Ministry of Education of the Republic of Belarus
Educational Institution
Belarusian State University
of Informatics and Radioelectronics

UDC 004.912

Zhang Caigui

METHOD OF SEMANTIC BLOCK DEFINITION IN TEXT

ABSTRACT
for a master's degree

Speciality 7-06 0612 03 Information management systems

Academic Supervisor
German Yuliya Olegovna
Candidate of Technical Sciences,
Associate Professor

Minsk 2025

INTRODUCTION

With the rapid expansion of digital content in various fields, from academia to business, the need for efficient information retrieval systems has become increasingly pressing. Document-based question answering (QA) systems serve as a vital tool in this context, allowing users to obtain precise answers from extensive collections of unstructured text. As the volume of information grows, traditional search methods often fall short, failing to deliver contextually relevant responses. This highlights the necessity for more advanced systems that can comprehend and interpret natural language queries against a backdrop of complex documents.

The development of a document-based QA system involves several critical processes: document processing, text vectorization, question matching, and answer extraction. Document processing ensures that information is structured and accessible, while text vectorization converts textual data into numerical formats that machine learning models can understand. Question matching aligns user queries with pertinent sections of the document corpus, and answer extraction identifies and presents the most relevant responses. Each component plays a crucial role in enhancing the system's overall effectiveness. The similar approach is described in [1-A].

Despite significant progress in the field, existing document-based QA systems face several limitations. Many systems struggle with understanding nuanced language, leading to inaccurate or irrelevant answers. Additionally, the integration of context and user intent often remains superficial, resulting in a lack of depth in responses. Furthermore, current solutions may not adequately handle diverse document formats and varying information types, restricting their applicability across different domains.

To address these challenges, our aims to enhance the accuracy and relevance of responses by incorporating advanced natural language understanding techniques and improved contextual analysis. By leveraging state-of-the-art machine learning algorithms and innovative document processing methods, we aspire to create a more robust and user-friendly solution. The purpose of this master's thesis is to develop a method of semantic block definition in text, thereby improving the accuracy of document-based question answering systems.

GENERAL CHARACTERISTICS OF THE WORK

The aim of this research is to develop a local document-based Q&A system that efficiently retrieves answers from text documents. This research is relevant due to the growing need for rapid information retrieval in various fields such as education, business and scientific research, which require rapid processing of large amounts of text data. The object of the study is the accuracy and relevance of answers derived from specific text documents. The research subjects include natural language processing techniques, semantic similarity metrics and information retrieval methods.

For the purpose of this research, it is assumed that the following tasks will be performed:

– Analyzing existing methods and approaches. This stage includes reviewing relevant scientific literature and existing technical solutions in the field of Q&A systems and text retrieval.

– Development of retrieval algorithms. The task at this stage is to improve the relevance of answers using semantic similarity metrics and keyword extraction techniques.

– Implementing the algorithm. This involves not only developing the algorithm, but also ensuring that it is implemented correctly in the form of executable code.

– Evaluation and optimization. The performance of the system will be evaluated and optimized to improve accuracy and response time.

All of the above tasks are aimed at achieving the main goal of this research –creating an effective local document-based Q&A system that accurately retrieves information from text.

Provisions for defense:

1. The combined approach using semantic similarity metrics alongside with keyword extraction methods significantly enhances the relevance of retrieved answers, ensuring they are aligned with user queries.

2. The implementation of a context-aware retrieval mechanism allows to improve answer accuracy and efficiency.

The topic of the dissertation work corresponds to the list of priority areas determined by the Decree of the President of the Republic of Belarus dated 07.05.2020 (No. 156) “On priority areas of scientific and technical activity in the Republic of Belarus for 2021-2025” (direction of development of the information society, electronic state and digital economy).

The main provisions and results of the research were discussed at various conferences:

Proceedings of the 60th scientific conference of graduate students, undergraduates and students, BSUIR, Minsk – 2024;

Information Technologies and Systems 2024 (ITS 2024): Proceedings of the international scientific conference, Minsk, November 20, 2024, BSUIR;

Internauka: Electronic Scientific Journal. Part 1. – 2025. – № 13(377).

SUMMARY OF THE WORK

The focus of this work is to research and enhance a document-based Q&A system to ensure high accuracy and relevance in retrieving answers from text. This research proposes modifications to address the limitations of traditional retrieval methods and conducts a detailed analysis of the performance of the improved system.

The document retrieval process utilizes advanced semantic similarity techniques to better assess the relevance of answers to user queries. The initial retrieval results are refined using a combination of keyword extraction and vector embeddings, which allows the system to optimize the matching of questions to relevant documents. This approach significantly enhances the accuracy of the answers provided.

To further improve the system’s efficiency, a context-aware mechanism is implemented. This mechanism dynamically adjusts the retrieval strategy based on user interactions, thus addressing the challenges of contextual relevance and information retrieval from large datasets.

Experiments were conducted to evaluate the system's performance using various datasets, including randomly generated questions and actual user queries from different domains. The evaluation considered multiple scenarios, such as varying query complexity and dataset sizes, and demonstrated good results in terms of accuracy, response time, and user satisfaction.

Overall, this research contributes to the development of a powerful document-based Q&A system that effectively meets the demands of real-time information retrieval.

CONCLUSION

This thesis explores in detail the implementation of a Q&A system based on semantic similarity and keyword matching, emphasizing the innovativeness and advantages of the system over traditional approaches.

First, the Q&A system utilizes an advanced deep learning model, SentenceTransformer, to achieve efficient semantic understanding. Compared with traditional keyword-based retrieval methods, semantic matching is able to capture the underlying intent of user queries more accurately. This enables the system to not only handle literal matches, but also understand more complex semantic relationships to provide more relevant and accurate answers.

Second, the document processing flow has been carefully designed using a strategy of splitting documents into multiple chunks. This approach improves processing efficiency while ensuring that the system can respond to user queries in a shorter time. The maximum length limit of each block effectively reduces the computational complexity, enabling the system to quickly retrieve the most relevant information.

In addition, keyword matching, as a complementary mechanism, further enhances the robustness of the system. When the semantic similarity is low, effective keyword matching can still provide valuable information. This dual mechanism combines the advantages of semantic understanding and keyword retrieval, enabling the system to remain efficient and accurate in multiple situations.

Finally, the modular design of the system makes it easy to understand and maintain with good scalability. This design concept not only enhances the usability of the system, but also provides users with a more intuitive interaction experience.

In conclusion, the Q&A system described in this thesis demonstrates significant innovation in the field of information retrieval, overcoming the limitations of traditional methods by combining semantic understanding and keyword matching to provide more efficient and accurate answering capabilities. The realization of this system provides a new perspective on the development of Q&A technology and demonstrates the great potential of modern natural language processing in practical applications.

LIST OF PUBLISHED WORKS

1-A. Caigui, Zh. One local language model / German, Y. O., German, O. V., Nasr, S. N., Hengrui, Z., & Caigui, Zh. // Internauka: Electronic Scientific Journal. Part 1. – 2025. – № 13(377). – P. 30–35.

2-A. Zhang Caigui. How to extract meaning from a phrase / Zhang Caigui, Yu. German // Information Technologies and Systems 2024 (ITS 2024) = Information Technologies and Systems 2024 (ITS 2024): Proc. of the international scientific conference, Minsk, November 20, 2024 / BSUIR, Minsk – 2024. – P. 191–192.

3-A. Zhang Caigui. Fingerprint recognition / Zhang Caigui, Ouyang Shiyun // Information technology and management: Proc. of the 58th scientific conference of graduate students, undergraduates and students, Minsk, April 18-22, 2022 / BSUIR, Minsk – 2022. – P. 32-33.