Context Refining for Retrieval Augmented Generation

Juan Yu

Abstract

Due to the diverse training data, large language models (LLMs) may not be able to give an accurate answer to a domain-specific question. Retrieval Augment Generation (RAG) is a framework that provides a solution by grounding LLMs on specific evidence documents. On top of RAG, this project takes a step further and focuses on refining context from documents that are already retrieved and provide evidence for answering questions.

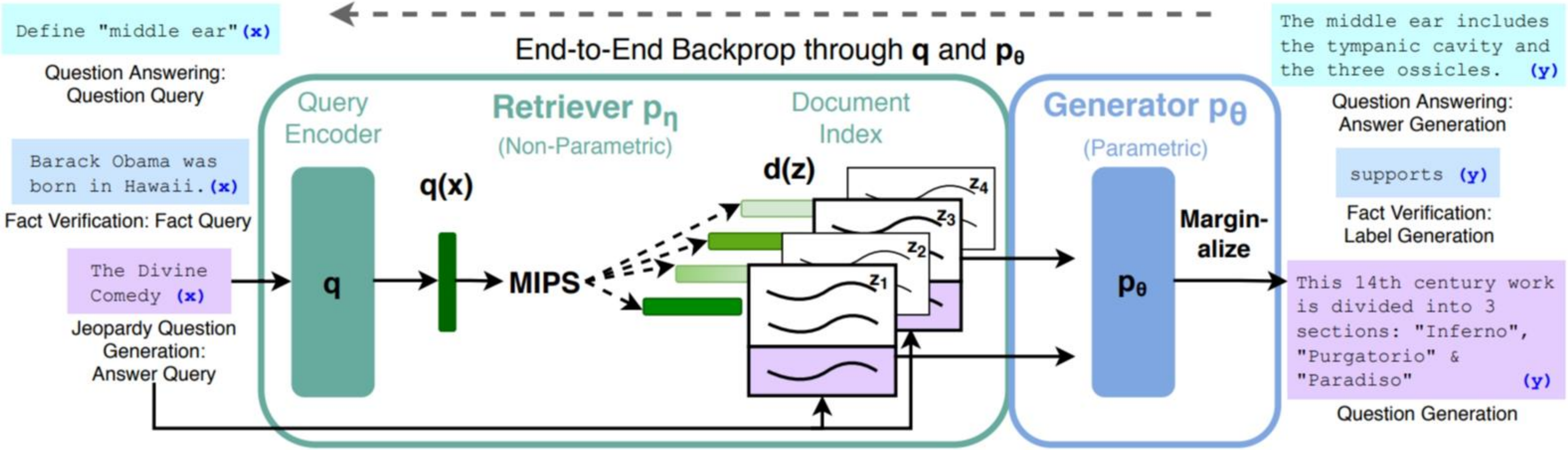
Since no gold-standard context data is available, the evaluation will be based on the final outputs. Llama 2, the latest state-of-art open-source model, will be used to generate the final outputs. Exact Match will be used as the metric to measure the accuracy of the final outputs, and comparison will be made between inputs with refined context and inputs with un-refined evidence documents.

Background

As we know, large language models (LLMs) are trained on **massive, diverse data**. The diversity of the training data means that LLMs may not be able to answer a domain-specific question accurately. Therefore, it is usually necessary to ground the content generated by an LLM on some specific knowledge.

Literature Review

- *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,* Lewis et al., (2020)



Define "middle ear" (x)
Question Answering: Question Query
Barack Obama was born in Hawaii. (x)
Fact Verification: Fact Query
The Divine Comedy (x)
Jeopardy Question Generation: Answer Query
End-to-End Backprop through q and p_θ
Query Encoder
q
q(x)
Retriever p_η
(Non-Parametric)
MIPS
Document Index
d(z)
z_4
z_3
z_2
z_1
Generator p_θ
(Parametric)
p_θ
Margin-alize
The middle ear includes the tympanic cavity and the three ossicles. (y)
Question Answering: Answer Generation
supports (y)
Fact Verification: Label Generation
This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso" (y)
Question Generation

Literature Review

- *Dense Passage Retrieval for Open-Domain Question Answering,* Karpukhin et al., (2020)

  - Dense representation of queries and candidate contexts
  - Query encoder +context encoder
  - Training to maximize similarity with positive contexts and minimize similarity with negative contexts

Problem Statement

The classic RAG framework retrieves top-k documents that are relevant to a given query, augments the query by adding the retrieved documents to it, and then inputs the augmented query to a generating LLM for final output.

The retrieved documents could still be too broad for answering a specific question and may easily exceed the maximum input length for an LLM. For example, in the original paper, top k (k>5) documents, each of which contains 100 words, were retrieved from Wikipedia dump.

This project aims to use informational retrieval and natural language processing techniques to refine context from evidence documents, so as to shorten the length of inputs and improve the accuracy of outputs.

Methodology

- **General idea:** evidence documents are split into chunks, and top-k chunks with the highest similarity scores with a query will be selected as the context.

- **Splitting methods:**

  -by sentence (1 sentence per chunk)

- **Similarity scoring methods:**

  -TF-IDF

  -BM25

  -Pretrained DPR document encoder and query encoder

Experiments - Data

- **TriviaQA:** a dataset consisting of question-answer pairs + evidence documents (six documents per pair on average)

- **Download:** https://huggingface.co/datasets/trivia_qa/viewer/rc

Experiments - Evaluation

- **General idea:** evaluate the accuracy of final outputs based on gold answers, and compare accuracy between inputs with refined context and inputs with un-refined evidence documents.
- **Generator:** llama 2 (trained between January 2023 and July 2023)
- **Baseline:** evidence documents + queries as inputs
- **Metrics:** accuracy based on Exact Match

Conclusion and discussion

- This project focuses on further refining context from evidence documents based on input queries.
- Refined context greatly reduces input length while still strongly supporting a generator to output desirable answers to queries.
- Due to the flexibility of human language, encoders fine-tuned on domain-specific knowledge bases may work better for domain specific tasks. Limited by data availability and computing resources, this project will not be able to fine tune encoders. However, the framework can be applied across different domains.

References

- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474. https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf
- Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. In *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 6769-6781). (EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference). Association for Computational Linguistics (ACL). https://aclanthology.org/2020.emnlp-main.550.pdf
- Wang, Z., Araki, J., Jiang, Z., Parvez, M. R., & Neubig, G. (2023). Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*. https://arxiv.org/pdf/2311.08377.pdf
- Joshi, M., Choi, E., Weld, D. S., & Zettlemoyer, L. (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*. https://arxiv.org/pdf/1705.03551.pdf