Knowledge Graph-Driven Neuro-Symbolic System for Intelligent Document Matching

Jans Aasman[1]

[1]*Franz Inc. - AllegroGraph*

Abstract

This paper presents a novel Neuro-Symbolic framework for intelligent document matching that integrates Knowledge Graphs with Large Language Models (LLMs) to address challenges in various domains, including healthcare, aircraft maintenance, and legal documentation. Traditional methods relying solely on taxonomies face limitations due to diverse document authorship and the complexity of semantic searches. The proposed approach leverages the reasoning capabilities of Knowledge Graphs, the semantic richness of taxonomies, and the adaptive retrieval strengths of LLMs. This combination enhances precision, reduces costs, and facilitates the automated matching of documents by efficiently managing embeddings within a vector database. The framework demonstrates significant improvements in data management and insight generation, with potential applications across multiple industries.

Keywords

Knowledge Graph, Neuro-Symbolic AI, LLM, UMLS

1. Introduction

Automated and intelligent matching of information in Knowledge Graphs, which contain various types of documents, is a crucial and prevalent use case across numerous domains. Common examples include automated systems that locate the relevant legal policy documents based on actual police reports, software that connects maintenance records for aircraft with governmental policy documents, aircraft repair manuals, or prescribed maintenance checklists, as well as automated systems that match clinical trials with suitable patients for improved healthcare outcomes.

Traditionally, taxonomies have played a crucial role in working with unstructured text within Knowledge Graphs. They are typically used to enable semantic search, enrich documents through entity and relation extraction, and, in the context of deployment experience, assist with intelligent matching between documents.

We routinely work on use cases similar to those described above. In these scenarios, we are fortunate to have comprehensive SKOS and OWL taxonomies that cover almost every aspect of the domain, such as UMLS for medical use cases or the FAA ontology for aircraft maintenance. However, we increasingly find that using taxonomies alone for intelligent document matching presents many challenges, especially when precision is critical, and it is too costly to involve humans in the process.

2. Industry Challenges

The first challenge we experience in industry use cases is that the different types of documents are authored by individuals with diverse backgrounds and perspectives, leading to varied terminologies. For example, in the context of aircraft, FAA policies are written by policymakers who incorporate legal and safety perspectives into their documentation. In contrast, mechanics performing maintenance approach the task from a procedural and technical standpoint. Similarly, in the healthcare field, clinical trials are authored by researchers or scientists who focus on specialized medical phenomena, whereas

Posters, Demos, and Industry Tracks at ISWC 2024, November 13–15, 2024, Baltimore, USA

jans.aasman@franz.com (J. Aasman)

https://franz.com/ (J. Aasman)

clinical notes are written by doctors and nurses who consider the patient holistically from a clinical care perspective.

The second challenge is the lack of time and budget to create all the necessary altLabels and extraction rules for the concepts we need to find or extract.

The third challenge is the risk of becoming too general when conducting semantic searches in a knowledge graph—deciding when to stop going up the 'skos:broader' chain or deeper down the 'skos:narrower' chain. For instance, UMLS includes taxonomies like ICD10, LOINC, MeSH, and SNOMED CT, which cover similar phenomena but were developed for markedly different purposes.

3. Knowledge Graph Driven Neuro-Symbolic Solutions

To address these challenges in our commercial projects, we created a Neuro-symbolic framework for intelligent document matching that respects the importance of taxonomies but also uses 'LLM embeddings' an equal partner in our efforts to solve matching problems.

In our PatientGraph[1,2] project we utilize this Knowledge Graph driven Neuro-symbolic approach to integrate and analyze complex EMR and biomedical data. This approach combines the reasoning strength of LLMs, the detailed semantic understanding provided by Knowledge Graphs, and the adaptive information retrieval prowess of Retrieval Augmented Generation (RAG) models. By blending symbolic reasoning with deep learning, PatientGraph not only captures the explicit knowledge contained within the EMR and biomedical data but also infers new knowledge, enabling a more intuitive exploration and analysis process. This approach significantly streamlines the management, exploration, and interpretation of vast amounts of biomedical data, opening the door to discovering new insights and opportunities in the Healthcare field.

The PatientGraph solution demonstrates several critical aspects of embedding storage and utilization within a Knowledge Graph, closely integrated with a vector database. The solution comprises the following key components:

1. Efficiently storing embeddings for taxonomy concepts within the Knowledge Graph, leveraging a vector database for optimal performance.
2. Enhancing the vector embeddings with precise metadata for entity types, significantly improving precision.
3. Employing a Large Language Model (LLM) to extract relevant terms and phrases from unstructured text.
4. Generating embeddings for these extracted terms and phrases, while also incorporating metadata about these terms into the vector store in Knowledge Graph.
5. Matching terms and phrases against existing taxonomy embeddings to ensure consistency and accuracy.
6. Comparing and aligning taxonomy terms with each other to maintain the integrity and coherence of the taxonomy.

We posit our Knowledge Graph driven Neuro-symbolic approach delivers higher precision at a much lower cost.

For future work, we plan to extend this solution to apply to many different domains. The first versions were based on healthcare, aircraft maintenance, and legal documents, but we see a way forward to make this an automated document matching architecture with minimum cost for creating large domain taxonomies.

The integration of Neuro-symbolic AI with Knowledge Graphs offers a solution to the complex challenges of intelligent document matching. This approach demonstrates significant potential in various domains, especially healthcare, by enhancing data management, enabling new insights, and improving the accuracy and efficiency of document matching processes.

References

[1] OpenText Hacknosis Announcement - https://www.prnewswire.com/news-releases/just-what-the-doctor-ordered-2024-opentext-hackathon-to-focus-on-sustainability-in-healthcare-302195540.html

[2] R. Bajracharya, R. Wallace, J. Aasman and P. Mirhaji, "Entity Event Knowledge Graph for Powerful Health Informatics," 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI), Rochester, MN, USA, 2022, pp. 456-460, doi: 10.1109/ICHI54592.2022.00068.