University *of Ljubljana*
Faculty *of Computer and Information Science*

Integrating Structured Knowledge into Large Language Models

Konstantin Bojchevski, Matic Conradi, and Matjaž Kumin

Abstract

This project explores integrating structured knowledge into Large Language Models (LLMs) to improve reasoning. We focus on incorporating knowledge graphs (KGs), particularly Slovenian linguistic data, to enhance question answering. We explore direct input augmentation, graph embedding-based integration, and attention-based fusion techniques. This report presents initial ideas and methods for integration, evaluating their impact on model performance. The project leverages the Slovenian Digital Dictionary Database to provide structured linguistic information to the LLM.

Keywords

LLMs, Knowledge Graphs, Slovenian Linguistics, Integration, Reasoning

Advisor: Slavko Žitnik

Introduction

Large Language Models (LLMs) have significantly advanced natural language understanding and generation. However, these models often struggle with complex reasoning tasks that require structured knowledge. This project explores techniques for integrating knowledge graphs (KGs) into LLMs to enhance their ability to accurately answer complex questions.

Our focus is on incorporating structured Slovenian linguistic data into LLMs via KGs. By leveraging semantic relationships within the Slovenian Digital Dictionary Database and other linguistic sources, we aim to improve model performance in accuracy and reasoning. This project will evaluate different integration techniques and their impact.

Related Work

Several studies have explored combining knowledge graphs and LLMs to improve reasoning and language understanding. Key works include:

- **GraphGPT: Graph Instruction Tuning for Large Language Models** [1]: Tang et al. (2024) explored tuning LLMs with graph-structured data, demonstrating the benefits of structured knowledge in language generation.
- **Combining Knowledge Graphs and Large Language Models** [2]: Kau et al. (2024) investigated hybrid approaches for combining KGs and LLMs.
- **Deep Bidirectional Language-Knowledge Graph Pretraining** [3]: Yasunaga et al. (2022) introduced pretraining methods integrating bidirectional knowledge graphs into LLMs.
- **Graph Language Models** [4]: Plenz and Frank (2024) overview the use of graph-structured information with LLMs.
- **Chain-of-Knowledge: Integrating Knowledge Reasoning into Large Language Models by Learning from Knowledge Graphs** [5]: Zhang et al. (2024) explored integrating knowledge reasoning into LLMs by learning from knowledge graphs.
- **InfuserKI: Enhancing Large Language Models with Knowledge Graphs via Infuser-Guided Knowledge Integration** [6]: Wang et al. (2024) proposed enhancing LLMs with KGs via infuser-guided knowledge integration.

Initial Ideas

The primary dataset for this project will be sourced from the Slovenian Digital Dictionary Database. This dataset contains structured linguistic information crucial for our integration techniques. We hypothesize that explicitly providing semantic relationships, word definitions, and contextual usages will enable the LLM to better understand and reason about complex linguistic queries. Additionally, we may incorporate other Slovenian linguistic resources and adapt existing datasets

where necessary. The dataset will be preprocessed and merged with other sources, ensuring its compatibility with knowledge graph structures. Initial ideas include:

- **Semantic Enrichment:** Using the dictionary's hierarchical structure (e.g., word classes, part-of-speech tags) to provide richer semantic information to the LLM. This allows the model to better understand the relationships between words.
- **Contextualization:** Embedding definitions and example usages within the knowledge graph to provide the LLM with broader context for interpreting and reasoning about queries. This includes example sentences from the dictionary.
- **Knowledge Graph Traversal:** Constructing complex queries that require the LLM to traverse the knowledge graph to retrieve specific relationships or definitions to arrive at an accurate answer. This tests the model's ability to utilize the KG for reasoning.
- **Cross-Lingual Transfer:** Investigating whether the structured knowledge from the Slovenian dictionary can be used to improve the performance of LLMs on related tasks in other Slavic languages.

Methods

We will employ the following methods to integrate KGs into LLMs, focusing on the Slovenian linguistic context:

1. **Dataset Creation**: We will construct a dataset of complex questions pertaining to Slovenian linguistics. These questions will be designed to require reasoning about the relationships defined in the Slovenian Digital Dictionary Database. Each question will be paired with a relevant KG subset extracted from the dictionary.
2. **Direct Input Augmentation**: We will incorporate KG data directly into the input prompts. This will involve crafting prompts that include relevant facts, definitions, and relationships extracted from the knowledge graph, providing explicit hints to the LLM. Different prompt engineering strategies will be explored.
3. **Graph Embedding-Based Integration**: We will convert KGs into vector embeddings using techniques like node2vec or graph convolutional networks (GCNs). These embeddings will then be fed into the LLM, potentially by concatenating them with the word embeddings or integrating them into the attention mechanism. We will experiment with different embedding dimensions and architectures.
4. **Attention-Based Fusion Techniques**: We will modify the LLM's attention mechanism to incorporate KG-derived features. This could involve using the KG embeddings to bias the attention weights, allowing the model to focus on the most relevant nodes in the knowledge graph when processing the input query.
5. **Evaluation**: We will assess the effectiveness of each technique using metrics such as accuracy, precision, recall, and F1-score. We will also analyze the types of errors made by the LLM with and without KG integration to understand the strengths and weaknesses of each approach. Statistical significance testing will be used to validate the results.

References

[1] Jiabin Tang, Qinkai Zheng, Yuchen Yan, Hongliang Fei, Jianfei Cai, and Li Guo. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

[2] A. Kau, X. He, A. Nambissan, A. Astudillo, H. Yin, and A. Aryani. Combining knowledge graphs and large language models. *arXiv preprint arXiv:2407.06564*, 2024.

[3] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, 35:37309–37323, 2022.

[4] Moritz Plenz and Anette Frank. Graph language models. *arXiv preprint arXiv:2401.07105*, 2024.

[5] Y. Zhang, X. Wang, J. Liang, S. Xia, L. Chen, and Y. Xiao. Chain-of-knowledge: Integrating knowledge reasoning into large language models by learning from knowledge graphs. *arXiv preprint arXiv:2407.00653*, 2024.

[6] F. Wang, R. Bao, S. Wang, W. Yu, Y. Liu, W. Cheng, and H. Chen. Infuserki: Enhancing large language models with knowledge graphs via infuser-guided knowledge integration. *arXiv preprint arXiv:2402.11441*, 2024.