CDRAG: An RAG framework for enhancing cross-document reasoning of an LLM

PIPATPAJONG Thapakorn[1]   Supervisor: Prof. James CHENG[1]

[1]The Chinese University of Hong Kong

Introduction

- Retrieval-Augmented Generation (RAG) is the popular method to extend the Large Language Model (LLMs) knowledge by retrieving relevant external documents to assist in generating answers.
- However, traditional RAG has struggled in handling the complex query that requires cross-document reasoning.
- An CDRAG (Cross-Document Retrieval-Augmented Generation) is proposed, a framework that enhances LLMs by improving the knowledge graph with Named Entity Recognition (NER) to capture more details.

Research Objectives

- Design the GraphRAG pipeline that enables complex question-answering and cross-document retreival.
- Construct a knowledge graph that effectively captures the detail in the documents.
- Compare the CDRAG framework with other existing RAG frameworks, such as traditional RAG and Knowledge Graph Generation (KGGen).

Method: The CDRAG Pipeline

The CDRAG framework consists of 2 important processes.

- Graph Construction: Texts are converted into graph with entities as nodes and relationships as edges. Enhancing the representation by using NER.

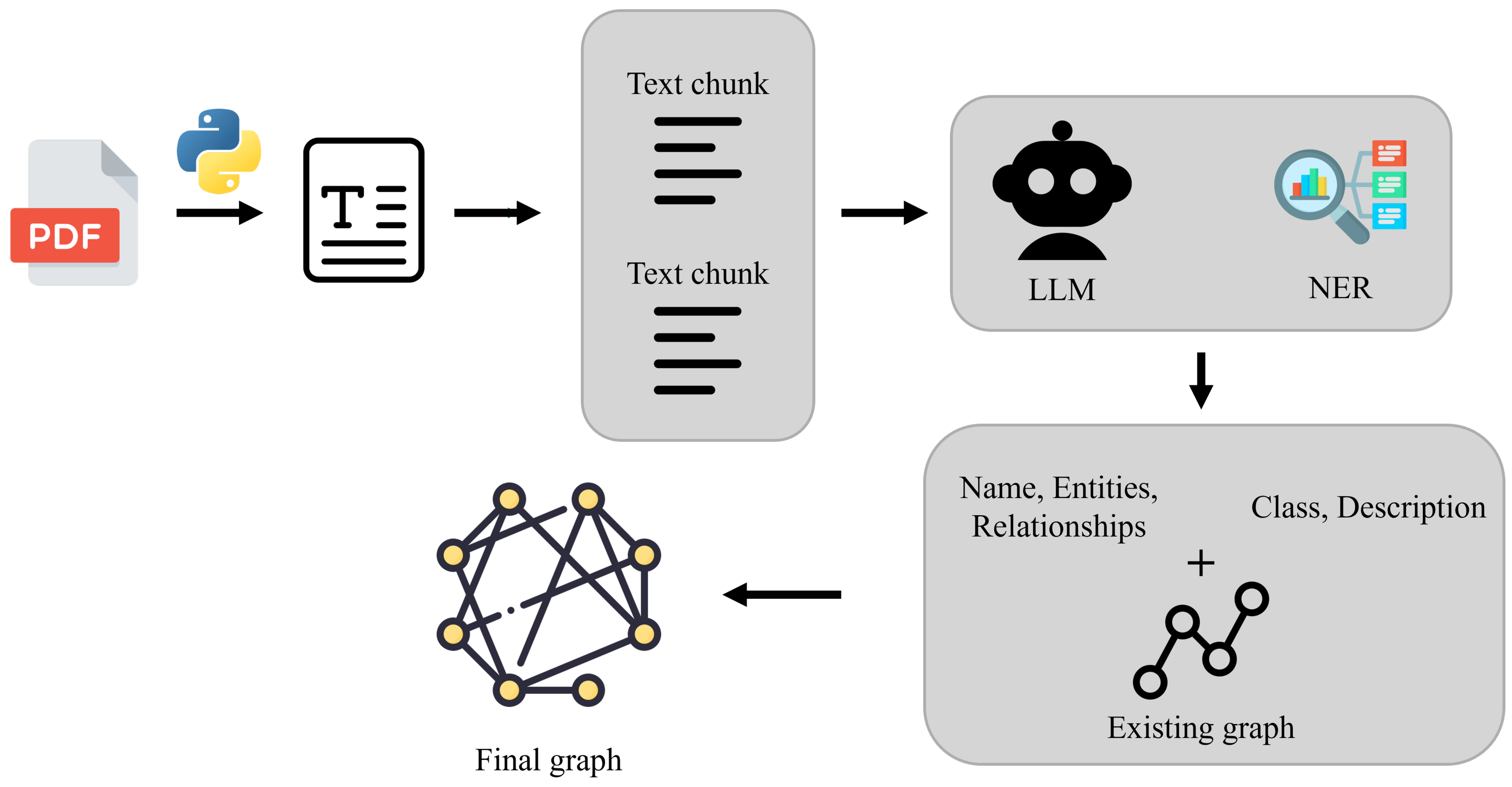

Figure 1. The CDRAG graph construction pipeline

- Information Retrieval: Turning query into vector for similarity search and explore the neighbors in the respective communities.

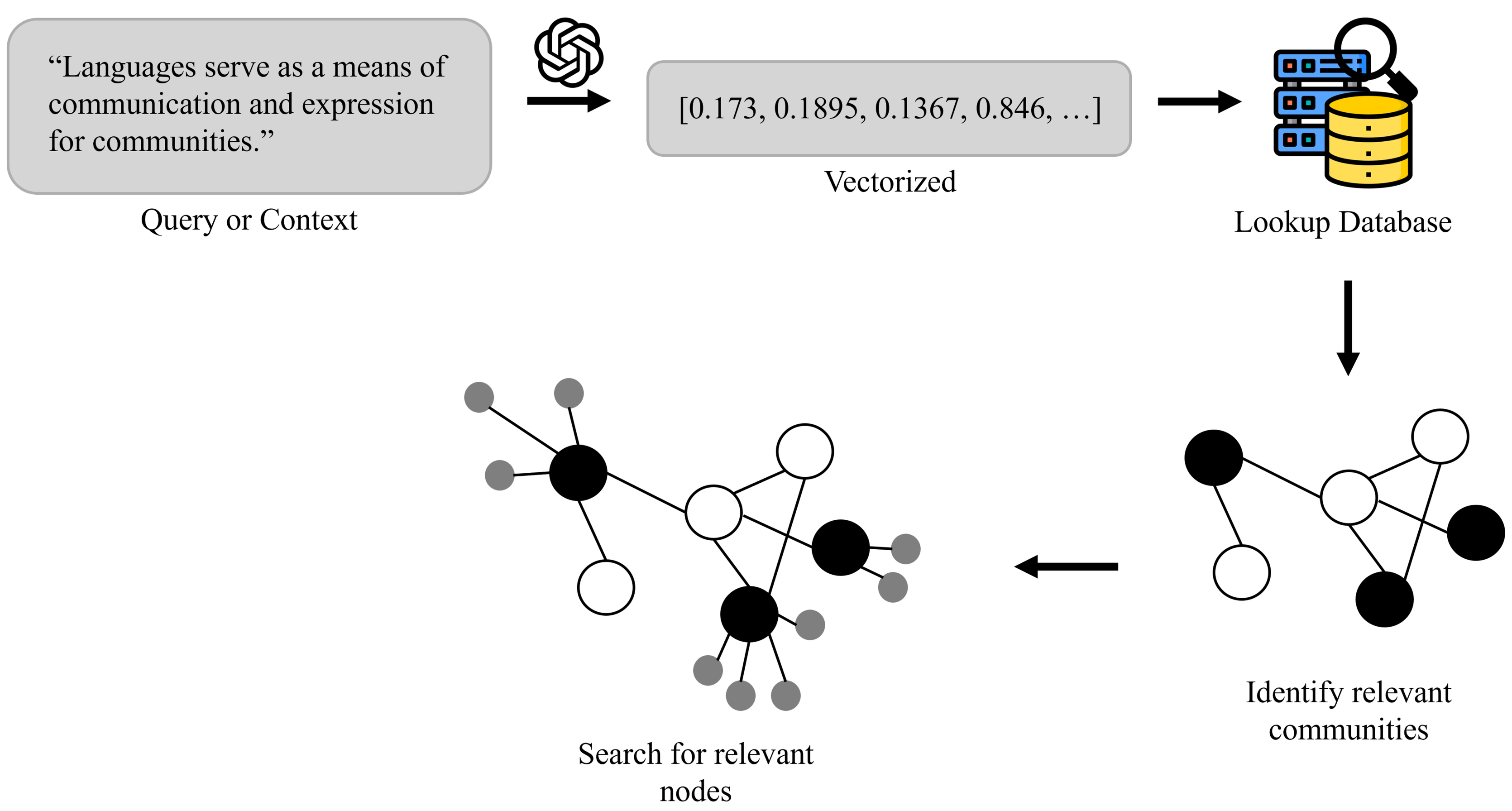

Figure 2. The CDRAG information retrieval pipeline

Experimental Results

Dataset:

- The Measure of Information in Nodes and Edges (MINE) dataset contains 105 articles, ~1000 words each, which covers multiple fields: art, science, history, ethics, and psychology.
- The relevancy of the retrieval in each RAG method will be measured through this dataset. Three RAG methods are tested.

Results:

- The tradition RAG struggles to retrieve the relevant context from multiple documents with only 83.05%, almost 10% lower than KGGen.
- KGGen and CDRAG are both graph-based method, which also have similar accuracy, with CDRAG slightly ahead by 2.60% due to richer details construction.
- CDRAG also perform more consistent than KGGen, which no domains have under 60% while KGGen has some topics around 50% accuracy.

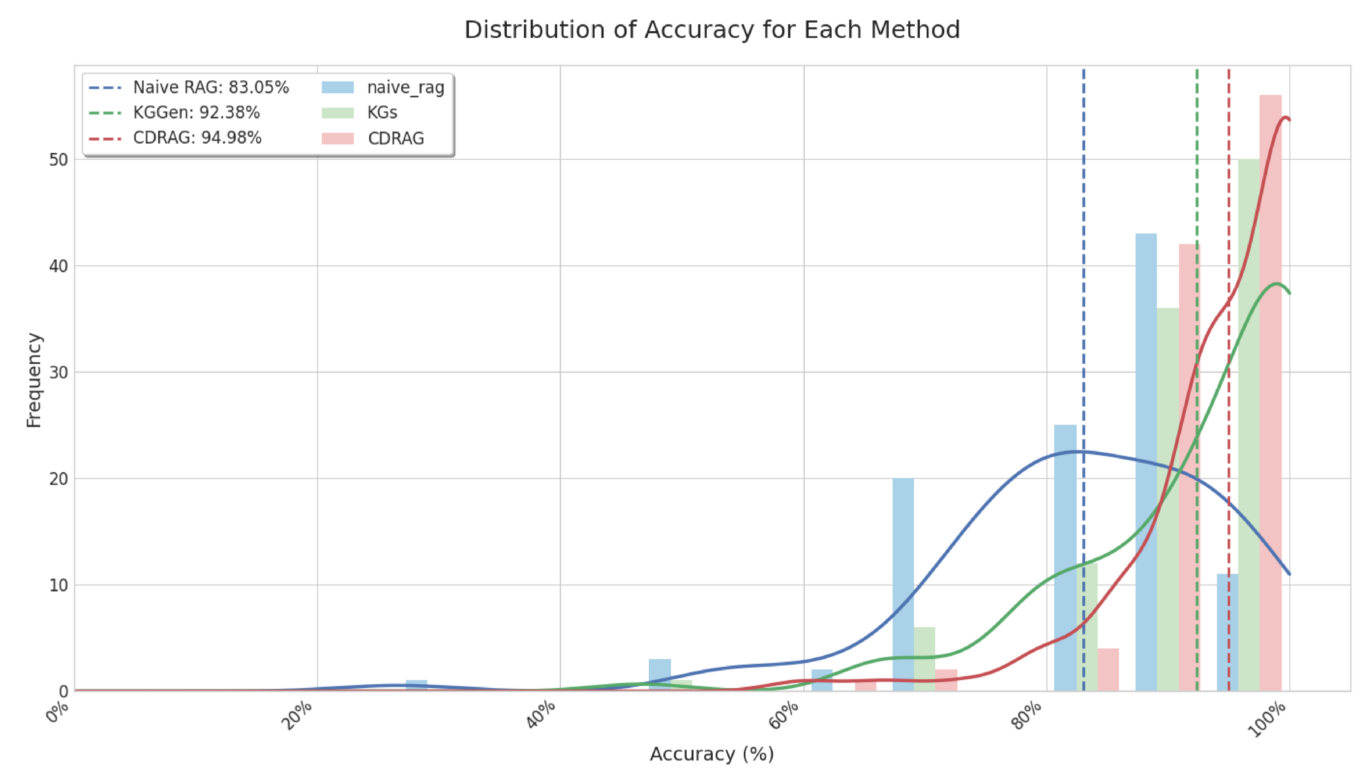

Figure 3. The benchmark results

Conclusion and Future Works

- CDRAG bridges the gap in cross-document reasoning capabilities of a retrieval mechanism by utilizing knowledge graph to represent information.
- CDRAG achieves superior performance to the existing method by incorporating Named Enity Recognition to capture more details in the documents.
- The future works may explore the CDRAG to much larger dataset to test the retrieval mechanism on the real data.
- Moreover, the graph may extend further to represent images, and videos on top of just plain text representation.

References

[1] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

[2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[3] Belinda Mo, Kyssen Yu, Joshua Kazdan, Proud Mpala, Lisa Yu, Chris Cundy, Charilaos Kanatsoulis, and Sanmi Koyejo. Kggen: Extracting knowledge graphs from plain text with language models. *arXiv preprint arXiv:2502.09956*, 2025.

[4] Hairong Zhang, Jiaheng Si, Guohong Yan, Boyuan Qi, Pinlong Cai, Song Mao, Ding Wang, and Botian Shi. Rakg: Document-level retrieval augmented knowledge graph construction. *arXiv preprint arXiv:2504.09823*, 2025.