QA Modeling

Dr. Sabah Mohammed
Prof and Chair,Dept. of Computer Science
Lakehead University
Thunder Bay, ON
sabah.mohammed@lakeheadu.ca

Kushal
M.S. Computer Science
Lakehead University
Thunder Bay, ON
kkushal@lakeheadu.ca

Tharun Sekar
M.S. Computer Science
Lakehead University
Thunder Bay, ON
tsekar@lakeheadu.ca

Supprethaa Shankar
M.S. Computer Science
Lakehead University
Thunder Bay, ON
sshanka3@lakeheadu.ca

***Abstract*—The integration of graph-based reasoning and retrieval-augmented generation (GraphRAG) offers promising potential for medical applications, especially in handling complex, interconnected datasets. This project aims to develop a GraphRAG application to support clinical decision-making by extracting relevant information from structured medical data on heart attacks. Using Neo4j to model entities and relationships and a JSON dataset to structure questions and documents, this approach enables efficient, evidence-based query responses. Our methodology leverages advanced graph modeling to enhance retrieval accuracy, offering healthcare professionals a reliable tool for information access. Early results indicate improved retrieval relevance, demonstrating GraphRAG's potential to streamline clinical workflows in real-world applications.**

***Index Terms*—GraphRag, Neo4j, JSON**

I. Introduction

Recent advancements in artificial intelligence, particularly in the domains of graph modeling and retrieval-augmented generation (RAG), offer new possibilities for handling complex medical data. In healthcare, where vast, interconnected datasets are frequently used, clinicians and researchers face challenges in quickly retrieving relevant information from this data. Applications that leverage graph-based models can efficiently represent and explore relationships within medical data, offering a path toward enhanced evidence-based decision support.

A. Problem Statement

Traditional models for clinical data retrieval often rely on simple, text-based search methods, which struggle with the complexity and interconnected nature of medical knowledge. There is a need for models that can better capture the relationships among medical entities, such as symptoms, treatments, and conditions, to improve the accuracy and relevance of retrieved information in real-world medical contexts.

This project aims to develop a GraphRAG application that utilizes graph-based reasoning to enhance information retrieval for clinical decision support. By modeling medical data on heart attacks in Neo4j, this project seeks to improve the relevance and efficiency of query responses. The use of a structured JSON dataset allows for consistent and context-aware information retrieval, tailored to address specific clinical questions.

To address the challenges of complex data retrieval in healthcare, this project proposes a GraphRAG model that leverages Neo4j for entity and relationship management. By using structured JSON data to simulate clinical queries, this approach facilitates the extraction of relevant, context-aware information from medical datasets. This model is designed to support clinicians by offering a precise, graph-enhanced retrieval mechanism that aligns with medical best practices.

II. Literature Review

The domain of question-answering (QA) systems has seen significant advancements, particularly with the integration of deep learning (DL) and transformer-based architectures. This review consolidates insights from recent literature, highlighting the progress, challenges, and future directions.

A. Medical Textual Question Answering Systems

A comprehensive review of deep learning techniques for medical QA systems by MDPI, 2021 categorizes key approaches, such as autoencoders, convolutional neural networks (CNNs), and recurrent neural networks (RNNs), including advanced hybrid models. The study emphasizes the transformative potential of DL in medical QA while underscoring challenges in model accuracy and data availability. Future work is directed towards improving these metrics and expanding datasets to enhance system robustness.

B. Neural QA Systems for Programming Subroutines

A domain-specific implementation discussed by IEEE, 2021 employs encoder-decoder models with attention mechanisms for answering programming-related questions. Training on 10.9 million tuples, the model demonstrates strengths in answering Java subroutine queries but reveals limitations in handling complex language constructs. This proof-of-concept highlights the potential for extending neural QA systems to broader domains.

C. Survey on QA Systems

A broader survey on QA systems by Arxiv, 2021 identifies emerging trends, including multimodal QA and open-domain systems. Key advancements include the use of transformer-based models such as BERT and GPT, which offer scalability and explainability. The paper advocates for resource-efficient models and refined evaluation mechanisms to cater to multilingual and user-centric applications.

D. Cross-lingual Open-Retrieval QA

The paper on XOR QA by Arxiv, 2020 explores answering questions in one language using documents in another, leveraging cross-lingual models like XLM-R and mBERT. The research highlights the challenge of language asymmetry and proposes multilingual transformer architectures combined with dense retrieval models for improved performance in such settings.

E. Community QA and DL Models

A systematic review by IET Research, 2023 evaluates 133 articles on community QA (CQA), demonstrating how DL models like BERT and CNN outperform traditional machine learning methods. Challenges in user identification and cross-platform performance remain, while the paper calls for more nuanced analyses of ML and DL model efficacy.

F. QA on Scientific Articles

The QASA system by MLR Press, 2023 introduces a benchmark dataset targeting reasoning over scientific articles. Using associative selection and evidential rationale generation, the system outperforms models like InstructGPT. Future directions include handling multimodal data and refining reasoning pipelines for generalizability across domains.

G. Comparative Analysis of Transformer Models

A study by IEEE, 2023 benchmarks various transformer-based QA models, such as BERT and RoBERTa, on the SQuAD dataset. RoBERTa outperformed BERT variants, while future work includes integrating XLNet for better multilingual performance and exploring generative capabilities with larger models like GPT-3.

H. Hybrid BERT-BiLSTM for QA

The introduction of a hybrid BERT-BiLSTM model by IEEE, 2024 incorporates refined embeddings like SemGloVe to improve semantic understanding in QA. The model surpasses baseline methods, with proposed enhancements focusing on incorporating RoBERTa and expanding applications to sentiment analysis and document classification.

I. Fine-tuning QA Models

Research on fine-tuning extractive QA models for German business documents (Arxiv, 2023) showcases the effectiveness of fine-tuning on small datasets to extract complex information. Proposed advancements include integrating multimodal features and optimizing prompts for better QA model performance.

J. Biomedical QA Chatbots

The use of transformer-based models like BioBERT and RoBERTa for biomedical QA chatbots (IEEE, 2023) emphasizes maintaining semantic accuracy. The study suggests enhancing datasets, expanding chatbot interactivity, and developing user-friendly visual platforms to improve adoption in medical research contexts.

K. GRAPH BASED RAG

Graph-based retrieval-augmented generation (RAG) has recently shown promise for enhancing medical question-answering systems. In Medical Graph RAG: Towards Safe Medical Large Language Models via Graph Retrieval-Augmented Generation (Wu et al., 2024), the authors propose a framework, MedGraphRAG, to improve the transparency and accuracy of large language models (LLMs) in clinical contexts. By leveraging hierarchical graphs built from medical documents and using a U-retrieve strategy, MedGraphRAG enhances context-aware response generation, producing more accurate, evidence-based answers. This model significantly improves upon previous RAG methods by integrating medical knowledge sources, supporting clinicians with zero-shot, evidence-traceable responses.

Similarly, the framework discussed in A Framework for Medical Question Answering (IJRCAIT) builds on ontology-based methods to refine precision in medical question answering, especially in retrieving specific diagnostic and treatment information. This approach combines rule-based and supervised learning techniques, yielding high response specificity and relevance. Benchmarking against conventional models, this framework demonstrates the effectiveness of domain-specific ontology applications for improved accuracy in complex medical queries.

Finally, Zero-shot Graph-based Retrieval in Medicine (Wu et al., 2024) highlights the potential of zero-shot retrieval using a graph approach, particularly for low-cost, scalable solutions. By integrating structured and unstructured medical data, this model enhances LLM performance in medical information retrieval without requiring additional training. It shows a notable advantage in applications needing quick deployment and minimal resource investment, making it suitable for diverse clinical scenarios.

Together, these studies illustrate how graph-enhanced RAG frameworks contribute to more accurate, context-aware, and resource-efficient medical question-answering systems. They underscore the importance of integrating knowledge graphs to advance precision and support real-world clinical needs, paving the way for more robust, evidence-based AI in healthcare. By combining structured knowledge graphs with retrieval-augmented generation, these models address key challenges in precision, contextual relevance, and scalability. They underscore the potential for graph-enhanced frameworks to advance medical information systems, offering evidence-based, accurate responses that are essential for reliable clinical support. Together, they highlight a trajectory toward AI-driven healthcare solutions that meet rigorous standards for safety, efficiency, and contextual accuracy, supporting both cost-effective and high-quality clinical decision-making.

III. Implementing Different Models

In this study, multiple transformer-based and traditional models were implemented and evaluated on two datasets: GBaker/MedQA-USMLE-4-options and Shubham09/medqa, using BLEU scores as the primary evaluation metric. The

findings highlight significant variations in model performance across different datasets and offer insights into how these results differ from existing literature.

A. Evaluation on GBaker/MedQA-USMLE-4-options

The BLEU scores obtained for the GBaker/MedQA-USMLE-4-options dataset demonstrate varying degrees of performance across models:

BERT Large achieved the highest BLEU score of 0.0839, showcasing its ability to generate outputs with better alignment to the reference answers compared to other models. BioBERT performed moderately well with a BLEU score of 0.0552, leveraging its domain-specific pretraining on biomedical data. RoBERTa, XLM-R, and SPARQL-based models produced significantly lower BLEU scores (0.0116, 0.0067, and 0.0012, respectively), indicating challenges in adapting to the structured and domain-specific nature of the dataset. **Comparison to Literature** In existing studies, such as those on biomedical QA systems (BioBERT and SPARQL-based models), these models have demonstrated strong performance on general biomedical tasks. However, their relatively lower BLEU scores here suggest difficulty in handling the fine-grained complexity and multiple-choice format of the USMLE dataset. This may highlight gaps in generalization when transitioning from benchmark datasets like SQuAD or PubMedQA to specialized medical datasets.

B. Evaluation on Shubham09/medqa

The Shubham09/medqa dataset presented even greater challenges for the models:

RoBERTa, XLM-R, BERT Large, and BioBERT all produced BLEU scores close to zero, with values ranging from 2.5e-37 to 3.8e-27. These results suggest significant limitations in these models' ability to generate answers aligning with the reference answers for this dataset. CountVectorizer scored 0.0, reinforcing the inadequacy of traditional frequency-based methods for complex QA tasks. **Comparison to Literature** Transformer-based models like RoBERTa and XLM-R have shown high accuracy in multilingual QA tasks and biomedical datasets in prior studies (Comparative Study of BERT Models and Cross-lingual QA). However, the near-zero BLEU scores for Shubham09/medqa suggest that these models may struggle with specific nuances in the dataset, such as its language patterns, question complexity, or reference answer structure. The disparity emphasizes the need for domain-specific fine-tuning or more robust pretraining data.

C. Overall Observations

The results across both datasets highlight critical differences between implemented models and their reported performances in the literature:

*a) **Dataset Sensitivity:** Models like BioBERT and BERT Large, despite their high domain relevance, underperformed on Shubham09/medqa due to possible mismatches in dataset structure or language.:*

*b) **Generalization Gap:** While existing studies report high BLEU and F-scores for these models on datasets like SQuAD or PubMedQA, the results indicate a substantial drop in performance when applied to less-structured or niche datasets.:*

*c) **Domain-Specific Adaptations:** The findings underscore the importance of domain-specific adaptations, such as pretraining on related corpora or fine-tuning on similar datasets, to improve model efficacy.:*

IV. Implementing Graph RAG

In implementation of GraphRAG, a graph-based retrieval-augmented generation (RAG) model is employed to enhance information retrieval from medical documents, specifically using a Neo4j graph database. This approach involves constructing a graph where nodes represent medical documents or concepts and edges capture the relationships between them. This structure allows for contextually relevant document retrieval by leveraging entity connections within the graph, leading to precise, evidence-based responses for medical queries. This implementation addresses the need for accurate, context-aware information in clinical applications by enhancing retrieval precision through graph modeling.

In this implementation, we transitioned from a traditional setup, where a knowledge graph and vector store operated independently, to an integrated GraphRAG (Graph-based Retrieval-Augmented Generation) framework. The enhancements focused on combining the strengths of both tools—relationship-based reasoning from the knowledge graph and similarity-based retrieval from the vector store—to generate more complete and accurate answers.

1) Initial Setup: The previous system consisted of two separate components:

Knowledge Graph: Stored relationships like causal links ("A causes B") or treatments ("X treats Y"). It allowed for structured queries and exploration of interrelated concepts. **Vector Store**: Contained embeddings for longer text passages, enabling similarity-based searches to retrieve relevant textual information. While both systems provided valuable insights, they operated independently. Answers were often incomplete as they failed to combine relational knowledge with textual details.

A. Transition to GraphRAG

To address this limitation, several changes were implemented to create an integrated GraphRAG system:

1) Combined Search:: The system now queries both the knowledge graph and the vector store simultaneously when a question is posed. This approach mirrors consulting two experts—one specializing in connections and relationships, the other in detailed contextual information—and synthesizing their knowledge for a unified answer.

2) Smart Context Mixing:: Relationships retrieved from the graph (e.g., "A causes B," "X treats Y") are combined with detailed textual information from the vector store (e.g., passages explaining symptoms or treatment protocols). The integration

ensures that answers reflect both high-level connections and in-depth explanations, making the responses more contextual and informative.

3) Enhanced Answer Generation:: The Large Language Model (LLM) used in the pipeline now processes: Graph-derived knowledge: Interrelationships and causal connections. Vector store insights: Detailed passages and supporting evidence. This dual input enables the LLM to generate comprehensive answers that are both relationally grounded and textually rich.

4) Smarter Evaluation:: A scoring system was introduced to evaluate generated answers based on: **Accuracy:** Alignment with known correct answers. **Completeness:** Inclusion of both graph relationships and vector store details. **Relevance:** Appropriateness of the response to the question posed. The evaluation framework ensures continuous improvement by comparing generated answers with benchmark solutions.

B. Benefits of GraphRAG

The GraphRAG approach represents a significant shift in how knowledge is retrieved and synthesized:

1) Old Approach:: Independent use of a knowledge graph or vector store, akin to asking a librarian or consulting a dictionary separately.

2) New Approach (GraphRAG):: Simultaneously consulting both the librarian and the dictionary, resulting in answers that reflect a holistic understanding.

By integrating these tools, the system provides more accurate, complete, and relevant responses, leveraging all available information sources effectively. This unified approach improves the user experience and reliability of the generated answers, particularly in complex domains like healthcare and biomedical research.

ACKNOWLEDGMENT

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [**?**]. Refer simply to the reference number, as in [**?**]—do not use “Ref. [**?**]” or “reference [**?**]” except at the beginning of a sentence: “Reference [**?**] was the first . . .”

REFERENCES

[1] one