Building a Question-Answering System

Silma Khan

New York City College of Technology

CST 4802 Information Retrieval

Professor Yuksel

FALL 2024

Table of Contents

1. Abstract

In this project, I have built a question-answering (QA) system leveraging both lexical search methods (BM25) and semantic search techniques (SentenceTransformer models). The system uses Wikipedia as its corpus to answer user queries about topics specifically in the areas of artificial intelligence, natural language processing, and information retrieval. This report outlines the methodology, tools, and algorithms used to construct the QA system, evaluates its performance using precision, recall, and F1-score, and proposes improvements based on the results. The project serves as a hands-on application of information retrieval concepts and a demonstration of the potential and limitations of Question and Answering systems.

2. Introduction

The field of information retrieval focuses on the efficient retrieval of relevant information in response to one or multiple users queries. Modern Question and Answering systems often combine lexical search methods, such as BM25, with semantic search methods, which use machine learning models to understand meaning behind the users' queries. The primary objective of this project is to design and implement a Question and Answering system capable of answering user queries with high relevance and accuracy using a corpus derived from Wikipedia.

This report details the design and development of the Question and Answering system. We will discuss the outline for the preprocessing steps for corpus preparation, the indexing process using BM25, the semantic refinement process using SentenceTransformer models, and the evaluation metrics used to measure performance. Finally, we will discuss the results and propose enhancements for future iterations.

3. Methodology

In this methodology section, we will describe and discuss the step-by-step process used to build the Question and Answering system. It begins with data collection, where Wikipedia pages were selected as the primary source of information, but only a subset with certain topics. Following this, the data underwent preprocessing to ensure it was clean and ready for indexing. The methodology also explains the implementation of BM25, a lexical retrieval algorithm, to create a baseline for retrieving relevant sentences based on keywords. Semantic search, powered by SentenceTransformers, refined these results by evaluating the contextual meaning of sentences. The section concludes with details on how user queries were handled and how answers were generated by combining the strengths of lexical and semantic search methods.

3.1 Data Collection

Wikipedia was chosen as the corpus for this project due to its comprehensive and structured information. The wikipedia-api library was used because it allows for efficient access to Wikipedia content, ensuring flexibility and compliance with Wikipedia’s usage policies. The topics "Information Retrieval," "Natural Language Processing," and "Artificial Intelligence" were selected due to their relevance to the use of this project along with what we have learned in class. These topics served as a foundation to test the Question and Answering system's performance. However, the system is designed to handle any topic, making it adaptable to various domains with the addition of more Wikipedia pages or external datasets.

3.2 Text Preprocessing

Text preprocessing is a critical step in preparing raw text data for information retrieval systems. It involves cleaning and structuring the data to ensure consistent and meaningful results. For this project, preprocessing began with sentence tokenization, which divides the text into individual sentences using NLTK's sent_tokenize function. This granularity allows the system to retrieve specific sentences rather than large blocks of text. Additionally, text cleaning was applied to remove non-ASCII characters and excessive whitespace while preserving punctuation. This ensured that the processed data retained its semantic integrity, which is crucial for subsequent lexical and semantic analysis and allowing for human understanding of the output of the text.

- **Sentence Tokenization:** Using NLTK's sent_tokenize function to split text into sentences.
- **Text Cleaning:** Removing non-ASCII characters and unnecessary whitespace while preserving punctuation and structure to maintain context.

3.3 Lexical Search with BM25

BM25, short for Best Match 25, is a ranking algorithm widely used in information retrieval systems. It operates on the principles of term frequency and inverse document frequency (TF-IDF) but incorporates probabilistic models to improve retrieval precision. BM25 scores sentences based on the presence and frequency of query terms based on a given users query, with parameters *k1* and *b* controlling key aspects of its behavior. Specifically, *k1* adjusts the term frequency saturation, determining how much additional weight frequent terms should receive, while *b* regulates document length normalization, accounting for the length of the sentence to avoid bias towards longer texts. In this project, BM25 was employed as a baseline retrieval method, identifying sentences with high lexical similarity to the user query. By tuning its parameters to *k1=1.5* and *b=0.75*, the model balanced the importance of term frequency and sentence length, ensuring relevant candidate sentences were retrieved efficiently. The values *k1=1.5* and *b=0.75* were chosen after considering standard recommendations and balancing performance. A higher or lower *k1* might overemphasize or underweight frequent terms, while an inappropriate *b* value could lead to biases in sentence ranking based on their length. These values ensured optimal retrieval efficiency and relevance for this specific corpus use of my chosen topics and query types used in this project. BM25 was employed as a baseline retrieval method, identifying sentences with high similarity to the user query. By tuning its parameters to these specific values, the model effectively balanced the importance of term frequency and sentence length, ensuring relevant candidate sentences were retrieved efficiently.

3.4 Semantic Search with SentenceTransformers

Semantic search enhances the traditional lexical retrieval by understanding the contextual meaning of words. SentenceTransformers are machine learning models that generate dense vector representations (embeddings) of sentences, capturing their semantic essence or meaning according to the word. In this project, the *all-MiniLM-L6-v2* model was utilized due to its efficiency and strong performance in semantic similarity tasks. By computing embeddings for both the query and candidate sentences, the system measured their cosine similarity to refine the ranking provided by BM25. This approach ensured that the final answers were not only lexically similar but also contextually relevant given the different queries for testing.

3.5 Query Handling and Answering Generation

User queries, specifically the pre-set ones that I have set, were processed through the hybrid system that follows:

1. **Initial Retrieval:** BM25 retrieved the top 10 candidate sentences based on lexical similarity to the query.
2. **Semantic Refinement:** SentenceTransformer ranked the candidates by semantic similarity to the query.
3. **Final Output:** The top 3 most relevant sentences were returned as the answers. This multi-step approach leveraged the speed of lexical search and the depth of semantic understanding to deliver precise and meaningful results.

4. Evaluation

The evaluation of the Question and Answering system was conducted using a test dataset containing three queries along with their expected answers. These queries included "What is natural language processing?", "What is information retrieval?", and "What is artificial intelligence?" To measure the performance of the system, we used precision, recall, and F1-score as evaluation metrics. Precision represented the proportion of retrieved sentences that were relevant, while recall quantified the proportion of relevant sentences that were retrieved. The F1-score, as the harmonic mean of precision and recall, provided a single measure to assess the overall performance of the system.

The results revealed that the system achieved a precision of 0.1111 (11.11%), indicating that only a small fraction of the retrieved sentences were relevant. The recall was slightly higher at 0.1667 (16.67%), demonstrating that the system retrieved a limited portion of the relevant sentences available in the corpus. Consequently, the F1-score stood at 0.1333 (13.33%), reflecting the challenges faced in balancing relevance and coverage. These results highlighted the limitations of the current approach, including its struggle to filter irrelevant sentences and capture all relevant ones.

5. Findings and Challenges

The results of this project highlighted several important findings. BM25 demonstrated its effectiveness in retrieving sentences with high keyword overlap, making it a strong baseline for lexical search. However, its inability to capture semantic meaning resulted in irrelevant results for queries that required deeper contextual understanding. The integration of SentenceTransformer models significantly enhanced the relevance of retrieved answers, showcasing the value of semantic similarity in question-answering systems. The approach successfully combined the strengths of lexical and semantic retrieval, but its effectiveness was constrained by the quality of BM25's candidate sentences and the coverage of the corpus.

Several challenges emerged during the development and evaluation of the system. The limited size of the corpus, which focused only on three topics, reduced the availability of relevant answers, impacting recall. Queries with abstract or multi-faceted meanings posed additional difficulties, as they required a nuanced understanding that neither BM25 nor SentenceTransformer could fully address. Threshold selection for semantic similarity also proved to be a critical factor in balancing precision and recall, requiring further experimentation to optimize.

6. Proposed Improvements

The contents of this section outlines and discusses various enhancements to address the limitations identified in the current QA system. It focuses on expanding the corpus to provide a more diverse set of data, utilizing advanced models for improved semantic understanding, and fine-tuning the BM25 algorithm for better lexical retrieval. Additionally, strategies to increase the size of the candidate pool and refine the test dataset are discussed, offering a roadmap to improve precision, recall, and overall performance. These improvements aim to make the system more robust, adaptable, and accurate in addressing user queries.

6.1 Expand Corpus

Incorporating more Wikipedia pages and external datasets can improve corpus diversity and increase the likelihood of retrieving relevant answers. A more diverse corpus allows the system to address a wider range of user queries effectively. For instance, expanding the dataset to include topics outside the current scope, such as historical events, scientific breakthroughs, or the pop culture events, could significantly enhance the system's adaptability. Additionally, leveraging external datasets like open-access academic papers or curated question-answer pairs can provide structured and domain-specific knowledge, filling gaps where Wikipedia content may be limited. Such expansion will help balance relevance and recall while ensuring the system's utility in diverse applications.

6.2 Advanced Models

Using more advanced SentenceTransformer models, such as multi-qa-mpnet-base-dot-v1, could enhance semantic similarity calculations. These models are designed to provide superior embeddings, enabling a deeper understanding of complex semantic relationships. By utilizing such models, the Questioning and Answering system could improve its ability to discern subtle contextual differences between sentences, resulting in more accurate and meaningful answers to user queries. The use of models trained on domain specific data could further refine performance for specialized applications, such as legal or medical question answering.

6.3 BM25 Tuning

Experimenting with BM25 parameters and integrating synonyms or stemming can improve lexical retrieval. Adjusting k1 and b parameters further allows for fine-tuning the balance between term frequency importance and document length normalization. Additionally, incorporating stemming ensures that variations of a word are treated as equals, while a synonym dictionary can capture related terms that BM25 may otherwise overlook. These enhancements can address the system's current gaps in recall and relevance, particularly for queries with diverse vocabulary usage.

6.4 Larger Candidate Pool

Increasing the number of candidate sentences retrieved by BM25 before semantic refinement can improve recall. The current approach limits BM25 to providing the top 10 candidates, which may exclude relevant sentences that rank slightly lower. Expanding this pool to 20 or more candidates would give the SentenceTransformer model a broader base to refine, increasing the likelihood of retrieving relevant answers. While this may introduce more noise, the semantic model's ranking mechanism is expected to mitigate such issues effectively.

6.5 Improved Test Dataset

Expanding the test dataset with more queries and detailed expected answers can provide a better evaluation framework. A robust dataset should include diverse query types, such as factual, explanatory, and multi-faceted questions, to test the system's adaptability. Each query should be accompanied by a set of expected answers that capture varying valid responses. By simulating real-world usage scenarios, this expanded test dataset can reveal additional performance bottlenecks and guide future enhancements to both lexical and semantic components

7. Conclusion

This project demonstrated the construction and evaluation of a Questioning and Answering system using BM25 and SentenceTransformer models. The results highlighted the potential of combining lexical and semantic retrieval methods to address complex user queries. While the system showed promise, the evaluation metrics, particularly the low precision and recall, revealed significant room for improvement. These shortcomings emphasized the challenges of balancing relevance and coverage in information retrieval.

Future work will focus on several key areas. First, expanding the corpus to include a broader range of topics and integrating external datasets will enhance the system's adaptability and relevance. Second, employing more advanced models, such as domain-specific SentenceTransformers, is expected to refine semantic similarity calculations further. Additionally, refining the BM25 parameters and experimenting with larger candidate pools will address gaps in recall and precision. Finally, an improved evaluation framework, featuring a diverse and comprehensive test dataset, will provide more accurate insights into system performance and guide iterative improvements.

By addressing these areas, the QA system can evolve into a more robust and effective tool, capable of handling a wider variety of queries with greater accuracy and relevance..

References

1. "Wikipedia API Documentation." *PyPI*, Python Software Foundation, https://pypi.org/project/wikipedia-api/.
2. "NLTK Documentation." *Natural Language Toolkit*, NLTK Project, https://www.nltk.org/.
3. "Rank-BM25 Documentation." *PyPI*, Python Software Foundation, https://pypi.org/project/rank-bm25/.
4. "SentenceTransformers Documentation." *SBERT.net*, UKP Lab, https://www.sbert.net/.
5. Trotman, Andrew, et al. "Improvements to BM25 and Language Models Examined." *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2014.
6. Reimers, Nils, and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." *arXiv preprint arXiv:1908.10084*, 2019.
7. Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
8. Vaswani, Ashish, et al. "Attention Is All You Need." *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
9. Mikolov, Tomas, et al. "Efficient Estimation of Word Representations in Vector Space." *arXiv preprint arXiv:1301.3781*, 2013.
10. Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint arXiv:1810.04805*, 2019.