Review 1: "Exploring Large Language Models for Knowledge Graph Completion"

MohammadJavad Ardestani
University of Alberta
Ardestan@ualberta.ca

1 First Path

1.1 Category

The paper focuses on knowledge graph completion (KGC) with a specific emphasis on addressing the issue of incomplete knowledge graphs. It introduces the Knowledge Graph Large Language Models (KG-LLM) framework, which incorporates LLM to bring innovation to the modeling of triples within knowledge graphs.

1.2 Context

The paper provides context by addressing the limitations of KGC techniques, citing various existing methods. Some approaches, solely rely on the structural information present in triples (Bordes et al., 2013; Yang et al., 2015; Wang et al., 2017). While this enables the utilization of KG information in a low-dimensional continuous space, it remains constrained to the structural aspects of triple facts. In contrast, other efforts, aim to enhance KG Representation by incorporating textual and contextual information (Xiao et al., 2017; Xu et al., 2016). These endeavors seek to uncover semantic relevance and improve overall representation beyond the limitations of structural information.

1.3 Contributions

The paper builds on the authors' prior exploration of utilizing the BERT Language model for embedding KG information. While acknowledging some enhancements to BERT's efficiency and performance, the authors highlight its relative limitations compared to contemporary LLM. Recognizing the superior capabilities of LLMs, such as in-context learning, instruction following, and step-by-step reasoning, the paper introduces a novel approach for KGC. This involves mapping the KGC problem to a sequence-to-sequence framework, treating entities, relations, and triples as textual sequences. Their experimental results demonstrate that even relatively small LLMs, such as LLaMA and ChatGLM, can surpass state-of-the-art (SOTA) models after fine-tuning.

1.4 Correctness

While their experimental results showcase the superiority of their fine-tuned KG-LLMs over SOTA models, Language Models are widely recognized for their effectiveness in sequence-to-sequence tasks (Song et al., 2019). The fundamental concept of employing LLM for this task appears both correct and innovative. However, a notable concern arises from the absence of discussion on how the authors plan to address potential overfitting issues of an LLM to a specific KG and what generalization techniques might prove beneficial. Providing clarification on these aspects would enhance the completeness and robustness of their proposed approach. Additionally, the comparison of their model to ChatGPT and GPT-4, based on only 100 data points, raises questions about the robustness of the comparison.

1.5 Clarity

The paper exhibits clear and effective writing, with a well-organized structure that facilitates the understanding of their models. The information flow for explaining their models is coherent and accessible to readers. The inclusion of sample examples illustrating model input and output enhances comprehension. However, the lack of specifics on the implementation details and the dataset used for fine-tuning LLMs, as well as the KGs employed in the fine-tuning process, leaves a notable gap in the presentation.

2 Second Path

2.1 Summary

The authors introduce KG-LLM, a novel approach employing LLMs to augment the comprehension and completion of KGs. In their proposed methodology, entities, relations, and triples are treated as sequential text, transforming the task of KGC into a sequence-to-sequence problem.

The authors provide limited insights into the fine-tuning process. Despite not explicitly mentioning the KG used for fine-tuning, an examination of their code reveals a segmentation of each KG dataset into training and testing sets. Subsequently, the LLMs underwent fine-tuning on specific training partitions of KG datasets. Their performance was then assessed on the corresponding test sets using three language models: ChatGLM with 6B parameters, and LLaMA with 7B and 13B parameters.

The paper delineates their solution to three fundamental problems within KGC. They formulate each sequence of triples into three types of prompts corresponding to each task. For example, consider the sentence "Biden is president of U.S.A" outlined as follows:

- **Entity Prediction:** Given a head/tail entity and a relationship, the objective is to predict the tail/head entity associated with the given head/tail entity. Prompt: "What/Who/When/Where is president of U.S.A?", expected output: "Biden".
- **Relation Prediction:** The goal is to predict the relationship when given a head entity and a tail entity. Each prompt includes a list of possible relations, requiring the model to select the appropriate one. Prompt: "What is the relationship between Biden and U.S.A? Please choose your answer from: is president of | played for | was born", expected output: "is president of".
- **Triple Classification:** This task entails classifying a given triple as either correct or incorrect. Prompt: "Is this true: Biden is president of U.S.A?", expected output: "Yes, this is true".

For the evaluation of performance, the authors manually labeled the sequence outputs of the LLMs. In general, if the LLMs' output contains the expected words, the predicted output is considered correct; otherwise, it is deemed a wrong prediction.

Concerning the entity prediction problem, YAGO3-10 (Dettmers et al., 2018) and WN18RR (Dettmers et al., 2018) were utilized. Based on the Hits@1 metric, KG-LLaMA-13B outperformed existing SOTA models, although GPT-4 demonstrated superior performance for 100 random datapoints. For the relation prediction problem, YAGO3-10 (Dettmers et al., 2018) was employed, and KG-LLaMA-7B outperformed other models based on the Hits@1 metric. In triple classification, WN11 (Dettmers et al., 2018) and FB13 (Bollacker et al., 2008) were selected. KG-LLaMA-13B either outperformed or exhibited similar performance to other existing models.

2.2 Follow up Questions

If we train a language model on one KG dataset, how might its performance be affected when evaluating it on a different KG? What implications could this have for the model's generalization across diverse knowledge domains?

Can the proposed KG-LLM method effectively manage different types of KGs and scenarios beyond those explored in their experiments? Are there specific situations where KG-LLM might encounter challenges or limitations?

2.3 Terms I Did Not Understand

- Translational Distance Models

2.4 Relevant References

"LambdaKG: A Library for Pre-trained Language Model-Based Knowledge Graph Embeddings" and "LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities" are other recent papers focused on KGC using LLMs.

2.5 Relevance for the Project

This paper can be the foundation of our project, and we can enhance its methodology by incorporating a few-shot Chain of Thought prompting approach, which reduces the risk of overfitting compared to conventional fine-tuning methods. Moreover, we can employ LLMs to automatically extract the final predicted label from the output, eliminating the need for manual intervention in this process.

3 Acknowledgment

ChatGPT 3.5 was utilized as a tool for refining language and checking grammar in this report.

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

Han Xiao, Minlie Huang, Lian Meng, and Xiaoyan Zhu. 2017. Ssp: semantic space projection for knowledge graph embedding with text descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Jiacheng Xu, Kan Chen, Xipeng Qiu, and Xuanjing Huang. 2016. Knowledge graph representation with jointly structural and textual encoding. *arXiv preprint arXiv:1611.08661*.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.