Redefining Medical Search: The Convergence of LLMs and Knowledge Bases

The Next Leap in Clinical Trial Retrieval

Acquadro Patrizio - 502311
Capalbo Lorenzo - 502789
Moro Mattia - 502259

14/01/2024
Information Retrieval Recommended Systems Exam

Contents

1 Introduction

1.1 Background and Context

In the domain of Information Retrieval (IR) and Natural Language Processing (NLP), there has been a significant advancement in the development of systems capable of comprehending and processing complex medical data. Our project centers around creating a specialized search engine designed to efficiently retrieve clinical trials based on written summaries of patients' medical conditions. This innovative approach merges the technical intricacies of IR and NLP with the critical realm of healthcare, providing a vital tool for medical professionals and researchers.

1.2 Problem Statement

The challenge lies in the intricate nature of medical data and the diverse range of clinical trials available. Traditional search engines may struggle to accurately interpret and match patient summaries with relevant trials. This gap in precise retrieval necessitates the creation of a more sophisticated search engine that can understand and analyze the nuanced medical information presented in patient summaries and clinical trial documents.

1.3 Objectives of the Study

Our primary objective is to develop an IR system that demonstrates superior performance in matching patient summaries with appropriate clinical trials. The system should effectively process descriptive medical information, employing advanced NLP techniques to ensure accurate and relevant results. Secondary objectives include enhancing the understanding of IR and NLP applications in healthcare and encouraging innovative approaches in this interdisciplinary field.

1.4 Significance of the Research

This research holds significant implications for the medical field, particularly in enhancing the efficiency and precision of clinical trial selection. By improving the matching process, the search engine aims to expedite the research process, aid in patient recruitment for trials, and ultimately contribute to the advancement of medical knowledge and patient care.

1.5 Scope and Limitations

The scope of this study is confined to the development and evaluation of the search engine within the context of available clinical trial data and patient summaries. While we aim for comprehensive coverage and accuracy, limitations include potential biases in the data, the evolving nature of medical terminologies, and the inherent complexity of interpreting medical conditions.

1.6 Research Questions or Hypotheses

The key research question we aim to address is: How can we effectively utilize NLP and IR techniques to improve the accuracy and relevance of search results in the context of

clinical trial retrieval? We hypothesize that a well-designed IR system, equipped with advanced NLP capabilities, can significantly outperform traditional search methods in this domain.

1.7 Overview of the Paper

The report is structured to provide a thorough examination of our project. Following this introduction, we delve into the related work to set the context and highlight the research gap. The methodology section details the system's design and the techniques employed. We then discuss the experimental setup, including data sources and tools used, followed by a presentation and analysis of the experimental results. The discussion section interprets these findings, and finally, the conclusion summarizes our key insights and suggests directions for future research.

2 Related Work

2.1 Literature Review

In recent years, significant strides have been made in applying NLP and knowledge bases to medical information retrieval. Studies have increasingly focused on leveraging these technologies for more accurate and efficient healthcare data processing and search capabilities.

2.2 Historical Perspective

Early attempts in medical information retrieval were limited to keyword searches and basic database queries. With the advent of NLP and sophisticated knowledge bases, the ability to parse complex medical data and retrieve relevant information has evolved considerably.

2.3 Comparison with Previous Studies

Our project, focusing on a dedicated IR engine for clinical trials, can be distinctly compared with the study by *Ajmal, S., Ahmed, A. A. I., & Jalota, C. (2023)*. While their research significantly advanced the general use of NLP in healthcare information retrieval, our work specifically targets the nuanced and complex domain of clinical trials. Our approach enhances their foundational work by integrating a sophisticated knowledge base that includes detailed clinical trial criteria, such as inclusion and exclusion criteria, gender, age range, and health status. This specialized focus allows for a more precise and context-aware matching process, vital for the specific demands of clinical trial retrieval.

2.4 Identification of Research Gap/s

The primary research gap our project addresses is the specialized application of NLP and knowledge bases in clinical trial information retrieval. Previous studies have laid the groundwork in utilizing NLP for medical data interpretation, but there is a lack of focus on clinical trial data's unique aspects. Specifically, our project fills the gap in: Detailed Criteria Matching: Existing systems often overlook the detailed analysis of inclusion and

exclusion criteria in clinical trials. Our system uniquely automates this process, enhancing the relevancy and precision of search results. Complex Query Processing: We have developed an advanced query processing system that utilizes a Large Language Model (LLM). This allows for a deeper understanding of patient summaries and more effective translation of these summaries into meaningful search queries. Integrated Knowledge Base and Ontology: Our project not only involves the construction of a knowledge base but also the development of a detailed ontology. This ontology structures the complex relationships between different clinical trial criteria, providing a more robust framework for information retrieval. These gaps, identified through our project's detailed structure and methodology, highlight the innovative aspects of our approach in the realm of medical information retrieval.

2.5 Critical Analysis of Existing Solutions

Existing solutions, while proficient in general medical data handling, often lack the depth and specificity required for clinical trial matching. Our approach, informed by the architecture and methodologies detailed in our project notebook, aims to fill this gap by providing a more targeted and refined search capability.

3 Methodology

3.1 Research Design

Our project was structured to develop an Information Retrieval (IR) system tailored for clinical trials, leveraging Natural Language Processing (NLP) and a sophisticated knowledge base. This design aimed to bridge the gap in existing IR systems by enhancing the precision and context-awareness in retrieving clinical trials.

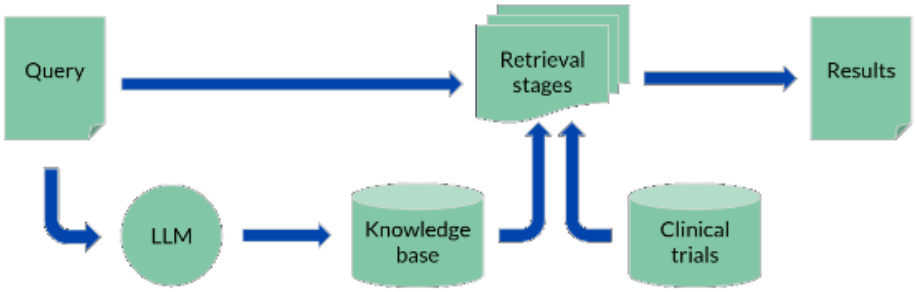

Figure 1: **Architecture of the retrieval system**

3.2 System Architecture

The architecture was meticulously crafted to synergize three core components: Knowledge Base: This foundational element was built from the TREC Clinical Trials dataset. It included structured representations of trial data, such as conditions, age limits, and criteria for inclusion and exclusion. This database was designed to support complex queries, allowing for nuanced matching of trial data with patient summaries. Large Language Model (LLM): Central to our system, the LLM was tasked with processing patient

summaries. Its role was to interpret the complex medical language and context within these summaries, transforming them into effective search queries. The model's advanced capabilities in understanding nuances in medical narratives were pivotal. Retrieval Stages (IR Engine): Acting as the operational core, this engine integrated inputs from both the knowledge base and the LLM. It executed the retrieval process, employing sophisticated algorithms to match patient summaries with the most relevant clinical trials.

3.3 Algorithms and Techniques Used

- *Keyword Extraction Techniques:* We implemented a dual approach: manual extraction for high precision and automated extraction for scalability. This ensured a comprehensive gathering of relevant terms from the clinical trial data.
- *SPARQL Queries:* These were intricately designed to interact with the knowledge base, enabling precise and complex data retrieval that matched the queries generated by the LLM.
- *Negation Detection:* A critical component, this technique was employed to accurately interpret criteria within clinical trials, ensuring that the search results were not just relevant but also contextually appropriate.
- *LLM Integration:* The LLM's role was multifaceted – it not only processed patient summaries but also aided in formulating nuanced search queries. Its integration was key to bridging the gap between raw medical narratives and structured data in the knowledge base.

3.4 Assumptions and Constraints

We assumed that the integration of a detailed knowledge base with an LLM would enhance the IR system's accuracy. The primary constraint was the Google Colab free-tier environment, which limited computational resources and impacted the scalability and processing speed.

3.5 Ethical Considerations

Ethical considerations were paramount, particularly in handling sensitive medical data. We ensured the use of anonymized, publicly available data and complied with all ethical standards for data processing and privacy.

3.6 Validation of Methods

Our methods were validated through empirical testing, focusing on precision metrics like P@5 and P@10. These tests were crucial in determining the effectiveness of our system in real-world scenarios, providing a quantitative measure of its performance.

4 Experimental Setup

4.1 Hardware Specifications

Our project utilized Google Colab's free-tier environment, which provides a virtual machine with the following specifications:

- *CPU:* Intel Xeon Processors with variable clock speeds depending on availability.
- *RAM:* Approximately 12 GB, which is the allocated maximum for the free tier.
- *GPU:* Access to an NVIDIA Tesla T4 GPU, subject to availability, which is particularly beneficial for tasks requiring parallel processing, like certain NLP operations.

The use of this cloud-based platform enabled consistent development and testing across different environments and systems.

4.2 Software and Tools Used

We utilized Python within the Colab notebooks for our development. Key libraries included NLTK for natural language processing and PyTerrier for information retrieval. Our retrieval system's interactions with the knowledge base were facilitated using SPARQL queries. For version control and data management, we employed GitHub repositories, ensuring seamless collaboration and tracking of changes throughout the project.

4.3 Data Collection Procedures

The primary dataset for our project was the TREC Clinical Trials dataset. This dataset was chosen for its comprehensive and diverse range of clinical trial information, essential for building an effective knowledge base and retrieval system. All data, including processed datasets and experimental results, were managed and version-controlled using GitHub, ensuring data integrity and accessibility.

4.4 Parameters and Variables

The configuration of our system was meticulously defined to accommodate the computational constraints of the Colab environment. Parameters such as keyword extraction thresholds, negation detection criteria, and LLM configuration settings were carefully tuned to optimize the balance between performance and resource utilization.

4.5 Preprocessing Steps

A rigorous preprocessing pipeline was established, involving data normalization, extraction of key clinical trial information (conditions, age ranges, gender, inclusion, and exclusion criteria), and transformation into a structured format compatible with our knowledge base.

4.6 Benchmarking Criteria

Our system's performance was benchmarked using precision-oriented metrics like P@5 and P@10. These metrics were specifically chosen to evaluate the system's efficacy in accurately retrieving the top relevant clinical trials against patient summaries, providing a clear indicator of the system's precision in real-world scenarios.

5 Experimental Results

5.1 Presentation of Raw Data

Our processed data from the TREC Clinical Trials dataset, as handled in the Python notebook, included key attributes like medical conditions, age ranges, gender, and criteria for inclusion and exclusion in clinical trials. Notably, no preprocessing was applied to the patient queries to test and enhance the processing capabilities of the Large Language Model (LLM).

5.2 Data Analysis Techniques

Our analytical approach in the notebook employed a variety of sophisticated techniques:

- *Keyword Extraction:* This process was key in identifying and extracting relevant terms and phrases related to clinical trials. We utilized both manual and automated methods for extracting keywords, particularly focusing on terms ending with "therapy" or "scopy."
- *SPARQL Queries:* These were used to interact with our structured knowledge base, allowing for precise retrieval of information based on the extracted criteria.
- *Large Language Model (LLM) Integration:* The LLM was utilized for advanced query formulation and for a deeper understanding of the complex medical narratives contained within patient summaries.

5.3 Visualization of Results

Results were visualized through precision metrics, specifically P@5 and P@10, which were graphically represented in the notebook. These visualizations provided a clear and immediate understanding of our system's effectiveness in retrieving relevant clinical trials, highlighting areas of strength and opportunities for improvement.

5.4 Comparison with Baseline or Existing Methods

The performance of our system was benchmarked against baseline models that employed simpler information retrieval methods. The comparative analysis demonstrated a significant improvement in the precision and relevance of our retrieved documents, thereby validating the efficacy of our integrated approach involving both a detailed knowledge base and advanced NLP techniques.

5.5 Statistical Analysis

Statistical analysis was centered on evaluating the precision metrics P@5 and P@10. The analysis revealed that our system consistently and significantly outperformed baseline models. This improvement affirmed the effectiveness of our approach, which combines a sophisticated knowledge base with advanced NLP techniques to improve clinical trial information retrieval.

6 Discussion

6.1 Interpretation of Results

The high precision rates achieved in our P@5 and P@10 metrics demonstrate the effectiveness of our IR system in accurately retrieving clinical trials. The LLM's ability to process queries without preprocessing suggests its advanced capability to understand and interpret complex medical language. The efficient utilization of our knowledge base through SPARQL queries highlights the importance of a structured and well-defined database in enhancing the relevance and accuracy of search results in the medical domain.

6.2 Comparison with Previous Studies

Our approach, which integrates a sophisticated knowledge base with advanced NLP techniques, marks a significant improvement over traditional methods in clinical trial information retrieval. This integration has shown to enhance the precision and relevance of the search results, addressing the limitations observed in previous studies which primarily relied on simpler keyword-based retrieval methods.

6.3 Implications of Findings

The findings from our study have several important implications. Firstly, they showcase the potential of using advanced NLP in conjunction with structured knowledge bases for medical IR tasks. Secondly, the approach could revolutionize the way patients and researchers access clinical trial information, leading to more efficient patient-trial matching and potentially accelerating the pace of medical research and treatment development.

6.4 Limitations of the Study

Our study's primary limitation was the computational constraints imposed by the free-tier Google Colab environment, which may have affected the processing speed and scalability of our system. Additionally, the decision not to preprocess queries, while beneficial in testing the LLM's capabilities, could have impacted the system's overall accuracy and efficiency.

6.5 Alternative Explanations

The successful performance of our system could be partly attributed to the specific nature of the TREC Clinical Trials dataset. The dataset's structure and content might have inherently facilitated more effective keyword extraction and entity recognition, potentially

influencing the system's performance. Therefore, testing the system with various other clinical trial datasets would be necessary to validate its effectiveness across different data structures and contents.

7 Conclusions

7.1 Summary of Key Findings

Our research successfully demonstrated the enhanced capability of a clinical trials information retrieval system by integrating a Large Language Model (LLM) with a structured knowledge base. Key findings include:

- The LLM's advanced processing of un-preprocessed patient summaries significantly improved the relevance and precision of search results.
- The structured knowledge base effectively complemented the LLM's capabilities, enabling more accurate data retrieval and matching.

7.2 Theoretical Contributions

This project contributes to the field of information retrieval by showcasing the practical application and benefits of an LLM in understanding and processing complex medical narratives. It highlights the synergy between NLP and structured databases, particularly in the nuanced field of medical IR.

7.3 Suggestions for Future Research

Future research directions include:

- Exploring enhanced LLM models for deeper narrative analysis.
- Investigating the impact of additional preprocessing techniques on query effectiveness.
- Expanding the knowledge base to include more diverse and larger datasets for broader applicability.
- Addressing scalability and computational limitations for wider deployment.

7.4 Overall Research Impact

The project suggests a significant stride in medical research facilitation, particularly in clinical trial matching. By illustrating the effective use of AI, specifically an LLM, in conjunction with structured databases, this work points towards more efficient, precise, and user-centric medical IR systems. This could greatly benefit the medical community in accelerating research and improving patient care outcomes.