Brance Founding Applied AI Researcher Task

Name:

Nandana Sreeraj

Linkedin Profile:

https://www.linkedin.com/in/nandana-sreeraj/

Date on which Challenge was received:

5th July 2023

Date on which solution was submitted:

10th July 2023

Contents

Chapter 1

Problem Statement

The task involves building a chatbot that utilizes the RAG framework to provide accurate and personalized answers to user questions.

The task requires the development of a chatbot using the Retrieval Augmented Generation (RAG) methodology. A knowledge document is provided. The objective is to create a chatbot that can effectively respond to questions given by the user by leveraging the information contained within the document.

A RAG module has to be defined crucially to handle user inquiries by employing a two-phase approach. The first phase is the retrieval phase, where the module retrieves relevant contextual information from the knowledge document based on the user's question. This retrieval process aims to ensure that the chatbot has access to the most important and relevant information to provide accurate and meaningful responses.

The second phase is the generation phase, where the RAG module utilizes a Large Language Model (LLM) to generate appropriate answers. The retrieved knowledge serves as a foundation, allowing the LM to craft responses that specifically address the user's question. By combining the power of retrieval and generation, the chatbot can offer a more tailored, customized and informative conversational experience.

The retrieval phase locates the relevant information from the knowledge document, where as the generation phase leverages the Large Language Model (LLM) to generate responses that cater to the user's specific needs.

Chapter 2

Approach

2.1 Detailed Approach

In my chosen approach, I am building a personalized chatbot using LLama-Index and LangChain. By leveraging LangChain, I am using HuggingFace Model Hub, which offers a wide range of language models that can enhance my bot's capabilities.

I am utilizing GPT-Index to construct and query a vector index for efficient information retrieval. This index allows my chatbot to quickly search and retrieve relevant information based on user queries. By organizing the knowledge base using GPT-Index, I can swiftly search and retrieve relevant information based on user queries. This not only improves the responsiveness of my chatbot but also ensures that the responses it generates are accurate and contextually appropriate.

In the approach, Retrieval Augmented Generation (RAG) is employed to enhance the capabilities of the KnowledgeBot. RAG is a framework that combines the retrieval phase and the generation phase to provide more accurate and contextually relevant responses.

In the retrieval phase, the LLama-Index and LangChain components play a significant role. The LLama-Index is utilized to build and query a vector index, which stores the docu-

ment embeddings. These embeddings capture the semantic information of the documents. When a user query is received, the vector index is searched to retrieve the most semantically similar document vectors. This retrieval step helps identify the relevant documents that serve as context for generating accurate responses.

Once the relevant documents are retrieved, they are passed as context to the LangChain component. LangChain orchestrates the interaction between the retrieval and generation phases. The retrieved documents serve as valuable information for the Language Model (LLM) to generate responses that are personalized and contextually appropriate to the user's query.

In the generation phase, the Language Model (LLM) from Hugging Face is employed to produce the final response. The retrieved documents, serving as context, are provided to the LLM to generate personalized and contextually relevant answers.

Using the Hugging Face LLM, the model utilizes the retrieved documents and the user's query to understand the context and generate responses that are tailored to the specific question. The LLM leverages its language generation capabilities to produce coherent and informative answers based on the combined knowledge obtained from the retrieved documents.

By combining LLama-Index, LangChain, and GPT-Index, I aim to create a chatbot that excels in both language processing and information retrieval. This holistic approach enables my chatbot to comprehend user queries effectively and provide highly accurate and timely responses. By leveraging powerful tools like LLama-Index and LangChain, I can enhance the conversational experience and deliver a chatbot that meets the needs and expectations of users

2.2 Assumptions

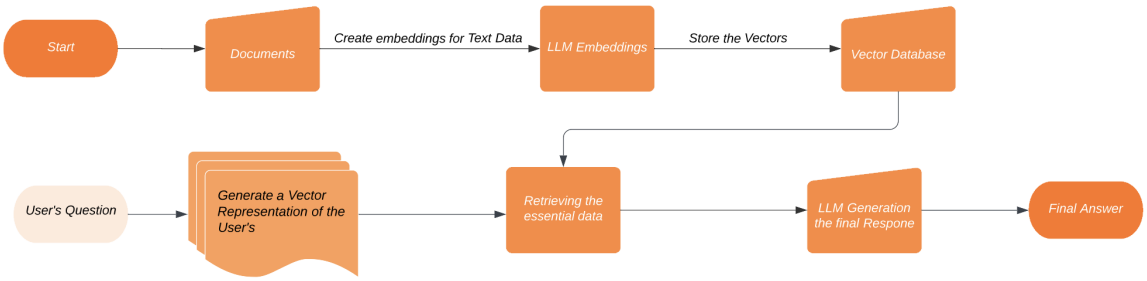
- It is assumed that the documents retrieved during the retrieval phase are relevant and contain valuable information pertaining to the user's query.
- The document embeddings generated during the document embedding phase are assumed to accurately capture the semantic information of the documents.
- The LLM is assumed to possess adequate semantic understanding and language processing capabilities

Chapter 3

Solution

- Document Embedding And Vectorization.
  - The embedding step transforms documents into a vector representation called embedding using HuggingFaceEmbeddings.
  - To build the vector index, textual data is encoded into a high-dimensional vector using HuggingFaceEmbeddings.
  - These vector representations capture the semantic information of the document.
  - The LLama index organizes and stores these vectors, allowing for efficient searching and retrieval based on similarity measures.
- Processing the user given query
  - When the user puts up the question, the LLM is used to transform the question into a vector, similar to how we processed the text in the previous step.
- Retrieving essential data and Answer Generation
  - Based on the question, the vector database is searched to identify the most semantically similar vectors to the question vector.
  - The vector database does not contain actual text, it contains only the embeddings and an identifier.

– The LLM generates an answer based on the provided information.



Chapter 4

Evaluation Matrix: Rouge Score

ROUGE scores evaluate the quality of the generated text based on the overlap between the generated text and the reference text. It calculates various metrics such as ROUGE-N (N-gram overlap), ROUGE-L (Longest Common Subsequence), and ROUGE-S (Skip-bigram) to measure the similarity between the generated text and the reference text.

Rouge scores provide insights into the quality and effectiveness of the LLM's generated responses, allowing us to make informed decisions about model selection or optimization.

- ROUGE-1: ROUGE-1 measures the overlap of unigram (single-word) units between the generated text and the reference text. It computes the precision, recall, and F1 score based on the count of matching unigrams between the two texts. ROUGE-1 focuses on the exact word matches and evaluates the quality of the generated text in terms of unigram overlap.
- ROUGE-2: ROUGE-2 measures the overlap of bigram (two-word) units between the generated text and the reference text. It calculates the precision, recall, and F1 score based on the count of matching bigrams. ROUGE-2 captures the co-occurrence and order of word pairs and assesses the quality of the generated text in terms of bigram similarity.

- ROUGE-L: ROUGE-L measures the longest common subsequence (LCS) between the generated text and the reference text. It focuses on capturing the longest sequence of words that appears in both texts, regardless of their order. ROUGE-L evaluates the overall content and captures the semantic similarity between the generated and reference text, considering their structural differences.

Chapter 5

Support for Audio Questions From User

I have used Whisper AI for Speech to Text Conversion. Whisper AI is used in enabling applications and services to transcribe spoken language into written form. It leverages advanced deep learning models and neural network architectures to accurately recognize and convert speech into text. Whisper AI's speech-to-text capabilities offer accurate and scalable solutions for converting spoken language into written text

Chapter 6

Future Scope

The future scopes for the approach utilizing RAG in the chatbot include improving personalization and enhancing document context understanding. We can also expand the knowledge base. A more refined evaluation process can be employed. We can also improve performance through iterative evaluation and fine-tuning.