Knowledge Graph Entity Type Prediction Using Large Language Models

M1 PLDAC janvier-mai 2025

Nombre d'étudiants

Projet pour un binôme

Encadrants

- Yuhe Bai, Yuhe.Bai@lip6.fr
- Hubert Naacke, Hubert.Naacke@lip6.fr

Project description

Knowledge Graphs (KGs) represent entities and their relationships in a structured form and are widely used in search engines, recommendation systems, and question-answering systems. However, challenges such as incomplete entity types and data sparsity hinder their full potential. The Entity Typing (ET) task involves predicting missing entity types in a knowledge graph, such as inferring that *Barack Obama* has the type *President*, using relational triples like *Barack Obama is leader of U.S.* as input, as illustrated in Fig. 1. Recent advancements in Knowledge Graph Embedding (KGE) and reasoning techniques have improved KG representation and completion by mapping entities and relations to vector spaces. While traditional entity type prediction models based on Graph Neural Networks (GNNs), embeddings, or logic rules have limitations in scalability and expressiveness, the rise of Large Language Models (LLMs) presents new opportunities. By integrating LLMs with KGs through fine-tuning, this project explores cutting-edge methods to enhance entity type prediction.

This project focuses on the SSET framework [2], with a particular emphasis on Step 1, which leverages fine-tuned BERT models for KG entity type prediction. The primary objective is to improve the model's performance and efficiency by optimizing semantic input representation and exploring advanced fine-tuning techniques like Low-Rank Adaptation (LoRA). You will experiment with crafting input sentences that better encode both semantic and structural information from KGs, enhancing the model's reasoning capabilities. You will investigate

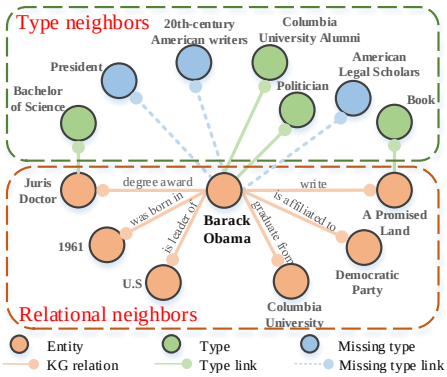

Figure 1: A KG with its entity type information (source [1]).

the following directions which aim to generate richer textual sequences suitable for LM fine-tuning:

- Rather than treating each KG triple independently (as in [2], step 1), consider which triples from an entity's neighborhood could be grouped together. Should all direct neighbors be included, or only a subset of them? Are indirect (multi-hop) neighbors also relevant? Propose a method to identify and select the most relevant triples for a given KG entity.
- After selecting a set of triples to use as input, how can you generate a meaningful textual sentence from them? How should the textual descriptions of the entities and relations be handled to ensure clarity and coherence? Some relations in certain datasets may not be easily comprehensible, and representing them in a more semantic and human-readable manner could improve the overall quality of the input.

Additionally, the project involves testing improvements on benchmark datasets such as FB15kET and YAGO43kET, evaluating the impact of these changes on prediction accuracy and efficiency.

References

[1] Zhiwei Hu, Víctor Gutiérrez-Basulto, Zhiliang Xiang, Ru Li, and Jeff Z. Pan. Transformer-based entity typing in knowledge graphs. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *ENMLP*, pages 5988–6001, 2022.

[2] Muzhi Li, Minda Hu, Irwin King, and Ho-fung Leung. The integration of semantic and structural knowledge in knowledge graph entity typing. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6625–6638, 2024. Full text available at `https://aclanthology.org/2024.naacl-long.369/`.