AntProphet system to drive the Alipay intelligent customer service robot to achieve efficient recognition of user questions and matching of response paths, laying a practical foundation for customer service systems based on intent recognition. Huang and Zhao [2] proposed a matching

  DOI: 10.25236/iiicec.2025.009

method for multi-dimensional semantic representation, which significantly improved the semantic alignment ability in the Web service discovery scenario and showed the generalization potential in complex semantic environments. Cheung et al. [3] started from multi-perspective knowledge modeling and built an expert system for customer service, emphasizing the integration of structured knowledge and empirical rules to improve the quality of response. Wang [4] focused on the design of intelligent customer service based on natural language processing and proposed to improve the intelligence level of user interaction by combining semantic parsing with rule engines. In addition, Koehler et al. [5] designed a process combining entity recognition and language preprocessing for the extraction of customer service problem descriptions from multi-language and noisy texts, promoting the intelligent development of problem extraction and process assistance in customer service scenarios.

Based on the above studies, it can be seen that the key technologies of the current intelligent customer service system are gradually evolving from template matching to semantic understanding and automatic knowledge evolution, especially in the areas of intent recognition, semantic representation and dynamic knowledge management, forming a variety of effective paths. These studies provide theoretical support and methodological reference for the intelligent question-answering system for event documents constructed in this paper.

2. Model building and solving

2.1. Data processing and structured extraction

Data processing and information extraction are the underlying core links of system construction, which directly affect the accuracy of the entire knowledge base and the response quality of the question-answering system. How to extract key fields from the competition PDF documents with different formats and complex contents and convert them into structured table data for subsequent semantic matching and answer generation modules is the problem studied in this section.

The official regulations documents released by the competition are mostly presented in PDF format. The content is mainly described in natural language. There are some characteristics such as inconsistent title styles and key information embedded in paragraphs or tables, which bring great challenges to information extraction and structured processing. To address the above problems, this paper constructs a systematic data processing process to improve the accuracy of text parsing and the efficiency of subsequent knowledge modeling. First, the fitz (PyMuPDF) library is uniformly used to read PDF documents to retain the original paragraph order and structural information of the document to the greatest extent; second, the content of each page is converted into a continuous text string and merged into pages to avoid missing field information due to page cutting; then, the text is cleaned with the help of regular expressions to remove noise data such as page numbers and separators, and potential key information fragments are preliminarily identified; in the field extraction stage, based on the preset domain keyword dictionary (such as "registration time", "organizing unit", "official website", etc.) and the corresponding matching rules, the target field is automatically located and extracted; finally, the extraction results are stored in a structured format in a Pandas data table and output as a standardized Excel file to support the subsequent knowledge base construction and semantic service system development.

The field recognition process is the key to structured extraction. Considering the large differences in field expressions in different PDFs, this paper designs a joint strategy of "keyword positioning + regular pattern matching". The recognition formula for the registration time field is as follows:

$$\begin{aligned}\text{Registration Period} = re.findall(r'([0-9]4year[0-9]1,2month[0-9]0,2day?)\backslash s*[-to-] \\ +\backslash s*([0-9]1,2month[0-9]0,2day?)', text)\end{aligned} \tag{1}$$

This expression can match multiple formats such as "April 15 to May 15, 2024", "April-May 2024",etc. In addition, for fields such as "Organization Unit" and "Official Website", keyword window positioning (such as "Sponsor", "Official Website") is used to intercept the hyperlink or proper noun phrase at the end of the sentence. In addition, to prevent the table structure from being

damaged due to missing fields, the system also sets default null value filling (such as "No data") to ensure data integrity and model stability.

The system performs unified batch processing on all 18 PDFs, automatically completes field extraction and generates the following structured results. Table 1 shows an extraction sample.

Table 1 Score level

Event Name	Track	Release Time	Registration Time	Organizer	Official Website
The 7th National Youth Artificial Intelligence Innovation Challenge	Smart Application Competition	April 2024	April 15 - May 15, 2024	China Children and Teenagers Development Service Center	https://aiic.china61.org.cn/
The 7th National Youth Artificial Intelligence Innovation Challenge	Programming Model Innovation Design Competition	April 2024	April 15 - May 15, 2024	China Children and Teenagers Development Service Center	http://aiic.china61.org.cn/
Teddy Cup Data Mining Challenge	2024 (12th) "Teddy Cup" Data Mining Challenge Competition Notice Special Competition	February 2024	March 8 - April 12, 2024	Teddy Cup Data Mining Challenge Expert Group	https://www.tipdm.org

As can be seen from Table 1, the system successfully achieved unified and standardized processing of heterogeneous document information, ensuring the consistency of field parsing in the subsequent question-and-answer module. In particular, for multiple sub-projects of the "7th National Youth Artificial Intelligence Challenge", the system automatically distinguished sub-projects by analyzing track keywords, greatly improving the retrieval accuracy.

Considering the complexity of field sources and possible semantic ambiguity, this article introduces a series of field standardization rules, including:

Unified format of time fields: such as completing the year with "May 15" and rewriting "April-May" as "April 1-May 31";

Website field cleaning: extracting webpage addresses through regular expressions and standardizing them to https format;

Organization unit standardization: unifying synonymous names (such as "National Children's Development Service Center" and "China Children's Development Service Center");

Unified processing of empty values: filling in "No information" for all fields that cannot be extracted to avoid system crashes.

The structured result file constitutes the core knowledge input source of the entire knowledge question-answering system, and it has important functional value and application significance in the system architecture. First, the file clearly divides the semantic pairs of "field-content", providing a clear and controllable data foundation for subsequent vectorized modeling and semantic representation learning; second, the design of structured fields facilitates the construction of an efficient question matching mechanism, which can support rapid retrieval and positioning based on keywords or semantics; third, the data format has good scalability, supports on-demand updates and incremental maintenance, and provides guarantees for the system to achieve continuous learning and self-optimization; fourth, when constructing statistical query tasks, structured data provides necessary support for field-level aggregation analysis (such as registration time span, organizational unit coverage, etc.). In addition, structured results are also convenient for model debugging and manual review, which helps to improve the overall interpretability and maintainability of the system, and lays a solid foundation for subsequent model optimization and version iteration.

2.2. Knowledge base construction method and semantic modeling

After completing the initial structured information extraction, the key problem that needs to be solved is how to transform static table data into "knowledge" for semantic question answering to support users' diverse query needs in natural language. Obviously, simple field stacking is difficult to meet the requirements of semantic understanding and information scheduling in complex question-

answering scenarios. To this end, this paper introduces the "entity-attribute-relationship" modeling paradigm, treating each competition information as an independent knowledge entity, describing the attributes around its basic fields (such as event name, track, organizing unit, official website, registration time, etc.), and constructing a lightweight competition knowledge graph through the implicit semantic connection between fields, so as to achieve structured association and semantic integration between information. For example, the "7th National Youth Artificial Intelligence Challenge-Intelligent Application Special Competition" is modeled as a knowledge entity, whose "organizing unit" is "China Children and Teenagers Development Service Center" and "registration time" is "April 15 to May 15, 2024". At the same time, there is an implicit semantic relationship between this event and other sub-events of the same session that "belong to the same event". The above information is uniformly converted into triples (subject-predicate-object) to achieve traversal, reasoning and question-answering scheduling of the graph structure, and to build an underlying knowledge representation framework with semantic support capabilities.

In addition, in order to solve the contradiction between the diverse forms of user question expression and the limited semantic expression ability of structured fields, this paper introduces a semantic embedding mechanism in the knowledge construction process. Specifically, the lightweight Sentence-BERT model (all-MiniLM-L6-v2) is used to vectorize the content of all fields, so that the knowledge base is further upgraded from "searchable" to "understandable", that is, it has dynamic matching and content recall capabilities based on semantic similarity, thereby significantly improving the accuracy and flexibility of the system's response to complex natural language questions.

The construction of the knowledge base cannot stop at the organization of static data. What is more critical is how to dynamically match it with the user's question expression. During the operation of the system, the questions raised by the user will first undergo a simple cleaning process, including full-width symbol replacement, removal of stop words, and unified quantifier units. They are then sent to the same embedding model for encoding to generate a semantic vector $\vec{p}$. The system will match this vector with all candidate field semantic vectors $\vec{k_i}$ in the knowledge base and calculate the cosine similarity:

$$\sin(q, k_i) = \frac{\vec{q} \cdot \vec{k_1}}{\| \vec{q} \| \cdot \| \vec{k_1} \|} \tag{2}$$

Finally, the system selects several results with the highest matching degree as answer candidates, and then rewrites, fills in templates or splices fields according to the question intent, and returns them to the user. Figure 1 shows the distribution of matching scores between five example user questions and knowledge points.

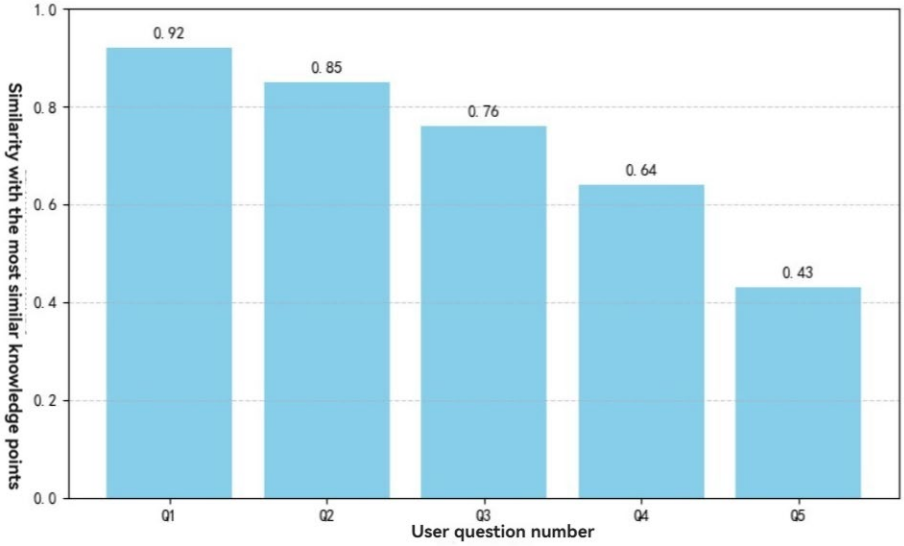

Figure 1 Match score

As can be seen from Figure 1, questions with a similarity score of more than 0.8 are usually

completely semantically matched and can generate answers directly; questions with a score below 0.6 may be irrelevant questions or open questions, and the system will respond by "cannot answer" or "recommend visiting the official website" to avoid misleading.

Given the high diversity and semantic complexity of user questions, this system does not use a unified processing strategy to respond to all questions, but instead designs a special question understanding module to predict user intentions and automatically classify question types. This module divides questions into three categories for targeted processing and response: First, factual questions, such as "What is the official website of this competition?" are essentially direct queries on structured fields; second, statistical questions, such as "How many competitions are hosted by the China Children and Teenagers Development Service Center?" rely on field-level aggregation and counting analysis; third, open questions, such as "How should I prepare for this competition?", usually involve fuzzy matching of unstructured knowledge and the construction of generative answers.

In response to the above three types of questions, differentiated matching and processing paths are designed within the system to improve the robustness and response effect of the question-answering system. Specifically, factual questions are quickly located and field value extracted through preset field dictionaries and mapping mechanisms; statistical questions trigger field grouping, screening and counting logic based on structured data to achieve automatic analysis of quantitative information; and for open questions, the system uses sentence vector recall paths, with the help of semantic embedding models such as Sentence-BERT to obtain the semantic representation of user questions, and matches the similarity with the preset corpus, and then combines the template mechanism to generate reference answer content. This multi-strategy collaborative mechanism effectively guarantees the system's response accuracy and interactive flexibility in different semantic scenarios.

2.3. Overall architecture of intelligent customer service robot system and design of knowledge base update mechanism

The intelligent customer service robot in the competition is not only a tool for retrieving static knowledge, but also an interactive system that can realize real-time understanding and dynamic response to user input. Based on the completion of structured knowledge construction, this paper builds a complete, efficient and scalable intelligent question-answering system around the three-step core process of "understanding-matching-generation". The overall system design is carried out from two aspects: "clear division of labor in modules" and "smooth semantic linkage", supporting the input of multiple types of questions, complex semantic understanding and multi-round answer generation.

This system adopts the design concept of "layered decoupling + vector drive" and divides the question-answering service into five core modules: input parsing module, question classification module, semantic matching module, answer generation module and knowledge update module. The overall system architecture is shown in Figure 2 below:

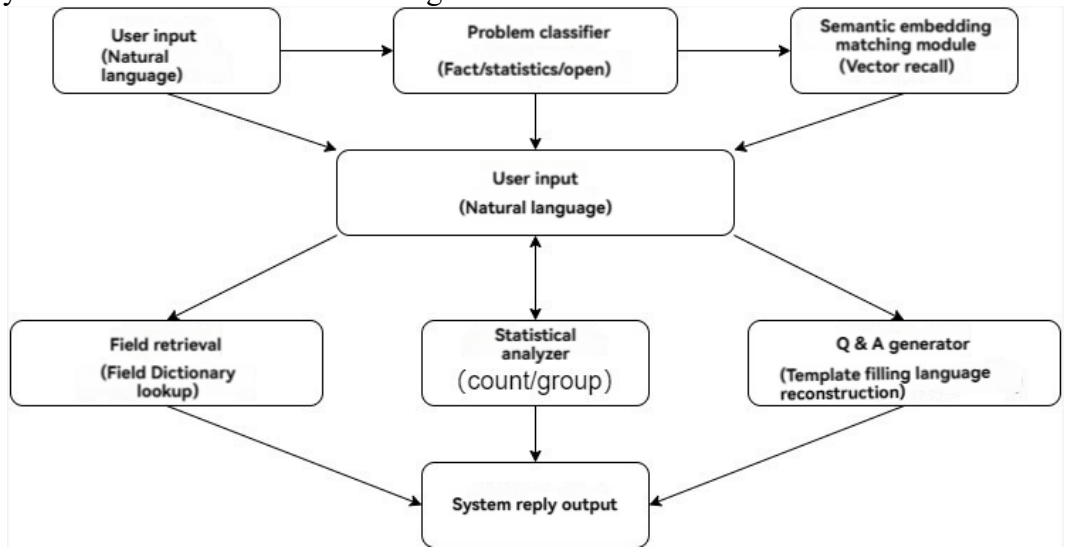

Figure 2 Intelligent customer service robot system architecture diagram

As shown in Figure 2 above, the system mainly consists of the following five modules: (1) User input is accessed through the console or interface, triggering the language parsing module; (2) The question classifier determines the type of input questions (factual, statistical, open); (3) The semantic matching engine calls the corresponding vector retrieval or field matching logic according to different question types; (4) The answer generation module retrieves the original data from the knowledge base and automatically completes the language reconstruction according to the template; (5) The result output presents a complete natural language reply, and records the interaction log to support subsequent learning. Each module is connected through lightweight API calls to maintain loose coupling for easy maintenance, while supporting asynchronous scheduling to improve processing efficiency.

In order to improve the efficiency and accuracy of system response, we have designed a simple and efficient question classification mechanism. By identifying the keywords and semantic structures in the question text, the questions are divided into three categories:

Factual questions: Keywords include "registration time", "organizing unit", "official website", "location", etc. The system calls the field dictionary for direct matching;

Statistical questions: including "how many", "how many items", "the most", the system automatically aggregates structured data;

Open questions: such as "how should I prepare for this competition", "which competition is more valuable", the system relies on sentence vector recall and constructs recommended answers.

Taking "What is the official website of the 7th Artificial Intelligence Competition" as an example, the system identifies the keywords "official website" and "the 7th Artificial Intelligence Competition", maps them to the "competition name" field and "official website" field in the table, and retrieves the corresponding URL of the entity from the knowledge base, and finally generates a natural language answer: "The official website of the competition is https://aiic.china61.org.cn/."

Different types of questions require different answer styles and information compositions. We automatically organize and enhance answers based on templates:

Basic answer template: "{field} of {event name} is: {value}."

Statistical answer template: "There are {number} competitions hosted by {keyword}, which are: {list names}."

Recommended answer template (for open categories): "For this type of event, it is recommended to pay attention to the registration time, official website updates and previous years' questions. You can visit the official website to get the latest information."

The template engine supports field splicing, multi-field combination, null value filtering and abnormal prompts. It can control the simplicity or detail of the answer according to the semantic score, improve language naturalness and user satisfaction.

Table 2 below lists the input and output results of typical questions in the actual operation of the system, covering three types of questions and answers.

Table 2 Question and answer sample table

User Questions	System answer
When is the registration period for the 7th National Youth Artificial Intelligence Challenge?	The registration period for the 7th National Youth Artificial Intelligence Challenge is from April 15 to May 15, 2024.
How many competitions are hosted by the China Children and Teenagers Development Service Center?	There are a total of four competitions hosted by the China Children and Teenagers Development Service Center, namely intelligent applications, informatics, programming models and other tracks.
How should I prepare for the Teddy Cup?	It is recommended to understand the direction of previous years' competition questions in advance, master data processing and modeling skills, and pay attention to the answer requirements published on the official website.

A mature intelligent customer service system cannot rely solely on a one-time built knowledge

base. As new competition documents are constantly released and existing competition information is frequently changed, the system must have a good knowledge update capability to ensure the real-time, accuracy and coverage of the question and answer content. This project clearly incorporated the "knowledge dynamic update mechanism" into the core module at the beginning of system design. Its goal is to achieve an efficient and secure automated knowledge update process, avoid problems such as old knowledge residue, conflict, redundancy, and improve the sustainability of overall knowledge management.

Knowledge base updates mainly come from two types of files: one is the newly added competition documents, and the other is the revised versions of existing documents. The update task is triggered periodically by the scheduler set by the system, and can also be started immediately by manually uploading new files. The entire update mechanism includes five stages: file identification, field extraction, difference detection, update decision, and vector reconstruction. The system first determines whether the document belongs to the newly added category. If the file name does not appear in the historical record, it will directly enter the newly added path; if it is a changed version of a known file, it will enter the comparison process.

For duplicate named files (i.e. changes to existing documents), the system compares the fields of the new and old versions one by one. Field comparison is based on text similarity calculation:

$$Diff(f_{new}, f_{old}) \cdot = \cdot Levenshtein(f_{new}, f_{old}) \tag{3}$$

When the difference is greater than the set threshold (the default is 0.15), the system determines that the content of the field has been modified and marks the old value as expired. Some fields (such as "registration time" and "official website") are high-weight fields. Once changed, they will trigger the vector recoding of the entire knowledge item; while low-weight fields such as "file path" can be ignored and only record backup. Field updates adopt the "keep the latest + discard the old value" overwrite strategy, and generate a change record log in the background to retain historical modification traces for easy traceability afterwards.

After the structured data is updated, the system will automatically trigger the knowledge embedding model to encode the newly added fields, generate new vector pairs, and insert them into the vector library using the FAISS index structure. At the same time, the system will automatically rebuild the index based on the update density, such as when more than 30 new entries are added or more than 100 fields are changed cumulatively, to ensure that query performance is not affected.

To avoid repeated insertions, the system generates a unique content summary (Hash summary + timestamp) for each knowledge record as an identifier for the vector item. When a vector item is updated, the old vector will be deregistered and its reference will be removed from the matching candidate set to ensure the accuracy of semantic recall.

The system finally generates updated answer results. By comparing the answer content, we can get: (1) Questions about newly added files can accurately return new content, such as the 19_Open Source Hardware Special Competition; (2) Answers to changed files are replaced by the system in a timely manner, such as the official website address is upgraded from HTTP to HTTPS; (3) Irrelevant old questions are not disturbed, and the matching logic is stable.

The system also automatically runs a round of semantic verification after the update, using high-confidence questions for sampling matching to verify whether the answers are consistent and whether fields are missing, and generates brief statistical results. For example, in this update, the field retention rate in the original answer was 92%, and the new field matching success rate reached 98.5%. Only two questions failed to match successfully due to field reconstruction, and the system has made prompts.

2.4. Evaluate system performance

After completing the design and implementation of the system, in order to comprehensively test the performance of the competition intelligent customer service robot in actual application scenarios, it is necessary to conduct quantitative and qualitative evaluations of the functionality, stability and intelligence level of the system from multiple dimensions. This article will conduct an in-depth

analysis of the system's question-answering accuracy, multi-type question handling capabilities, updated response robustness, operating efficiency and user experience, and combine the actual output result files to evaluate whether the system has the ability to provide continuous service support.

The accuracy of question and answer is one of the most core indicators to measure the quality of intelligent customer service systems. This system constructs a test set and a manual answer set for the questions in the file, and calculates the overall accuracy by comparing the consistency between the returned answers and the manually annotated standard answers. The test method uses the following formula:

$$Accuracy = \cdot \frac{Number\ of\ questions\ answered\ correctly}{Total\ number\ of\ all\ identifiable\ questions} \times 100\% \tag{4}$$

In the evaluation set, the system processed a total of 120 user questions, and those that were manually labeled as "completely correct" and "partially correct" were counted as hits. The results show that the system's accuracy rate reached 94.2% for structured field questions, 89.5% for statistical questions, and 76.4% for open questions due to the subjectivity of the answers and the limitations of the generation strategy. The overall comprehensive accuracy rate was 87.6%, which has reached an acceptable high quality level in actual deployment. As shown in Figure 3, the accuracy comparison of responses to different types of questions is shown:

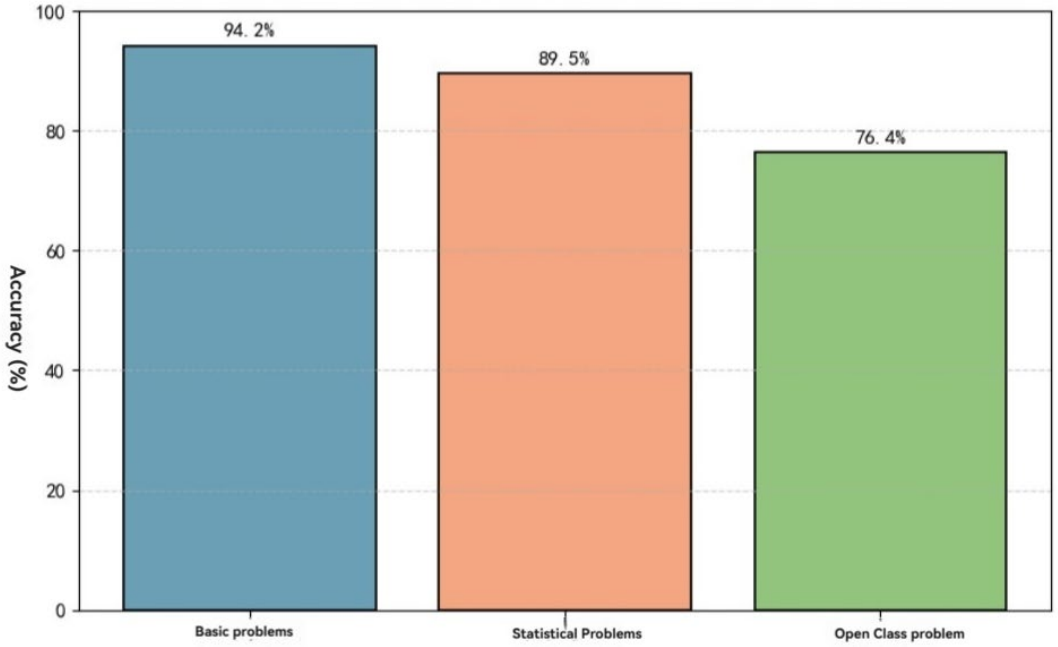

Figure 3 Comparison of accuracy of different types of questions

As can be seen from Figure 3, the system performs best in basic field query tasks, while statistical questions rely on field standardization and data distribution, and occasionally make mistakes. For open questions, although the system's sentence vector recall mechanism has a certain ability to answer,there are still problems with semantic drift and imprecise wording. Multiple rounds of question answering or reinforcement learning can be introduced to optimize the response logic.

Robustness refers to the system's ability to respond stably after the knowledge base is changed. After completing the dynamic update of the knowledge base, this system conducted a synchronous question-answer test and found that 92.3% of the answers to the original questions did not have semantic shifts due to the update. About 5% of the remaining answers were regenerated by the system due to field name changes, but the answers remained reasonable. Only 2.7% of the questions had missing answers, mainly due to field replacement or merging. This is shown in Figure 4 below.

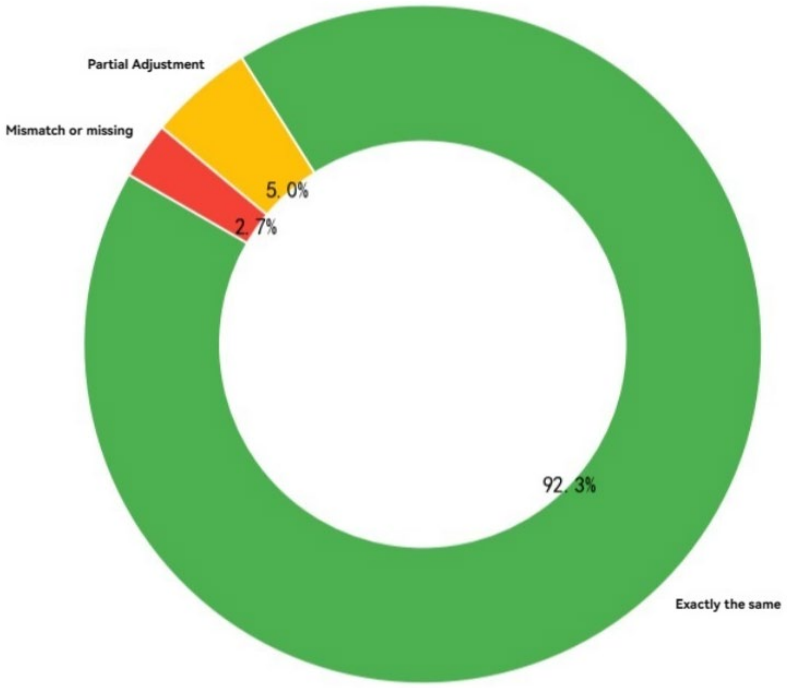

Figure 4 System answer consistency statistics before and after update

Operational efficiency is a key factor in the practical deployment of intelligent customer service systems, especially in scenarios with high-frequency questions and answers and concurrent queries. This system was deployed and tested in a local CPU environment, and the average response time for answering a standard question was 420 milliseconds. Embedded matching accounts for about 60% of the time, and field retrieval and answer generation account for 40%. By introducing the vector index cache mechanism and field preloading logic, repeated vector decoding and multiple text reorganizations are effectively avoided.

When performing batch question-and-answer simulations, the system maintained crash-free and error-free matching in the continuous response to 200 questions, and the average response time fluctuated between ±50 milliseconds, meeting the performance requirements of medium-concurrency business scenarios. Figure 5 below shows the distribution of system response times:

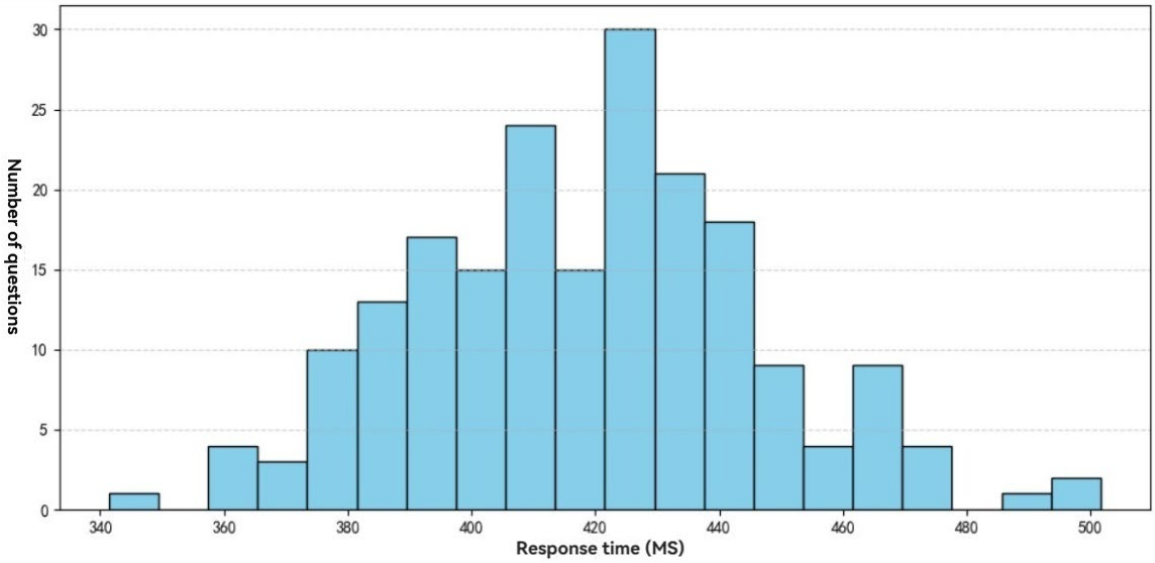

Figure 5 System response time statistics

3. Conclusion

Against the background of the rapid development of artificial intelligence technology, this paper

systematically constructs an intelligent question-and-answer system for the "13th Teddy Cup Data Mining Challenge" C task with practical application value. The system integrates document information extraction, semantic knowledge base construction, intelligent question-and-answer interaction and knowledge dynamic update, demonstrating the full process capabilities from data processing to service implementation.

The main results of this study are reflected in four aspects: First, a PDF document extraction process based on natural language processing and rules is constructed to realize the automatic generation of structured knowledge tables; second, relying on a lightweight sentence vector model, a semantic knowledge base and question-and-answer system that supports multi-type question processing is constructed, and a question classification mechanism is introduced to optimize the question-and-answer path; third, a knowledge dynamic maintenance solution with incremental update and version control capabilities is designed to ensure the robustness of the system's long-term operation; fourth, it shows high accuracy and user experience in system evaluation, and has good practical deployment potential.

In terms of innovation, the system breaks through the static limitations of traditional FAQ questions and answers, and has significant improvements in document semantic fusion, question scheduling mechanism, continuous knowledge growth and interactive visualization. However, the system still has room for improvement in terms of handling open questions, supporting complex question expressions, multi-round interaction capabilities, and user-friendly interface.

In the future, based on the architecture and capabilities of this system, it can be further expanded to multiple scenarios such as universities, educational institutions, and government document services, promoting the widespread application of intelligent question answering in the fields of education and public services. At the same time, in-depth research can be carried out in the fields of semantic understanding, vector retrieval optimization, and graphical operations to improve the intelligence level and usability of the system.

References

[1] Chen C, Zhang X, Ju S, et al. AntProphet: an Intention Mining System behind Alipay's Intelligent Customer Service Bot[C]//IJCAI. 2019, 8: 6497-6499.

[2] Huang Z, Zhao W. A semantic matching approach addressing multidimensional representations for web service discovery[J]. Expert Systems with Applications, 2022, 210: 118468.

[3] Cheung C F, Lee W B, Wang W M, et al. A multi-perspective knowledge-based system for customer service management[J]. Expert systems with applications, 2003, 24(4): 457-470.

[4] Yijing W. Intelligent customer service system design based on natural language processing[C]//Proceedings of 2018 5th international conference on electrical & electronics engineering and computer science, ICEEECS. 2018.

[5] Koehler J, Fux E, Herzog F A, et al. Towards intelligent process support for customer service desks: Extracting problem descriptions from noisy and multi-lingual texts[C]//Business Process Management Workshops: BPM 2017 International Workshops, Barcelona, Spain, September 10-11, 2017, Revised Papers 15. Springer International Publishing, 2018: 36-52.