Leveraging the Power of Complex Logical Query Answering with Natural Language

Mariana Meireles and Carsten Fortmann-Grote

1 Author Information

Career Stage: Mariana Meireles is a Researcher Software Engineer at the Max Planck Institute for Evolutionary Biology funded by the NFDI to work with ontologies. Dr. Carsten Fortmann-Grote is the Head of Scientific Computing Unit at the same institution.
Technical Skills: Deep learning, Knowledge Graphs (KGs), Large Language Models (LLMs) and the construction of biological ontologies.

2 Abstract

This project proposes a new approach to make ontologies more accessible to non-technical users. It combines the state of the art of Zero-Shot Logical Query Reasoning (ZSLQR)[1] with a natural language interface powered by Large Language Models (LLMs). By integrating ZSLQR and LLMs, we create an intuitive interface that translates natural language inputs into complex ontological queries.

Challenges: Develop a user-friendly interface that allows non-technical researchers to easily access and utilize biological ontologies without requiring knowledge of specialized the Complex Logical Query Answering (CLQA) language.

3 Methods

Ontologies are widely used across various fields, including by professionals who may not have a strong technical background. Our system aims to allow these users to perform sophisticated queries, equivalent to those possible with CLQA, without needing to learn CLQA's specialized language. This project consists of four parts: the integration of our ontology with ZSLQR, the generation of synthetic data based on valid CLQA inputs to our data, the fine-tuning of an open-source LLM model and the development of an user interface that ties all these parts together.

Public graph databases: Starting from a de-novo hybrid assembly and manual genome annotation [2], we curated a graph database [3] containing all known genomic features, their products and functions as well as cross references to domain specific databases. Though this is a specific use case the zero knowledge approach combined with the generic algorithms we will release for fine-tuning the LLM should make our results reproducible in any ontology.

LLM Fine-tuning: We will fine-tune an open-source language model to understand queries in the CLQA language. For that, the training process will be supplemented with synthetic data generation to improve the model's ability to handle real-world queries effectively.

Interface Development: We plan to adapt the Sparnatural[4] interface, integrating our LLM tool to provide an additional layer of accessibility. Users will have the flexibility to pose questions using natural language via our LLM tool or construct SPARQL queries visually through Sparnatural.

4 Goals

Through our participation in the Summer School, we aim to learn more about network research methodologies, with a particular focus on ensuring our data is high-quality. Additionally, we hope to learn about different approaches at the intersection of KGs and LLMs.

References

[1] et al. Galkin. Zero-shot logical query reasoning on any knowledge graph. *arXiv preprint arXiv:2404.07198*, 2024.

[2] Carsten Fortmann-Grote, Eric Hugoson, Joanna Summers, Loukas Theodosiou, and Paul B. Rainey. Genome Update for Pseudomonas fluorescens Isolate SBW25. 0(0):e00637–22.

[3] Carsten Fortmann-Grote and Paul B. Rainey. From genome annotation to knowledge graph: The case of pseudomonas fluorescens sbw25. In *30th Conference on Intelligent Systems for Molecular Biology 2022*, Bio-Ontologies COSI. International Society for Computational Biology, Zenodo.

[4] Thomas Francart. Sparnatural: a visual knowledge graph exploration tool. In *European Semantic Web Conference*, Cham, 2023. Springer Nature Switzerland.