Project Checkpoint-2

AIT-526 Sec-001

Team 14
Saivarun Tanjore Raghavendra
Mohammed Tareq Sajjad Ali
Suraj Poldas
Mano Harsha Sappa

1. Project Progress Demonstration

- **Objective**: The project focuses on creating a Q&A system tuned to answer technical questions related to construction techniques from the Infotunnel FHWA website, and more broadly, it aims to handle complex questions across multiple NDE (Non-Destructive Evaluation) techniques.
- **Current Progress**:
  - Developed a foundational Q&A system using Wikipedia as the initial data source, implementing key steps such as tokenization, preprocessing, query formulation, and question classification.
  - Integrated data scraping using BeautifulSoup to collect information, with document indexing and ranking to improve retrieval.
  - Successfully detects question types and fetches answers using Wikipedia's API, although limited to general knowledge.
- **Next Steps**: Transition the data source from Wikipedia to the Infotunnel FHWA database to better address construction-specific queries. Additionally, explore transformers or create a custom model using BERT to enhance passage filtering and ranking.

2. System Architecture/Framework

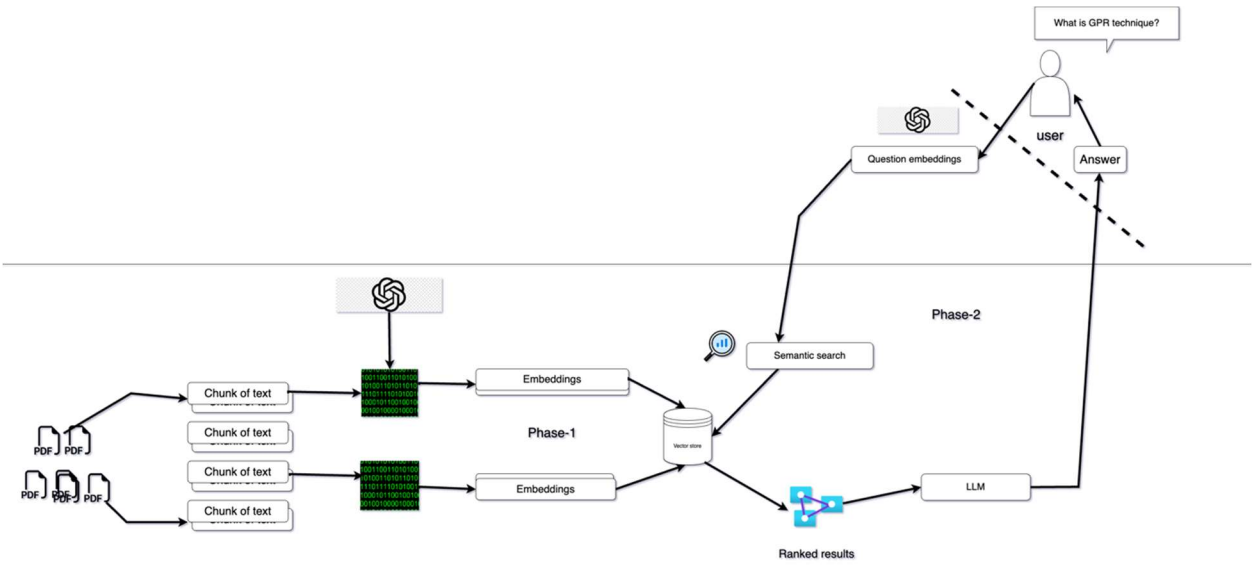

Fig: 1 Framework of Q&A System

Phase 1: Embedding Creation

- **Input Documents**: this process begins with numerous PDFs. First, these are separated into smaller "chunks of text" so that it processes and searches through the document.

- **Embedding Generation**: Each chunk of text then passes through an embedding model that converts the text into a vector representation-e.g., embeddings of its semantic meaning. This is likely to be done via a pre-trained language model.

- **Vector store**: The created embeddings are then persisted to a vector database, aka a vector store, that will facilitate efficient semantic searching. This database will allow retrieval by similarity of embedding and not simple keyword matching.

Phase 2: Question-Answering

- **Query:** This could be a user-generated question, for example, "What is GPR technique?"

- **Question Embedding:** This question is then processed to generate a question embedding using the same embedding model or one that is compatible such that the embedding would align in vector space with the document embeddings.

- **Semantic Search**: The question embedding is compared with the document embeddings kept in a vector store. This would go ahead and execute a semantic search on the vector store to return chunks of text most similar to the question, actually retrieving the most relevant information.

- **Ranked Results**: These results are ranked according to the relevance of the output with the user's question.

- **LLM Response Generation**: Ranked results are then fed into a large language model for the generation of a coherent, complete answer. An LLM uses context from the retrieved chunks of texts to formulate a response that answers a user's question directly.

- **Display Answer**: This is the final step, where the answer is generated and displayed to the user.

- **Framework Overview**: The system uses a modular approach, incorporating NLP and reasoning techniques with plans to combine traditional information retrieval with advanced methods:
    - **User Interface Module**: Accepts user queries.
    - **Processing Module**:
        - Applies Named Entity Recognition (NER) and pattern matching to identify question types.
        - Utilizes tokenization, query formulation, and question classification to structure queries effectively.
    - **Data Source Module**: Initially interfaces with the Wikipedia API but will be modified to pull data from FHWA resources and indexed NDE documents.
    - **Logging Module**: Records queries and responses for further analysis and refinement.
- **System Architecture Diagram**:
    - Diagram would include:

User Input → 2. **NER & Pattern Matching** → 3. **Data Source Query (API)** → 4. **Answer Generation** → 5. **Logging & Analysis**

3. SW/HW Development Platforms

- **Software Platforms**:
    - **Python**: Core language for system development.
    - **Packages**: nltk for NER, BeautifulSoup for data scraping, and standard libraries such as re, sys, and datetime.
    - **Frameworks**: Utilizes Jupyter Notebook for development and testing, with plans to incorporate LangChain and vector databases for semantic search.
- **Hardware Requirements**: Minimal computational power as the system primarily processes text queries.

4. Baseline Solution

- **Baseline Solution Description**: The baseline system pulls information from FHWA website based on keyword recognition and pattern matching. It logs questions and answers and attempts alternative responses if no direct answer is found. Currently, we are working on using Beautiful Soup for web scraping to gather more specific information; however, we are facing issues in the scraping process.
- **Errors and Limitations**: Challenges in Web Scraping, encountering errors in the scraping process with Beautiful Soup, hindering access to targeted information sources.

5. Error Analysis for Baseline Solution

- **Identified Errors**:
    - **Query Misinterpretation**: Occasional inaccuracies in understanding the technical intent due to general NER limitations.

6. Proposed Solutions

- **Description**:
  - Replace the data source with a domain-specific database (Infotunnel FHWA) and an indexed collection of NDE technique documents.
  - Implement BERT-based transformer models to improve passage filtering and ranking for more accurate retrieval.
  - **Suggestion**: During the presentation, it was recommended to use a combination of traditional techniques and a Retrieval-Augmented Generation (RAG) approach. This would involve leveraging traditional search methods alongside generative models to enhance response relevance, especially in technical contexts.
- **Expected Outcomes**: Improved accuracy in answers focused on construction and NDE techniques, with more relevant and context-aware responses.

7. Experimental Results

- **Initial Experiments**: Transitioning from Wikipedia to FHWA-specific data and NDE-related documents is anticipated. Performance evaluations will compare accuracy and relevance between general Wikipedia data and domain-specific FHWA data. Additionally, testing will assess improvements with transformer models and RAG-based approaches.

8. Analysis and Interpretation

- **Result Analysis**: Expect enhanced results in accuracy and relevance by focusing on FHWA-specific data and using transformer models for improved context understanding.
- **Challenges**: Implementing a refined NER model to accurately interpret technical jargon and integrating RAG to balance retrieval and generation.
- **Open Questions**: Further development needed to handle complex, multi-part technical queries, and assess RAG's impact on enhancing traditional methods for specialized question-answering.

References

[1] Tunnel. FHWA InfoTechnology. (n.d.). https://infotechnology.fhwa.dot.gov/tunnel/

[2] Dr .Liao's material .https://tinyurl.com/ytx2km9s

[3] Draw io. https://app.diagrams.net/