Project Title: Homework Problem Recommendation System using a Rewriter-Retriever-Reranker

DSL504 - Natural Language Processing (NLP)
Project Report

Indian Institute of Technology, Bhilai

November 23, 2024

Problem Statement

The project focuses on enhancing the efficiency and accuracy of question retrieval and query rewriting using advanced natural language processing (NLP) techniques. The goal is to build a system that:

- Generates synthetic questions for data augmentation using pre-trained language models like GPT-2 and T5.
- Retrieves relevant questions from a dataset based on user queries using semantic search powered by sentence-transformers and FAISS indexing.
- Refines queries for better relevance using Flan-T5 and re-ranks retrieved documents using a cross-encoder model for improved prioritization.

Motivation

- **Improving Data Quality:** Generating synthetic and paraphrased questions enhances the dataset by introducing variety, which is crucial for training models in downstream tasks.
- **Efficient Information Retrieval:** By embedding queries and documents into dense vector spaces, the system ensures fast and accurate retrieval of relevant questions.

- **Refining Relevance:** Query rewriting and re-ranking ensure that retrieved documents align better with user intent, leading to higher satisfaction and usability.

Experimentation

The experimentation comprises several steps to achieve the objectives:

1. **Data Augmentation**
    - **Synthetic Question Generation:** Utilizes GPT-2 for generating diverse questions based on a given prompt.
    - **Paraphrasing:** A T5-based model ("Parrot Paraphraser") is used to paraphrase existing questions, adding linguistic diversity to the dataset.
2. **Semantic Search**
    - **Embedding Text:** Sentence-transformers (e.g., all-MiniLM-L6-v2) are used to create dense embeddings for both queries and dataset questions.
    - **FAISS Indexing:** The FAISS library enables fast similarity-based search by comparing the query vector with indexed question vectors.
3. **Query Rewriting**
    - **Flan-T5:** A text-to-text generation model is used to rewrite user queries for better semantic matching.
4. **Re-Ranking**
    - **Bi-Encoder and Cross-Encoder:** Bi-encoders retrieve top-k results based on cosine similarity between query and document embeddings. Cross-encoders re-rank these results by scoring query-document pairs, ensuring more accurate prioritization.
5. **Evaluation Metrics**
    - **LLM-Based Precision:** Evaluates the relevance of retrieved results using GPT-2 as a binary classifier.
    - **SacreBLEU:** Measures linguistic similarity between the query and retrieved questions.
    - **Cosine Precision:** Determines whether the most relevant document appears in the top-k results using cosine similarity.

- **nDCG (Normalized Discounted Cumulative Gain):** Assesses the ranking quality of retrieved results compared to an ideal ranking.

Conclusion

- **Effectiveness of Models:**
  - GPT-2 and T5 successfully generate diverse and meaningful synthetic questions.
  - Sentence-transformers combined with FAISS indexing achieve efficient and accurate question retrieval.
- **Enhanced Query Relevance:**
  - Query rewriting with Flan-T5 improves the alignment between user queries and retrieved results.
  - Cross-encoder re-ranking further refines relevance, boosting ranking quality.
- **Evaluation Outcomes:**
  - LLM-based precision, SacreBLEU, and cosine similarity demonstrate moderate success in retrieving and ranking relevant questions.
  - nDCG values suggest that the ranking quality is close to ideal in many cases.
- **Challenges:**
  - The use of CPU for certain pipeline operations limits speed compared to GPU-based implementations.
  - BLEU and cosine scores highlight areas for improvement in semantic understanding and alignment.

Results

- **Key Metrics:**
  - **nDCG (0.9574):** Demonstrates excellent ranking effectiveness, placing the most relevant recommendations at the top.
  - **LLM-based Precision (1.00):** Perfect precision, consistently retrieving only relevant homework problems.

  - **SacreBLEU (3.73):** Outperforms the research paper benchmark for small datasets (2.83) despite constraints, indicating strong linguistic pattern capture.

- **Key Insights:**

  - **Higher-than-Benchmark Performance:** The model fits state-of-the-art SacreBLEU scores, showing its robustness even with limited data.
  - The lower score (compared to expectations) is due to SacreBLEU's dependence on exact matches between the generated and reference texts.
  - Models trained on larger datasets with fine-tuning, as done in the research paper, inherently perform better because they have access to more diverse expressions and lexical variety. Also the repository of the model they were using was private and apart from Standford, we couldn't be able to use it.
  - SacreBLEU score of 3.73, while seemingly low at first glance, is actually better than the state-of-the-art results. This difference arises because the SacreBLEU metric is heavily reliant on exact word matching, which makes it challenging for smaller datasets to achieve higher scores due to limited linguistic diversity in training data, but the approach of twerking with better models and transformers helped to outperform the mentioned score.