QueryBridge: The Largest and First Annotated Dataset for Question Answering over Knowledge Graphs

Abdelghny Orogat, Ahmed El-Roby
School of Computer Science
Carleton University

1 INTRODUCTION

In the era of Large Language Models (LLMs), there is an urgent need for Question Answering over Knowledge Graphs (QAKG) datasets that utilize LLM capabilities to interpret natural language questions and translate them into structured queries, such as SPARQL. Existing datasets like QALD and LC-QuAD are often limited in size and lack the necessary annotations to train LLMs effectively.

2 QUERYBRIDGE OVERVIEW

We introduce QueryBridge[1] as the largest annotated dataset for QA over knowledge graphs, comprising one million questions. This dataset serves as a vital resource for developing algorithms that interpret natural language in a structured context. QueryBridge supports QA tasks and is also applicable for entity extraction, relationship extraction, query generation, and multi-hop reasoning.

3 DATASET STRUCTURE

In developing the QueryBridge dataset, we update the Maestro system [1, 2] to add the annotation process. The resulting dataset, QueryBridge, is structured into a series of files, each containing JSON objects that represent annotated questions. An example of a JSON object from the QueryBridge dataset is illustrated in Figure 1. Fields include the seed entity, its type, the natural language question, the tagged version, the corresponding SPARQL query, graphical representations, and metrics like the number of triples and tokens.

4 QUERY SHAPES AND TAGS

The dataset categorizes questions into various shapes, each representing a distinct structural pattern. These include the *Single-Edge* shape for basic questions using a single triple pattern, and the *Chain* shape, consisting of a sequence of connected triple patterns. The *Star* shape identifies a seed node through multiple constraints, while the *Tree* shape adds complexity with interconnected stars. The *Cycle* shape links a seed entity through distinct paths, and the *Flower* shape depicts subgraphs with multiple attachments, often including other shapes. Lastly, the *Set* shape encompasses disconnected components for comparative questions. For more details on these shapes and their corresponding queries, refer to Table 1.

Each question is annotated with predefined tags. These include <qt> for question types (e.g., Who, What), <p> for predicates representing relationships, <o> for object entities, <s> for answer entities in Yes-No questions, and <cc> for coordinating conjunctions connecting different branches. In the presentation, we will demonstrate how these annotations help in understanding the question structure using a fine-tuned LLM.

[1]https://huggingface.co/datasets/aorogat/QueryBridge


```
{
  "seed_withPrefix": "http://dbpedia.org/resource/Robert_Gates",
  "seedType_withPrefix": "http://dbpedia.org/ontology/OfficeHolder",
  "questionString": "Who is the chancellor of College of William & Mary?",
  "questionStringTagged": "<qt>Who</qt> <p>is the chancellor of</p> <o>College
  ↪  of William & Mary</o>?",
  "query": "SELECT DISTINCT ?Seed
  ↪  WHERE{\n\t<http://dbpedia.org/resource/College_of_William_&_Mary> \t
  ↪  <http://dbpedia.org/ontology/chancellor> \t ?Seed  .\n}",
  ...
  "shapeType": "SINGLE_EDGE",
}
```


Figure 1: Example of a QueryBridge Question

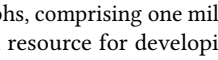
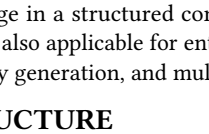
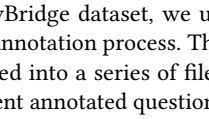
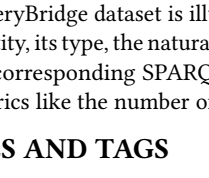
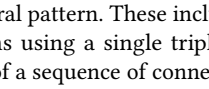
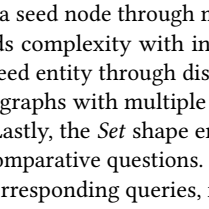
Query Shape		Query Template
Single Edge		`SELECT ?Seed WHERE { ?Seed P O. }`
Chain		`SELECT ?Seed WHERE { ?Seed P1 ?O1. ?O1 P2 ?O2. ... ?On-1 Pn On }`
Star		`SELECT ?Seed WHERE { ?Seed P1 O1. ?Seed P2 O2. ... ?Seed Pn On. }`
Tree		`SELECT ?Seed1 WHERE { ?Seed1 P1 ?Seed2. ?O1 P2 O2. ?O2 P3 O3. ... ?Seed2 P1.1 O1.1. ... }`
Cycle		`SELECT ?Seed WHERE { ?Seed P1 O. ?Seed P2 O. }`
Set		`SELECT ?Seed WHERE { ?Seed P1.1 n1. ?Seed P1.2 O1.2. S P2.1 n2. S P2.2 O2.2. FILTER(n1 > n2) }`

Table 1: Query Shape and Query Template Overview.

REFERENCES

[1] A. Orogat and A. El-Roby. SmartBench: Demonstrating Automatic Generation of Comprehensive Benchmarks for Question Answering over Knowledge Graphs. *Proceedings of the VLDB Endowment (PVLDB)*, 15(12), 2022.

[2] A. Orogat and A. El-Roby. Maestro: Automatic Generation of Comprehensive Benchmarks for Question Answering over Knowledge Graphs. *Proceedings of the ACM on Management of Data*, 1(2), 2023.