Advancing Feature Extraction in Healthcare through the Integration of Knowledge Graphs and Large Language Models

Fahmida Liza Piya, Rahmatollah Beheshti
University of Delaware
210 S College Ave
Newark, DE 19716
lizapiya@udel.edu, rbi@udel.edu

Abstract

The exponential growth of unstructured medical data presents significant challenges in extracting clinically relevant features for precise and timely decision-making in healthcare. Traditional methods often falter due to the inherent noise, ambiguity, and heterogeneity of such data. To address these limitations, we propose a novel hybrid approach that integrates domain-specific Knowledge Graphs (KGs) with Large Language Models (LLMs) within a Retrieval-Augmented Generation (RAG) framework. By constructing a KG that represents clinical entities—such as patients, admissions, diagnoses, procedures, and prescriptions—as nodes, with their interrelationships as edges, we capture the complex relationships inherent in medical data. Integrating this KG with LLMs enhances Named Entity Recognition (NER) across various diseases, supporting disease-agnostic applications. Our research focuses on refining the KG through advanced entity and relation extraction techniques, ensuring accurate and comprehensive modeling of clinical information. By embedding this enriched KG within the RAG framework, we aim to improve the precision of feature extraction, thereby providing more informative insights for clinical decision-making.

Introduction

EHRs are foundational to modern medical informatics, providing unprecedented volumes of data for clinical analysis (Piya, Gupta, and Beheshti 2024). Despite their potential, the unstructured components of EHRs, such as discharge summaries and clinical notes, remain underutilized. Traditional NLP techniques face significant challenges in processing this data due to its complexity, noise, and heterogeneity.

To address these limitations, this research proposes a novel approach that integrates domain-specific KGs and LLMs within a RAG framework. By combining the organizational strengths of KGs with the contextual understanding of LLMs, this method aims to enhance the extraction and analysis of clinical features from unstructured data.

The central question this thesis seeks to answer is: How can the integration of LLMs and KGs transform the extraction and analysis of clinical data from EHRs? Through leveraging advanced AI techniques, the objective is to develop a model that not only improves the precision of data extraction but also provides enriched insights to support clinical decision-making and enhance patient care.

Related Work

Entity extraction in healthcare has been widely studied, but most approaches either focus on structured data or rely on traditional NLP methods (Lee et al. 2020; Huang, Altosaar, and Ranganath 2019). Recent advancements in Large Language Models (LLMs), such as GPT, have demonstrated potential in processing and understanding unstructured clinical text (Yang et al. 2022). However, these models often lack the domain-specific expertise required for accurate interpretation, which can be addressed by incorporating Knowledge Graphs (KGs) to improve interpretability and contextual relevance (Edge et al. 2024; Wu et al. 2023). This research builds on existing methods and proposes a new framework that synthesizes these approaches. By integrating KGs with LLMs, we can enhance Named Entity Recognition (NER) across various diseases, supporting disease-agnostic applications. This research builds upon prior work by introducing a novel framework that synergizes these advanced methodologies to address the unique challenges of clinical data analysis.

Research Objectives

The primary objective of this research is to advance the state-of-the-art in healthcare data analysis by developing a novel hybrid AI framework that synergistically leverages Large Language Models (LLMs), Knowledge Graphs (KGs), and Retrieval-Augmented Generation (RAG). Specifically, this work aims to:

- Develop a precise and efficient LLM-based model to accurately identify and extract key entities from unstructured medical data.
- Develop a domain-specific KG extract nodes, edges, labels that depicts the inherent relation between each patient and their journey from the structured data, incorporating extracted entities and their associated attributes to create a semantically rich and detailed knowledge graph that enhances model understanding.
- Utilize the KG within a Retrieval-Augmented Generation (RAG) framework to provide contextual informa-

tion, thereby improving the quality and relevance of retrieved data for key feature extraction.

- Develop an LLM-based model to extract key features from the retrieved information, utilizing the contextual insights from the KG to identify and prioritize the most relevant features for analysis, thereby enhancing disease-agnostic applications.

By integrating these components, the research aims to develop an advanced AI framework capable of extracting valuable insights from unstructured medical data, leading to improved patient care and a comprehensive understanding of various disease agnostics.

Research Progress (as of November 2024)

In November 2024, we advanced our project by acquiring and preprocessing a comprehensive clinical dataset, focusing on key tables containing patient demographics, admissions, diagnoses, procedures, and prescriptions. Utilizing Neo4j, we constructed a knowledge graph representing clinical entities and their relationships, accurately reflecting patient journeys. We initiated integration of this knowledge graph with Large Language Models (LLMs) to enhance Named Entity Recognition (NER) tasks, conducting preliminary tests to assess performance improvements. This work has been done in collaboration with my thesis supervisor, Dr. Rahmatollah Beheshti, who has guided the theoretical framework of integrating KGs with LLMs.

Future Research Plan (By February 2025)

In the next phase of the project, I will focus on refining the RAG-based approach for entity extraction and validating the Knowledge Graph's role in disease progression prediction. The following milestones are planned by February 2025:

- Implement feedback mechanisms to enhance entity extraction accuracy using domain-specific knowledge.
- Expand the Knowledge Graph and integrate it with predictive models to evaluate disease outcomes.
- Conduct experiments to assess the system's performance on unseen patient data and generalize findings to other diseases.

Anticipated Thesis Contribution

The anticipated contributions of this thesis include the development of a novel framework that integrates Knowledge Graphs (KGs) and Large Language Models (LLMs) within a Retrieval-Augmented Generation (RAG) framework for healthcare analytics. This approach aims to enhance the extraction of clinically relevant features from unstructured medical data, thereby improving the accuracy of AI systems in healthcare and contributing to better patient care. By leveraging the structured knowledge of KGs and the contextual understanding of LLMs, this work seeks to provide a pathway for more effective utilization of unstructured clinical data, facilitating precise and timely decision-making in healthcare.

Timeline

- **November 2024:** Completed the initial entity extraction model and preliminary Knowledge Graph.
- **By February 2025:** Refined RAG system, extended Knowledge Graph, experimental validation and Paper Writing.

References

Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; and Larson, J. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Huang, K.; Altosaar, J.; and Ranganath, R. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Piya, F. L.; Gupta, M.; and Beheshti, R. 2024. HealthGAT: Node Classifications in Electronic Health Records using Graph Attention Networks. *arXiv preprint arXiv:2403.18128*.

Wu, X.; Duan, J.; Pan, Y.; and Li, M. 2023. Medical knowledge graph: Data sources, construction, reasoning, and applications. *Big Data Mining and Analytics*, 6(2): 201–217.

Yang, X.; Chen, A.; PourNejatian, N.; Shin, H. C.; Smith, K. E.; Parisien, C.; Compas, C.; Martin, C.; Costa, A. B.; Flores, M. G.; et al. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1): 194.