KG-Retriever: Efficient Knowledge Indexing for Retrieval-Augmented Large Language Models

Weijie Chen[1], Ting Bai[1*], Jinbo Su[2], Jian Luan[3], Wei Liu[3], Chuan Shi[1]

[1]Beijing University of Posts and Telecommunications,
[2]University of Science and Technology Beijing,
[3]Xiaomi Corporation.

Abstract

Large language models with retrieval-augmented generation encounter a pivotal challenge in intricate retrieval tasks, e.g., multi-hop question answering, which requires the model to navigate across multiple documents and generate comprehensive responses based on fragmented information. To tackle this challenge, we introduce a novel Knowledge Graph-based RAG framework with a hierarchical knowledge retriever, termed KG-Retriever. The retrieval indexing in KG-Retriever is constructed on a hierarchical index graph that consists of a knowledge graph layer and a collaborative document layer. The associative nature of graph structures is fully utilized to strengthen intra-document and inter-document connectivity, thereby fundamentally alleviating the information fragmentation problem and meanwhile improving the retrieval efficiency in cross-document retrieval of LLMs. With the coarse-grained collaborative information from neighboring documents and concise information from the knowledge graph, KG-Retriever achieves marked improvements on five public QA datasets, showing the effectiveness and efficiency of our proposed RAG framework.

1 Introduction

Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) have achieved initial success in generating accurate responses, especially in knowledge-intensive tasks. By using a retrieval component to incorporate relevant information from a vast corpus of external documents, the RAG technique has become the mainstream way to alleviate hallucination issues in the response generation of LLMs (Yang et al., 2023; Ding et al., 2023). Nonetheless, when engaging in intricate retrieval tasks, e.g., multi-hop question answering, the model encounters significant difficulties in extracting pertinent answers solely from a single document. It requires the model to navigate across multiple documents and generate comprehensive responses according to the fragmented information from multiple relevant documents. For example, in the multi-hop question "What are the trend and major factors contributing to dry eye syndrome in children, and what preventive measures can be taken?", most existing RAG-based LLMs inevitably suffer from incomplete retrieval knowledge due to the disability of reasoning over different documents. Recent RAG studies (Feng et al., 2024; Shao et al., 2023) attempt to disassemble the retrieval process into iterative retrieval steps by using the generated content from the last iteration as the query to retrieve relevant documents in the next round. Such approaches improve the inferential capabilities of LLMs and enhance the retrieval quality to some extent, but they still face the challenge of the escalating computational costs caused by multiple iterative retrieval steps. Besides, due to the discreteness of each iteration, such methods may still face poor retrieval performance in integrating information across different documents.

To enhance retrieval quality while maintaining RAG's efficiency, we propose a novel knowledge graph-based RAG framework with a hierarchical knowledge retriever, termed **KG-Retriever**. Specifically, KG-Retriever is built based on a Hierarchical Index Graph (**HIG**) that consists of a knowledge graph layer and a collaborative document layer (as shown in Fig. 1). In the knowledge graph layer, entities and relations in a document are extracted by LLMs, which enhances the internal information structuring of individual documents. In the collaborative document layer, the document-level graph establishes connections based on their semantic similarity, improving the cross-document knowledge correlations. The entity-level and document-level information in HIG

[*]Corresponding author.