Knowledge Graph Guided Semantic Evaluation of Language Models For User Trust

Kaushik Roy
Artificial Intelligence Institute
University of South Carolina
Columbia, SC, USA
kaushikr@email.sc.edu

Tarun Garg
Birla Institute of
Technology and Science
Pilani
f20160450h
@alumni.bits-pilani.ac.in

Vedant Palit
Indian Institute of Technology
Kharagpur
vedantpalit
@kgpian.iitkgp.ac.in

2023 IEEE Conference on Artificial Intelligence (CAI) | 979-8-3503-3984-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/CAI54212.2023.00108

***Abstract*—A fundamental question in natural language processing is - what kind of language structure and semantics is the language model capturing? Graph formats such as knowledge graphs are easy to evaluate as they explicitly express language semantics and structure. This study evaluates the semantics encoded in the self-attention transformers by leveraging explicit knowledge graph structures. We propose novel metrics to measure the reconstruction error when providing graph path sequences from a knowledge graph and trying to reproduce/reconstruct the same from the outputs of the self-attention transformer models. The opacity of language models has an immense bearing on societal issues of trust and explainable decision outcomes. Our findings suggest that language models are models of stochastic control processes for plausible language pattern generation. However, they do not ascribe object and concept-level meaning and semantics to the learned stochastic patterns such as those described in knowledge graphs. This has significant application-level user trust implications as stochastic patterns without a strong sense of meaning cannot be trusted in high-stakes applications.**

***Index Terms*—Knowledge Graph, Graph Neural Networks, Transformers**

I. Introduction

Recent studies have studied self-attention models such as transformers for their ability to encode underlying graph structures by drawing parallels with graph neural networks (GNNs) [1] Intuitively, there is a correspondence between the self-attention map in the transformer and the normalized adjacency matrix in GNNs. Also, there is a correspondence between GNN node representations and the output value vectors from a transformer. The multiple routings of the transformer output through layers of the transformer are similar to multiple graph convolution aggregations in a GNN. Thus, both transformers may be an effective way to learn graph contexts between language tokens. In this study, we aim to test this perceived equivalence rigorously.

Do transformers encode semantic graphs between input sequence tokens? We perform simple experiments that feed various graph path sequence inputs to transformers (we test with multiple KGs and LMs) and try reconstructing the input graph from transformer outputs. In our experiments, we find that in doing so, a high reconstruction error is observed for certain types of graph paths, *paths that require strongly typed real-world concept level knowledge* (e.g., Volvo is typically a type of high-performance car which is, in turn, a type of car). Several previous works have performed similar knowledge graph-based reconstruction experiments. However, they have measured link prediction performance alone and not path predictions [2]. Link prediction is a weak evaluation of knowledge graph semantics as the richness of concepts in a knowledge graph comes from graph paths connecting concepts comprising multiple relationships. Furthermore, they have not qualitatively analyzed the results of successful and failed outcomes. In this study, we quantitatively measure the ability of transformers to predict relationships and concepts in knowledge graph paths. We also qualitatively inspect the paths on which the model makes errors to evaluate their conceptual understanding capabilities.

II. Methdology

First, we extract masked graph paths from the knowledge graphs for processing by the language model. Figure 1 illustrates the masked graph path extraction process from the knowledge graph. Next, we predict the masked tokens and

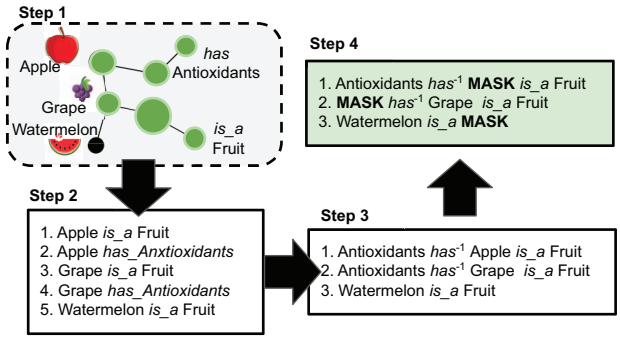

Fig. 1. Steps 1, 2, and 3 show the process of converting the knowledge graph links to paths. Step 4 shows the masked inputs to the language model that will predict the masked tokens. The links are connected two make longer paths through the use of inverse relationships, e.g., has^{-1}.

calculate the percentage of times the language models assign the correct token top five prediction ranks (measured using softmax over logits). Figure 2 illustrates this process. The final softmax logits obtained can be ranked in order of probability

979-8-3503-3984-0/23/$31.00 ©2023 IEEE
DOI 10.1109/CAI54212.2023.00108

values. For our evaluation metric, we calculate the percentage of times the correct answer is within the top five probabilities. We call this metric **%Top@5**.

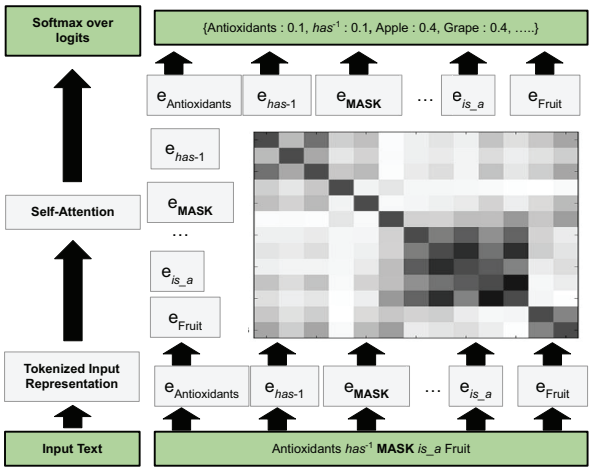

Fig. 2. The figure shows how the masked graph path inputs are processed through the self-attention transformer models to obtain softmax logit outputs.

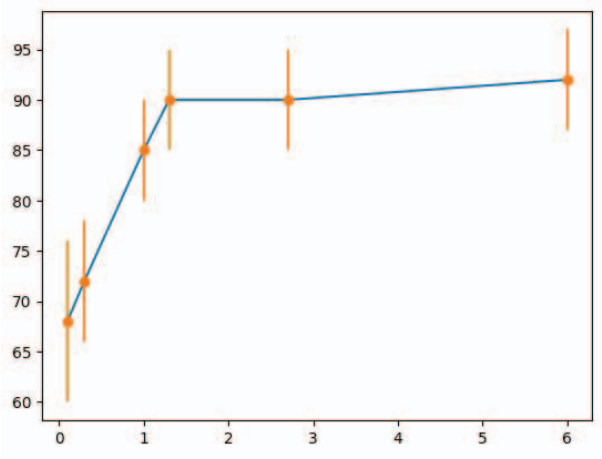

Fig. 3. The X-axis denotes the number of parameters in billions, and the Y-axis measures the **%Top@5**. The performance measured using **%Top@5** increases steadily with the number of model parameters. However, after a certain amount of parameters is reached ($\sim$ 1 billion), the performance starts to flat-line. The variance across different runs remains significant ($\sim \pm$ 5), although it also shows a decreasing trend with increased model parameters.

III. Experiments

We extract approximately 300K knowledge graph links from the knowledge graphs DBPedia, ConceptNet, Wiktionary, WordNet, and OpenCyc Ontology [3]. The relationships we find are *Antonym, DistinctFrom, EtymologicallyRelatedTo, LocatedNear, RelatedTo, SimilarTo, Synonym, AtLocation, CapableOf, Causes, CausesDesire, CreatedBy, DefinedAs, DerivedFrom, Desires, Entails, ExternalURL, FormOf, HasA, HasContext, HasFirstSubevent, HasLastSubevent, HasPrerequisite, HasProperty, InstanceOf, IsA, MadeOf, MannerOf, MotivatedByGoal, ObstructedBy, PartOf, ReceivesAction, SenseOf, SymbolOf, and UsedFor*. The data can be found at this link. For the language models, we use bert-base-uncased, bert-large, GPT-Neo small, medium, and large with 0.1B, 0.3B, 1B, 2.7B, and 6B parameters, respectively (B stands for billion) [4].

A. Quantitative Results

Figure 3 shows the quantitative results. We explain the results in the figure caption due to space limitations.

B. Qualitative Results

We manually inspect the knowledge graph paths at which the language models fail, which we will call *false paths*. Interestingly, the *false paths* almost exclusively involve knowledge of strongly typed objects and their properties as seen in the real world. Some examples include "volvo IsA car CapableOf slow_down", "retrograde_motion HasContext astronomy IsA physics", "handicapped SimilarTo unfit RelatedTo unhealthy", and "ultimate_frisbee IsA field_game IsA outdoor_game". The remaining examples are at this link. This finding is particularly interesting as it supports third-party observations about language models' fundamental lack of a conceptual world model when asked about physics-related questions (e.g., block-stacking) [5].

IV. Conclusion

This paper opens the black-box language models' ability to model knowledge graph semantics by proposing masked prediction tasks on graph paths. We do this to understand a language model's conceptual understanding and its bearings on application-level user trust issues. We introduce metrics for the evaluation of the results and also manually inspect the outcomes.

Our findings suggest that language models are models of stochastic control processes for plausible language pattern generation. However, they do not ascribe object and concept-level meaning and semantics to the learned stochastic patterns such as those described in knowledge graphs. This has significant application-level user trust implications for applications requiring concept-level understanding (e.g., healthcare) and physical simulations (e.g., war-time strategies). Our findings suggest that using language models alone, which are stochastic control models, to drive high-stake application-level decisions would be highly unsafe and irresponsible.

Acknowledgements

This work was supported in part by the National Science Foundation under Grant 2133842, "EAGER: Advancing Neuro-symbolic AI with Deep Knowledge-infused Learning and was carried out under the advisement of Prof. Amit Sheth." [6], [7].

REFERENCES

[1] E. Choi, Z. Xu, Y. Li, M. Dusenberry, G. Flores, E. Xue, and A. Dai, "Learning the graphical structure of electronic health records with graph convolutional transformer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 606–613, 2020.

[2] H. Ma and D. Z. Wang, "A survey on few-shot knowledge graph completion with structural and commonsense knowledge," *arXiv preprint arXiv:2301.01172*, 2023.

[3] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[4] R. Gozalo-Brizuela and E. C. Garrido-Merchan, "Chatgpt is not all you need. a state of the art review of large generative ai models," *arXiv preprint arXiv:2301.04655*, 2023.

[5] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

[6] A. Sheth, K. Roy, and M. Gaur, "Neurosymbolic ai-why, what, and how," *arXiv preprint arXiv:2305.00813*, 2023.

[7] A. Sheth, M. Gaur, K. Roy, and K. Faldu, "Knowledge-intensive language understanding for explainable ai," *IEEE Internet Computing*, vol. 25, no. 5, pp. 19–24, 2021.