Hybrid Q&A System with BM25, MiniLM, and FLAN-T5

Ahmed Mohamed Omar Rashad Omar El Hakeem
Omar Bahaa

Abstract

This report presents the development of a hybrid question answering (QA) system that combines classical lexical search with modern semantic embedding techniques and a generative language model for natural language answers. Specifically, we integrate BM25 for keyword-based retrieval, MiniLM embeddings for dense semantic similarity, and Reciprocal Rank Fusion (RRF) for hybrid ranking. For answer generation, we employ the FLAN-T5 model. The system was evaluated on a curated subset of StackExchange posts from the Artificial Intelligence and Data Science domains. Our experiments show that semantic search substantially outperforms lexical search alone, and the hybrid fusion achieves the highest accuracy in terms of MAP@10 and NDCG@10. This work is part of our final project for the Advanced Deep Learning and Generative AI course.

1 Introduction

Question answering (QA) is a central task in modern NLP applications. Traditional information retrieval systems rely heavily on keyword overlap (e.g., BM25), which fails when queries and documents use different terminology. On the other hand, semantic search models use dense vector representations to capture contextual similarity, but they may miss exact keyword matches or domain-specific terminology.

To address this, we build a hybrid QA pipeline combining both approaches. In addition to document retrieval, we generate fluent, factual answers using a fine-tuned generative model (FLAN-T5) which is grounded in the retrieved documents. The resulting pipeline allows users to input arbitrary natural language questions and receive relevant, accurate answers supported by citations.

Our system integrates IR and generative AI in a modular, extensible pipeline. This report documents the architecture, implementation, evaluation, and key findings from our experiments.

2 System Overview

The system consists of the following pipeline stages:

1. **Data ingestion and preprocessing:** Clean and structure the StackExchange dataset.
2. **Document indexing:** Build indices for BM25 and MiniLM embeddings.
3. **Query processing:** Accept user input and retrieve relevant documents.
4. **Fusion ranking:** Merge lexical and semantic results using RRF.
5. **Answer generation:** Use FLAN-T5 to produce a grounded natural language answer.

3 Dataset and Preprocessing

We collected 2,000 QA pairs: 1,000 from AI and 1,000 from Data Science domains on StackExchange. Each question was paired with its accepted answer. The final dataset contains the following fields:

```
id | title | body | answer | link | domain
```

We concatenated the title and body fields to form the full query context. All text was lowercased and stripped of HTML tags. Stopword removal and tokenization were applied for the BM25 index, while semantic search used raw text.

4 Retrieval Approaches

4.1 BM25 Search

BM25 is a probabilistic IR model widely used in search engines. It scores documents based on term frequency, inverse document frequency, and document length normalization.

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

We used the `rank-bm25` Python library with standard parameters: $k_1 = 1.5$ and $b = 0.75$.

4.2 Semantic Search with MiniLM

To capture deeper contextual similarity, we embedded documents and queries using `all-MiniLM-L6-v2` from SentenceTransformers. For each query, cosine similarity was computed against all document vectors.

$$\text{sim}(Q, D) = \frac{Q \cdot D}{||Q|| \cdot ||D||} \tag{2}$$

Embeddings were cached and vectorized using NumPy for efficient scoring.

4.3 Hybrid Search with RRF

We applied Reciprocal Rank Fusion (RRF) to combine BM25 and MiniLM rankings:

$$\text{RRF}(d) = \sum_{r \in R} \frac{1}{k + r(d)} \tag{3}$$

Where $r(d)$ is the rank of document d in the result list R, and k is a smoothing constant (set to 60). This method boosts documents ranked highly in both systems and helps balance lexical precision with semantic generalization.

5 Answer Generation

For answer generation, we fine-tuned a FLAN-T5 model on a few in-domain examples using HuggingFace Transformers. The model prompt format is:

```
Use the following passages to answer the question.
[1] Text from Doc 1...
[2] Text from Doc 2...
Question: What is deep learning?
Answer:
```

The output is a coherent, referenced answer using citations like "[1]".

This design ensures the language model does not hallucinate facts outside the given context, aligning it with RAG (Retrieval-Augmented Generation) principles.

6 Evaluation Setup

We created a manual set of 10 diverse queries with known relevant answers. Evaluation metrics included:

- MAP@10 (Mean Average Precision)
- MRR@10 (Mean Reciprocal Rank)
- NDCG@10 (Normalized Discounted Cumulative Gain)

We used the `ranx` library to compute metrics over all retrieval variants.

6.1 Results

```
BM25:
MAP@10  = 0.272
MRR@10  = 0.272
NDCG@10 = 0.399
```

```
Semantic:
MAP@10  = 0.733
MRR@10  = 0.733
NDCG@10 = 0.797

Hybrid (RRF):
MAP@10  = 0.750
MRR@10  = 0.750
NDCG@10 = 0.812
```

The hybrid approach consistently outperformed standalone methods. Semantic embeddings provide substantial gains in understanding user intent.

7 Example Queries and Outputs

Query 1: What is deep learning?

Retrieved Docs: 7 relevant posts on neural nets, AI definition

FLAN-T5 Answer:

"Deep learning is a subfield of machine learning based on artificial neural networks that enables computers to learn hierarchical features from data. It is used in tasks like image classification and NLP. [1][2]"

Query 2: Difference between AI and data science

Answer:

"AI focuses on intelligent systems that can simulate human thinking. Data science focuses on extracting insights from data using statistics and machine learning. [2][4]"

8 Limitations and Future Work

While our system performs well, it has some limitations:

- Limited document set (only 2K posts)
- FLAN-T5 is relatively small; larger models like Mixtral or GPT-4 could improve generation
- No real-time retrieval latency benchmarks were collected
- Citations are inline but not clickable in the current UI

Future work may include:

- Multilingual question answering
- Cross-domain generalization (e.g., medical, legal)
- Real-time performance testing and caching

9 Conclusion

This project demonstrates a hybrid IR-GEN system combining classic search, semantic understanding, and grounded text generation. Our findings indicate that semantic methods dramatically improve relevance, and combining them with lexical methods yields optimal results. The FLAN-T5 model effectively generates answers grounded in retrieved evidence, showing potential for real-world educational or assistant use cases.