Quantum Annealing for Knowledge Graph Alignment (Proposal)

Nuno Batista[1] Francisco Lapa Silva[1]

[1]University of Coimbra, Portugal

nuno.batista@uc.pt,
francisco.lapamsilva@gmail.com

1 Problem to Tackle

Knowledge Graphs (KGs) have become essential for structuring information, but a significant challenge lies in integrating knowledge from heterogeneous sources. General-purpose KGs, often built from encyclopedic sources such as Wikipedia, provide broad, foundational knowledge but lack the technical depth found in specialized domains. On the other hand, KGs constructed from scientific literature, such as arXiv papers, contain deep, domain-specific information but may lack general context. In rapidly advancing fields such as quantum computing, fusing these two types of knowledge is critical for comprehensive understanding and discovery. This project tackles the problem of *Knowledge Graph Alignment (KGA)*: the task of identifying and linking entities that refer to the same real-world object across two different KGs, one built from Wikipedia and another from arXiv. This is a non-trivial task due to significant differences in terminology, structure, and granularity between the sources.

2 Goals of the Project

The primary goals of this project are to:

1. Construct two KGs for quantum computing: one from Wikipedia (`kg_wiki`) and one from arXiv (`kg_arxiv`).
2. Design and implement a novel hybrid classical-quantum pipeline to align the entities between these two KGs.
3. Demonstrate the feasibility of using a quantum annealer to solve the KGA problem by reformulating it as a global combinatorial optimization task.
4. Produce a single, unified KG that is more comprehensive than its individual parts.

3 Proposed Approach: A Hybrid GNN-QUBO Pipeline

We propose a two-stage pipeline that leverages the strengths of both classical deep learning and quantum optimization.

3.1 Stage 1: Classical Representation Learning with Graph Neural Neworks

We will use SciBERT to extract triples from both texts and build the graphs `kg_wiki` and `kg_arxiv` . Subsequently, a Graph Neural Network (GNN) will be trained on these graphs to generate high-quality entity embeddings.

3.2 Stage 2: Quantum Global Optimization with QUBO

Our core novelty is to frame KGA as a global optimization problem. The entity alignment task is an instance of the NP-hard Quadratic Assignment Problem (QAP). We will reformulate this as a Quadratic Unconstrained Binary Optimization (QUBO) problem. The QUBO model will be constructed using the GNN-generated embeddings to define the objective function, rewarding both entity similarity and structural consistency. This QUBO instance will then be solved on a D-Wave quantum annealer, which finds optimal or near-optimal solutions to complex combinatorial problems.

4 Data and Tools

- *Data*: Core wikipedia articles on quantum computing (e.g., "Quantum computing", "Qubit") as well as arXiv abstracts from the `quant−ph` category related to the same core concepts
- *Knowledge Extraction and Management*: `Python` libraries such as `wikipedia` , `arxiv` , Hugging Face transformers (for `SciBERT`), `RDFLib` , `Protege` .
- *Classical Machine Learning*: `PyTorch` for GNN implementation.
- *Quantum Optimization*: `D−Wave Ocean SDK` to use the quantum annealer, `pyqubo` for QUBO formulation.

5 Connections Reviewed Literature

Mainstream KGA research focuses on embedding-based methods (like GNNs) or symbolic reasoning, but overlooks global optimization frameworks. This project is inspired by the observation that graph matching can be formulated as a QAP, and that a wide variety of NP-hard combinatorial optimization problems can be mapped to QUBOs. Our approach is novel for bridging this gap, applying the QUBO model to KGA and leveraging a quantum annealer as a solver, a method not present in existing KGA surveys. This hybrid classical-quantum structure is itself a key research area aiming to gain practical benefits from today's quantum hardware.

6 Connections to Course Topics

This project integrates key concepts from the following domains:

- Knowledge Representation: KG data modeling, ontology design, and RDF triple stores.
- Natural Language Processing: State-of-the-art entity and relation extraction from unstructured Wikipedia and arXiv text.