Intelligent Clinical Assistant for Personalized Responses and Clinical Summaries

Subash Neupane

Department of Computer Science & Engineering
Mississippi State University
sn922@msstate.edu

Abstract

Healthcare information is scattered across heterogeneous data sources, such as patient medical records, clinical guidelines, research literature, and online knowledge bases. Segmented information, both structured and unstructured, when integrated together using *context augmentation* - a knowledge fusion technique, has the ability to contextualize broader medical context. Current approaches lack knowledge aggregation that is necessary to generate personalized healthcare recommendations. I propose novel AI frameworks that leverage language models and hybrid retrieval techniques to aggregate multi source knowledge, enabling the generation of contextual and accurate medical response.

Introduction

Providing contextual and comprehensive medical information tailored to individual patients is critical for enabling effective care in the healthcare domain. However, existing approaches often struggle to deliver personalized responses due to the distributed nature of medical data across multiple sources like patient records, medical literature, and online resources. Take, for example, a *patient context* (a patient experiencing a Chronic Obstructive Pulmonary Disease (COPD) exacerbation) *"What specific management strategies or interventions would you recommend to improve respiratory symptoms and overall lung function?"* answering this accurately demands comprehending the patient medical history, while also incorporating relevant information about COPD exacerbations, treatment strategies, and respiratory management from various medical knowledge sources.

While researchers have made some progress in augmenting Large Language Models (LLMs) with external knowledge in the healthcare domain, they often rely on a single knowledge source or focus on specific medical tasks. Current research lack knowledge fusion techniques to create a broader medical context. To address this, we introduce *context augmentation* (Neupane et al. 2024). Context augmentation is a technique of aggregating disparate multi-source knowledge base both structured and unstructured leveraging dense retrieval techniques to establish a full spectrum of medical context. Augmented medical context when passed to an inference engine has the ability to generate contextually relevant and accurate responses. This in turn, also reduce the risk of hallucination, an open research problem inherent to LLMs.

Another aspect of my research is investigating the possibility of generating clinical summaries based on unstructured patient-doctor conversation using fine-tuned LLMs. I hypothesize that these summaries offer benefits to both patients and providers. For healthcare providers, such summaries reduces burnout, while for patient it improves patients' understanding of care plans.

The bulk of my dissertation will focus on developing efficient hybrid retrieval techniques (sparse and dense) and leveraging language models to address context augmentation and clinical summarization tasks. Additionally, I will explore knowledge representation techniques to model the patient journey as a reasoning layer. By utilizing knowledge graphs as reasoning layers, the system can infer relationships and validate the consistency of outputs, ensuring both safety and trustworthiness in clinical decision-making.

Existing Challenge and Dissertation Goal

Current AI systems in healthcare remain limited in generating personalized healthcare recommendation due to their reliance on a single knowledge source. The task of response generation in these systems can be mapped as $M : (Q, KB) \rightarrow A$ where, M is prediction model (either rule based or neural), Q is the question, KB represent the single knowledge base, and A is the answer. As I explained in preceding sections, healthcare information is scattered and fragmented multiple sources and there is a pressing need of knowledge fusion techniques to capture full spectrum of a patients' context. For example, consider a hypothetical patient context, John Smith, a 50-year-old male with type 2 diabetes, hypertension, and a sedentary lifestyle. A consultation note during his medical encounter contains information about his symptoms, diagnosis, treatment, concerns and care plans in unstructured format. To provide contextual and personalized recommendation, a system must have the ability to fuse multiple knowledge source together such as Mr.Smith's medical history C_p (patient context), up-to-date clinical guidelines C_g, and trusted medical text books C_m.

In order to address this challenge, my dissertation introduces *context augmentation*, a technique that integrates

multiple knowledge sources ($C_p \cup C_g \cup C_m$) to formulate a broader patient-specific augmented medical context A_c. This augmented medical context replaces the traditional knowledge base KB, allowing the inference engine (model) to generate personalized and contextual responses, as represented by $M : (Q, A_c) \rightarrow A$.

An additional focus of my dissertation addresses the challenge of generating clinical summaries in Subjective Objective Assessment Plan (SOAP) format from patient-doctor conversations, which remains the open research problem. Current approaches, like those using extractive and abstractive methods (Krishna et al. 2020), show progress but struggle with the noisy and unstructured nature of doctor-patient conversations (Zhang et al. 2021), resulting in summaries that may lack accuracy or fail to capture the full scope of the medical interaction (Giorgi et al. 2023). My research aims to address these limitations by introducing a hybrid approach that combines *retrieval-based filtering* with *fine-tuned models*. This approach sanitizes and refines the noisy input data before summarization, ensuring that the generated summaries notes are not only accurate but also relevant to the specific medical conversation.

Intelligent Clinical Assistant Construction

My dissertation includes the following Research Questions (RQ):

RQ1*: Can context augmentation improve healthcare recommendations?* – We propose MedInsight (Neupane et al. 2024), a novel multi-source context augmentation framework that aggregrate and fuse diverse knowledge sources such as patient-doctor conversations, medical history, clinical guidelines, and other relevant knowledge sources. Knowledge fusion with context augmentation technique enables intelligent clinical assistant AI systems to generate contextual, targeted, and personalized recommendations.

RQ2*: Is it possible to automatically generate clinical summaries from unstructured patient-doctor conversations?* Generating clinical summaries from unstructured conversations is challenging due to the variability of medical interactions. We propose a two-module architecture: retriever based filtering and inference. The first module sanitizes, filters and extracts relevant SOAP elements from the conversation, while the second module utilizes a fine-tuned language model as inference engine to generate clinical summaries based on the extracted information. This framework will be evaluated across various medical specialties to assess its effectiveness in generating accurate clinical summary, using quantitative metrics and qualitative assessments by Subject Matter Experts (SMEs).

RQ3*:Does integrating hybrid retrieval methods with context augmentation further improve healthcare recommendation?* The core of my research focuses on developing a hybrid retrieval approach i.e. combining dense and sparse retrieval methods to improve the likelihood of accurate and personalized response generation. I hypothesize that combining dense retrieval, which captures semantic similarities, with sparse retrieval, which focuses on keyword matching and term importance, will not only aid in understanding broader medical contexts but also ensure that critical, domain-specific terms are not overlooked. RQ3 will build on the context augmentation technique discussed in RQ1 and utilize the clinical summaries discussed in RQ2.

Objective	Timeline
RQ1	Sep 2023 - Mar 2024
RQ2	Apr 2024 - Dec 2024
RQ3	Dec 2024 - Mar 2025
Dissertation Development	Mar 2025 - May 2025

Table 1: Research Timeline.

Preliminary Work and Research Timeline

My preliminary work (Neupane et al. 2024) has focused on developing novel context-augmentation framework that has ability to generate personalized healthcare recommendation which aligns with RQ1. This work has been submitted to ACM Transactions (Healthcare Computing) and is currently under review. Additionally, I am investigating methods to generate clinical summaries, which will contribute to RQ2. I will also begin exploring hybrid retrieval methods, central to RQ3, which aim to enhance the accuracy and context-awareness of responses. These research efforts will be submitted to a high-impact AI conference. My research timeline is outlined in Table 1.

References

Giorgi, J.; Toma, A.; Xie, R.; Chen, S. S.; An, K. R.; Zheng, G. X.; and Wang, B. 2023. Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models. *arXiv preprint arXiv:2305.02220*.

Krishna, K.; Khosla, S.; Bigham, J. P.; and Lipton, Z. C. 2020. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. *arXiv preprint arXiv:2005.01795*.

Neupane, S.; Mitra, S.; Mittal, S.; Golilarz, N. A.; Rahimi, S.; and Amirlatifi, A. 2024. MedInsight: A Multi-Source Context Augmentation Framework for Generating Patient-Centric Medical Responses using Large Language Models. *arXiv preprint arXiv:2403.08607*.

Zhang, L.; Negrinho, R.; Ghosh, A.; Jagannathan, V.; Hassanzadeh, H. R.; Schaaf, T.; and Gormley, M. R. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. *arXiv preprint arXiv:2109.12174*.