Improved Search for Text Rich Knowledge Graphs

University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath

Dr. Rob Brennan[1] Prof. Gareth Jones [2] Swaraj Shaw

[1]School of Computer Science, University College Dublin [2]Dublin City University

Introduction

Pharmaceutical companies and other healthcare organizations are increasingly adopting knowledge graphs to store and analyze complex health records. This allows them to make better decisions, identify trends and patterns, and improve patient care.

In this research proposal, we will focus on developing new techniques and solutions for using natural language processing to improve the ranking of entities in knowledge graphs. We will perform a comparative study on graph-based ranking and embedding-based ranking.

Research Objective

The specific objectives of this research are to:

1. Develop new algorithms for ranking entities in knowledge graphs based on the text associated with them.
2. Create new tools for comparing knowledge graphs to each other and to unstructured text sources.
3. Develop new methods for identifying patterns in knowledge graphs that are relevant to safety management or other healthcare applications.
4. Work with industry partners to apply the research findings to real-world problems.

Methodology

The methodology for this project will be divided into three phases:

1. **Data collection and preprocessing**: The first phase will involve collecting and preprocessing a dataset of textually rich knowledge graphs describing healthcare settings. The data will be preprocessed to remove noise and inconsistencies.
2. **Model development**: The second phase will involve developing new methods and tools for graph ranking, comparison, and mining insights from knowledge graphs. I will leverage a mix of information retrieval, natural language processing, knowledge graphs, and user-centric systems research techniques.
3. **Evaluation and validation**: The third phase will involve evaluating and validating the proposed methods and tools. I will use a variety of evaluation metrics like BLEU Score, Accuracy, Precision Recall.
   I will follow a rigorous research process and will adhere to the highest ethical standards. I will make all data and code publicly available to ensure reproducibility of the research findings.

Proposed Architecture

The proposed architecture has two phases: data extraction and data integration. In the first phase, data is extracted from healthcare sources, such as electronic health records (EHRs), clinical trial data, and diagnosis codes. In the second phase, the data is integrated into a knowledge graph using graph traversal and ETL. Machine learning is used to reason over the data and extract insights.

The architecture is designed to be scalable and adaptable, so it can be used to build knowledge graphs of different sizes and with different types of data. It also uses a variety of machine learning techniques to address the challenges of building a healthcare knowledge graph, such as data silos, data sparsity, and the cold start problem.

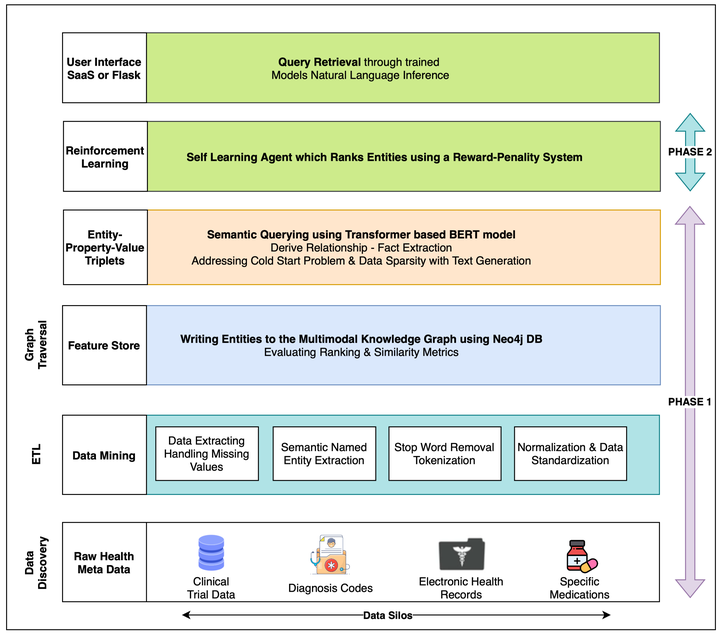

Figure 1. A flexible and scalable framework for building knowledge graphs for healthcare settings.

Implementation Approach

The implementation of above architecture addresses few shortcomings of BioBLP. These techniques include:

1. Using a single knowledge graph that integrates information from multiple sources: This will make it easier to find and integrate information from different sources.
2. Using a self-learning agent that ranks entities based on their relevance to user queries: This will address the cold start problem and improve the ranking of entities.
3. Using data mining techniques to identify patterns and trends in the data: This will help to fill in the gaps in the data and improve the accuracy of the inferences.

Advantages:

1. **Scalability**: The proposed architecture is designed to be scalable, so it can be used to store and manage large amounts of data.
2. **Reliability**: The proposed architecture is designed to be reliable, so it can be used to provide users with accurate and up-to-date information.
3. **Interoperability**: The proposed architecture is designed to be interoperable, so it can be used to integrate with other systems and applications.
4. **Graph traversal**: This will be used to find entities that are related to each other, or to find paths between entities.
5. **ETL**: This will be used to extract data from a variety of healthcare sources, such as electronic health records (EHRs), clinical trial data, and diagnosis codes. The data will then be transformed into a format that can be used to create a knowledge graph.
6. **Data discovery**: This will be used to identify the different types of data that are available, and to understand the relationships between the data. This information will be used to create a knowledge graph that is more accurate and comprehensive.
7. **User interface**: This will be the graphical interface that users will use to interact with the knowledge graph. It should be easy to use and navigate, and it should provide users with a variety of ways to search and explore the knowledge graph.
8. **Reinforcement learning**: This will be used to rank entities in the knowledge graph. The agent will learn to rank entities that are more relevant to user queries by receiving rewards for ranking entities correctly and penalties for ranking entities incorrectly.
9. **Entity-property-value triplets**: These are the basic units of information in a knowledge graph. They consist of an entity, a property, and a value. For example, the triplet "John Doe" - "has age" - "30" would represent the fact that John Doe is 30 years old.

Expected Timeline

Table 1. Expected timeline for the development of a healthcare knowledge graph

Phase	Duration (months)
Phase 1: Data collection and preprocessing	3
Phase 2: Knowledge graph construction	4
Phase 3: User interface development	1
Total	8

References