Enhancing search queries using large language models to optimize relevance and precision in information retrieval

Tarun Anoop Sharma, tsharma7@illinois.edu
Sanket Nagesh Babu Donty, sdonty2@illinois.edu
Bin Wu, binw3@illinois.edu
Keyuan Chang, keyuanc2@illinois.edu

Track: Research Track

Research Question

To study whether LLMs like ChatGPT can be used to augment user-made search queries to enhance search performance.

Significance

Today's search engines excel with common queries but face difficulties when queries are 'incomplete'. When users lack specific knowledge about the document they seek, because of a ‘vocabulary gap’, their queries often miss the terms needed to retrieve the right documents. This typically results in unsatisfactory search outcomes. Enhancing performance in these challenging scenarios will significantly help users in finding documents in cases of vocabulary gap.

Novelty

Our method introduces a novel approach by utilizing the latest large language models (LLMs) to understand user intent in incomplete queries. This enables the retrieval of relevant documents based on missing or hidden terms in incomplete queries caused by vocabulary gaps from the user. Till date, deep learning models have been studied to bridge the vocabulary gap in queries using representation learning [1]. Use of LLMs has also been explored in different stages of retrieval [2]. However, the idea of using LLMs to bridge the vocabulary has not yet been explored.

Approach

We propose the following:

- Pass context of the collection of documents in a ‘compressed’ format to the LLM.
- Provide the LLM with an incomplete query and the unsatisfactory documents that were returned for the query
- Prompt the LLM, given all this information, to discern the intent of the user and augment the query for better performance of the search engine.
- Evaluate the modified query and compare with the original query.

Evaluation

We will assess the effectiveness of our proposed strategies by comparing its performance against a standard baseline algorithm based on incomplete queries.

The evaluation process involves collecting multiple datasets comprising document collections and associated queries. We will identify and only include queries (incomplete queries) that fail to yield relevant recommendations within the top k results when processed by the baseline algorithm. This filtration ensures that our evaluation focuses on data where there is potential for meaningful improvement.

Timeline

04/15 – 04/22:

- Explore passing context of collection of documents to LLMs.
- Benchmark the existing performance of incomplete queries on datasets.
- Design front-end and backend for a user-facing application to test the proposed approach.

04/23 – 04/29:

- Start work on the front-end and backend for the application. Include UI features to allow users to enter a query, see a ranked list of relevant documents, click for a rerun using the proposed approach and compare the results.
- Experiment with different strategies to improve performance on incomplete queries. Document the findings of all the strategies implemented.

04/30 – 05/06:

- Finalize the best set of prompt strategies that result in the best performance for search with incomplete queries.
- Integrate the front-end, backend with the LLM model to build the complete system end-to-end.
- Test the whole system end-to-end and verify if the search results are indeed improved for incomplete queries on the collection.

Task Division

- Front-end Development: Keyuan Chang
- Backend Development: Bin Wu
- Integration of the application with the LLM model: Keyuan Chang, Bin Wu
- Exploration of passing collection context to LLMs: Tarun Anoop Sharma, Sanket Nagesh Babu Donty
- Experiment different strategies of prompting LLMs for query enhancement: Tarun Anoop Sharma, Sanket Nagesh Babu Donty

References

[1] X. Li, H. Jiang, Y. Kamei and X. Chen, "Bridging Semantic Gaps between Natural Languages and APIs with Word Embedding," in *IEEE Transactions on Software Engineering*, vol. 46, no. 10, pp. 1081-1097, 1 Oct. 2020, doi: 10.1109/TSE.2018.2876006 https://ieeexplore.ieee.org/abstract/document/8493293/

[2] Carraro, D., Bridge, D., *Enhancing Recommendation Diversity by Re-ranking with Large Language Models* https://arxiv.org/abs/2401.11506