Knowledge Graphs - Peeking Behind the Curtain of LLMs

Yuri Kinakin
University of California, Berkeley
yuri.kinakin@berkeley.edu

David Noble
University of California, Berkeley
david.noble@berkeley.edu

Abstract

A method of automated knowledge knowledge graph generation was attempted on biological and geoscientific prompts, using both base large language models and those fine-tuned on a domain specific corpus. Manual review of the graph nodes suggests better performance on the fine-tuned models, however, there currently exists no automated or semi-automated way of evaluating these graphs. There appears to be a negative correlation between node connectedness and domain relevance, suggesting that one way to improve the relevance of knowledge graphs generated by base models is to only retain nodes above a minimum connectivity.

1 Introduction

Over last ten years, significant progress has been made on large language models for a wide variety of natural language processing tasks (Min et al., 2021). While these models can achieve (and even occasionally supersede) human level performance, they are significantly more internally complex than historical approaches. Simple explanations of what each model has learned make interpretation difficult, and as these models are expected to be applied against an increasingly wide variety of tasks, having the ability to easily query their internal logic is increasingly important. Some of this can be achieved through the use of a knowledge graph (KG), which encodes semantic information and provides a way to visualize contextual relationships. Unfortunately, KGs have historically been complicated and onerous to create, requiring substantial manual input.

As novel approach for automated generation has been suggested by (Hao et al., 2022) which, as described, generates KGs with higher internal consistency. We propose to text the sensitivity of this approach on domain specific prompts by comparing graphs generated from base large language models (LLMs) and fine-tuned counterparts. Both a geoscientific and biological sciences specific fine-tuned version of the model was compared on the same prompts to check for increased domain relevant tuples.

2 Background

In their paper from 2022, (Hao et al., 2022) defined a process for the automated creation of prompts that accounts for the joint probabilty of masked language token results, as well as a consistency score for the resultant entity pairs that takes into account a weighted sum against all input prompts. This allows the ability to query a LLM with minimal prompting, generating a large number of (hopefully) relevant tuples. We hypothesize that models that are fine trained on a text corpus specific to the question prompt will return much "better" results.

One of the more obvious issues when presented with a large knowledge graph is to determine whether or not it is a reasonable representation of some knowledge base. What makes a knowledge graph "better" than any other? In other words, by what metric should a graph be judged? In (Hao et al., 2022), the authors utilized Amazon MTurk to evaluate the accuracy of the extracted knowledge through manual evaluation by a human. This is obviously a labour intensive and expensive process, so while one option is to take just a sample of triples from a larger graph to evaluate for correctness, an automated method would be preferred.

During a literature search, several alternative metrics are suggested in (Seo et al., 2022), with the overriding assumption made that the structure of a KG is representative of the ontology. Numerical quantification of this structure allows for the several metrics, which in general weight graphs with higher interconnectivity between classes as "better." The instantiation ratios that are suggested are not appropriate for the way in which the KGs are created in this paper, as this method naturally generates a balanced representation of relationships (irrespective of whether or not they are meaningful). The other metrics in (Seo et al., 2022) require evaluation and instantiation of subclasses, which are not a property of the KGs as generated in this work. As an alternative, (Yang et al., 2022) suggest comparing the balance and incident of negative to positive relations, however, as negative relationships were not generated during KG construction in this work they are unavailable to be used in a metric.

In general, it is asserted that efficient and effective metrics for evaluation of knowledge graphs is very much an unsolved problem, and the current state-of-the art utilizes problem specific solutions. In this paper, we will manually evaluate each of the nodes as either domain relevant or not, and only evaluate graph metrics in a qualitative way.

3 Methods

3.1 Fine Tuned Language Model

In order to compare the outputs from different models, it was necessary to first create a fine-tuned BERT model. Fortunately, the process to generate a geoscience specific corpus is well described in (Lawley et al., 2022), providing example notebooks and work flow in a GitHub repository.[1] These notebooks essentially scrape reports from provincial geological surveys and property assessments, convert the native PDFs to text, then clean the corpus. Additional to this work, the Geological Survey of Canada has released a comprehensive formatted dataset as (Raimondo and Lawley, 2022). Several minor adjustments were required in the code to account for changes in web addresses and APIs.

Due to the lower processing requirements, a DistilBERT model was used at the base for fine tuning. A total of 950 scraped reports were used, from which 51,634 groups of 512 tokens were generated for training data. Of these, 90% were used for training with 10% held back for testing. The masking probability was set to 15% for these data. After 3 epochs of 1452 steps, the model and test losses were similar with minimal change between epochs. The resultant DistilBERT model will be referred to as GeoBERT for the remainder of this paper.

Instead of training a new geoscience specific tokenizer, the one provided by the Geological Survey of Canada from their GEOSCAN repository [2] was utilized.

For the biology specific model and tokenizer there fortunately exist a plethora of excellent pre-trained models; this work utilized BioBERT, a BERT based model fine-tuned on a corpus of biomedical research documents (Lee et al., 2019).

3.2 Knowledge Graph Construction

The main construction of the graphs utilized the method proposed by (Hao et al., 2022). Fortunately, draft code has already been prepared in a GitHub repository [3] which was cloned into a CUDA enabled environment. As part of the process to operationalize this code, a custom Docker environment was created and is posted into the repository for this project.

To generate a knowledge graph using this approach, aside from selecting a masked model and several simple model parameters (notably, the number of tuples to consider and the number of prompts to generate), a list of model prompts is required. This takes the form of:

- An initial prompt: e.g. "[ENT1] is a mineral constituent of [ENT0] rocks."
- A list of seed tuples: e.g. [["granite", "feldspar"], ["basalt", "olivine"], ...]
- A list of additional prompts: e.g. ["The mineral [ENT1] is found in the [ENT0] rock .", ...]

[1] https://github.com/NRCan/geoscience_language_models

[2] https://ftp.maps.canada.ca/pub/nrcan_rncan/publications/STPublications_PublicationsST/329/329265/gid_329265.zip

[3] https://github.com/tanyuqian/knowledge-harvest-from-lms

The model returns a number of tuples for each category, each with an associated probability.

Proper tuning of the prompts and seeds was required; it was necessary to go through several iterations of prompt engineering to deliver a sensible output. The initial runs provided models outputs that were, upon a cursory visual inspection, not aligned with the intent of the prompt. A small number of highly specific prompts was ultimately found to be most effective, with the model runs under investigation variably consisting of 3-6 prompts and 500 - 1000 tuples.

Operationally, it was important to balance the benefits from evaluating a large number of tuples to find the most relevant pairs against computational requirements. As individual tokens need to be considered in pairs, the algorithm scales as $O(N^2)$ with an increasing number of tuples. Utilizing between 500 to 1000 tuples led to run times between 15 and 30 minutes on an NVIDIA RTX 3090.

3.3 Visualization and Evaluation

To view and manipulate the graphs, a notebook was created to load the data into a Neo4j database that takes output tuples as a list, creates a set of unique nodes, and assigns relationships according to the label assigned. The Data Science graph library in Neo4j was used to calculate node centrality, triangle counts, and strongly connected components.

4 Results and Discussion

When the same set of prompts were run through a base and fine-tuned LLMs, in both the geoscience specific and biological specific cases, quite different results were obtained. This is hypothesized to be a result of the prompts used to generate the knowledge graphs being designed to be specific enough to query the domain understanding of the LLM. An initial and cursory review of the graph nodes shows that many more of the words and relationships returned by the fine-tuned model are domain related. A human reviewed ranking of the words returned by the KGs returned by the base and fine-tuned models reported a difference of 57.5% domain specific words in the base model versus 92.7% domain specific words in the fine-tuned geoscience specific KG, and 53.2% in the base model and 98.1% in the fine-tuned version for the biology-specific KG.

4.1 Graphs

Initial evaluation of the graphs was visual, with Figure 1 and Figure 2 representing the raw outputs from the base and fine-tuned models respectively. The shapes of both of the resultant graphs are very similar and suggest little difference in the underlying structure of the graphs.

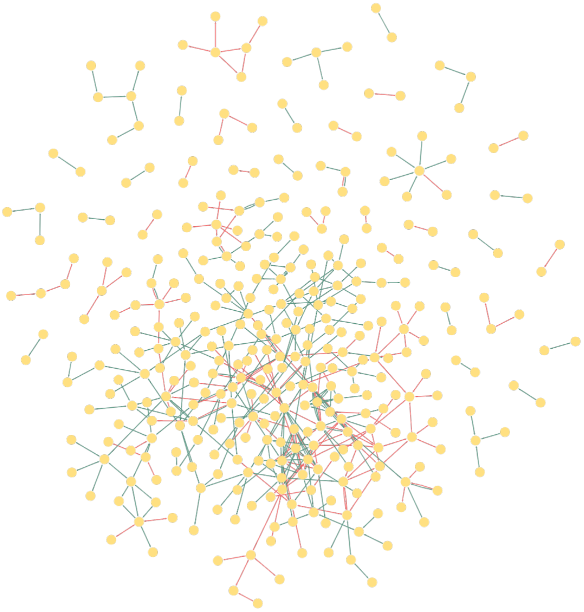

Figure 1: Knowledge graph obtained from querying a base DistilBERT LLM.

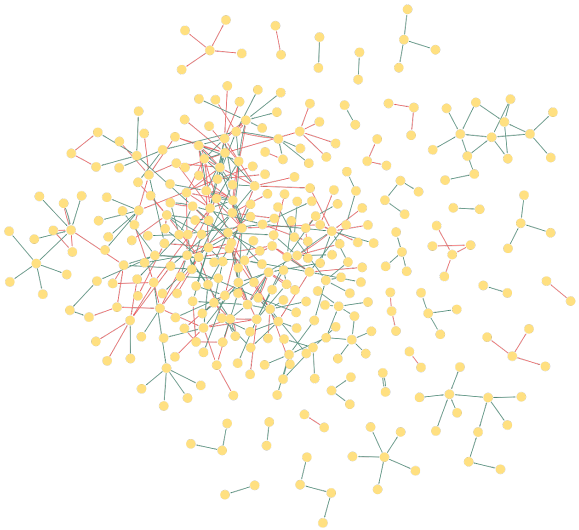

Figure 2: Knowledge graph obtained from querying the GeoBERT model.

As the tuples returned from the KG generation contain directionality, we can switch to a hierarchical representation of the same graphs (Figures 3 and 4). One represented thusly, we can see what appears to be a structural differences, in which the base model has a more circuitous representation with more internal cycles to the graph whereas the fine-tuned model returns a flatter representation.

While, as described in Section 2, the structural methods described in previous work are not likely directly applicable in the evaluation of these automated KGs, an initial review of several structural graph metrics was nevertheless undertaken as part of an exploratory data analysis. Centrality and triangle counts for the two geoscience graph types (base and fine-tuned) were calculated and ranked in descending order. The intent was to identify any differences in connectiveness between the two graph types (Figures 5 and 6). As the shape of these graphs is virtually identical, it adds weight to the supposition that simple automated techniques for extracting structure from graphs are unlikely to be useful in determining their domain relevance or correctness.

We next assessed the sets of KGs we generated with a collection of graph-based characteristics that we hypothesized would allow us to draw insight into the quality of the KGs. These characteristics included the assortativity coefficient, the number of bridges, average node connectivity, and global efficiency. Assortativity is calculated as the Pearson correlation coefficient of degree between linked nodes, which we hypothesized would show how well the graph could generate associations between paired terms. The number of bridges was included to indicate how many connections the graph would include between communities of well-connected nodes. Average node connectivity would simply indicate how densely the generated graph would draw relationships for each node. Finally, global efficiency quantifies the small-world behavior of the graph and indicates how easily one could travel between concepts in the graph ontology. Analyzing our generated KGs, we found that fine-tuning the models had no conclusive effect on any of these systemic graph characteristics. These are summarized graphically in Figure 7. The largest effect we found was that fine-tuned models generated graphs with node compositions that were much more domain-focused and relevant.

While no obvious correlation seemed to exist between structure metrics and domain relevance for the entire graph, this is not necessarily the case at all scales. We next evaluated the domain relevance for those highly connected nodes and compared that to nodes with very low connectivity values. The top 10 words with highest connectivities are entirely composed of geoscience relevant words (with some words appearing in both):

DistilBERT		GeoBERT	
Word	Score	Word	Score
zinc	38	iron	35
phosphate	33	gold	35
oxide	32	lime	33
nickel	32	quartz	33
chloride	31	granite	32
potassium	31	silver	31
sulphur	30	cobalt	31
silver	30	marble	30
gold	29	sodium	30
lead	29	coal	30

Table 1: Words with highest centrality scores from DistilBERT and GeoBERT KGs.

However, for those nodes with very low connectivity, the GeoBERT model is still performing quite well with a number of words being domain specific or related, while the base DistilBERT model contains at least 50% of words that are completely unrelated to the prompts. This is useful as, if the gold-standard evaluative metric continues to be human evaluation of automatically generated KGs, attention should be focused on the nodes with very low connectivitiy (and relationships arising from those nodes) as they are least likely to be relevant or coherent.

It was then hypothesized that utilizing a community detection algorithm might reduce the impact of some of the lower connected nodes through aggregation. A simple community detection algorithm was run on both geoscience specific graphs, with outputs from the top three strongly-connected networks shown as in Table 3. It is interesting to note that the number of domain specific words in the KG from the GeoBERT model all return domain relevant terms, while one of the base model communities is much more related to biologi-

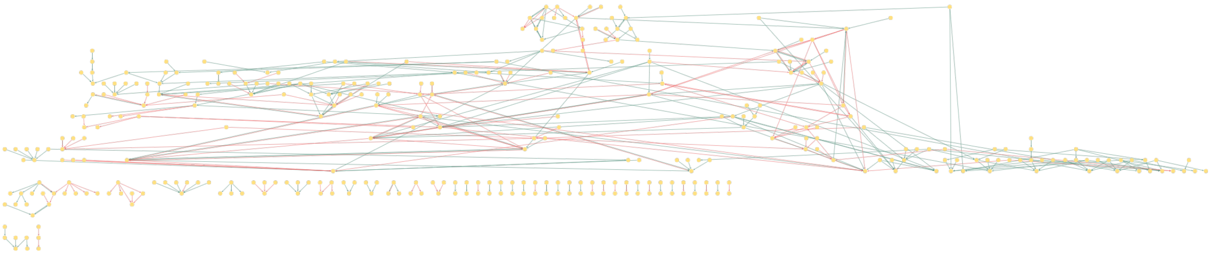

Figure 3: Knowledge graph from base DistilBERT model shown in hierarchical representation.

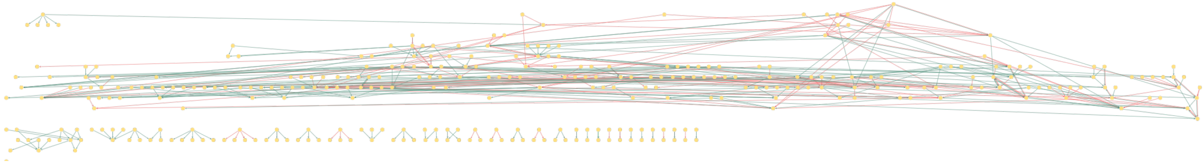

Figure 4: Knowledge graph from base GeoBERT model shown in hierarchical representation.

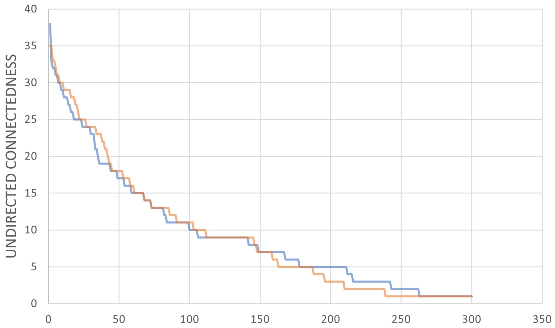

Figure 5: Ranked undirected connectedness values from the DistilBERT (blue) and GeoBERT (orange) for nodes up to 350.

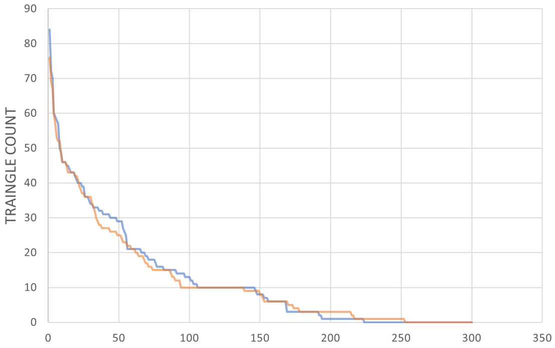

Figure 6: Ranked triangle counts for nodes from the DistilBERT (blue) and GeoBERT (orange) for nodes up to 350.

cal than geological terms.

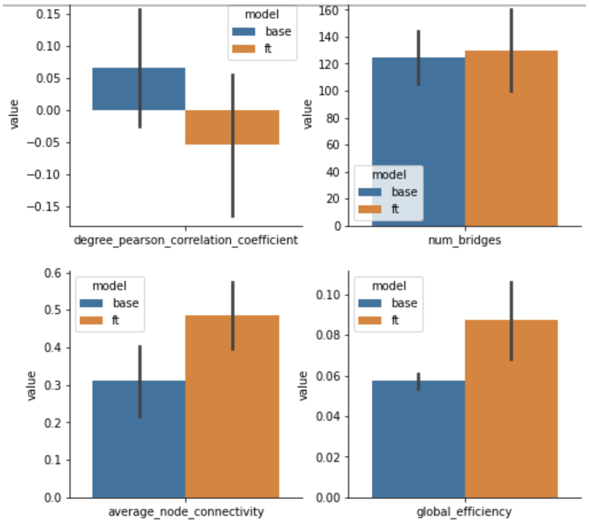

Figure 7: Summary comparison between base and fine-tuned graphs for whole graph characteristics.

5 Conclusion

The conclusion of this work is that, while the automated approach proposed by (Hao et al., 2022) is a method for generating generally coherent knowledge graphs, it is considerably improved by utilizing restricted prompts against a fine-tuned model trained on a domain specific corpus. This is a useful result, but verification required human review of the resultant graphs. This was only possible as these were small graphs with just several hundred nodes and relationships within each - on larger knowledge graphs a full manual review would be prohibitively expensive or otherwise unfeasible exercise. This work suggests a potential negative correlation between domain relevance and node connectedness - evaluation of knowledge graphs should therefore focus on

DistilBERT		GeoBERT	
Word	Score	Word	Score
soda lime	1	mineral exploration	1
mineral grains	1	pine point	1
mineral bone	1	economic potential	1
gum minerals	1	runoff	1
marine gastropod	1	volcanic glass	1
grape wine	1	industrial applications	1
sheep	1	native pb	1
lime juice	1	although lithium	1
rubber dioxide	1	economic significance	1
domain	1	science	1

Table 2: Words with highest centrality scores from DistilBERT and GeoBERT KGs.

DistilBERT

Score	Nodes
170	dioxide, orange, gram, rainbow, elemental, lead creek, vein tin, salt crystal, para
6	formulation, bath waters, salt formation, solutions, list
5	adrenaline, sugar, glucose, yeast, insulin

GeoBERT

Score	Nodes
154	aggregate, kim, natural gas, permafrost, gasoline, tin, ruby, crystalline, little
9	information, legend, noise, equation, title, metadata, data, compilation
6	deformation, zoning, magma, replacement, assimilation

Table 3: Three highest scoring strongly clustered components from the DistilBERT and GeoBERT KGs.

these sparsely connected nodes. Alternatively, an option to improve the resultant knowledge graphs obtained from this process would be an annealing step, retaining only the strongly connected nodes.

There are several next steps that are proposed as next steps for this research:

1. Optimizing code parallelization, allowing for the use of increasingly complex prompts with greater than two masked tokens.

2. Attempting runs with more than a maximum of 1,000 tuples. This should enable a richer potential vocabulary in the resultant tuples.

3. Attempt to reduce resultant graph complexity through clustering routines and algorithms.

References

Hao, S., Tan, B., Tang, K., Ni, B., Zhang, H., Xing, E. P., and Hu, Z. (2022). Bertnet: Harvesting knowledge graphs from pretrained language models.

Lawley, C. J., Raimondo, S., Chen, T., Brin, L., Zakharov, A., Kur, D., Hui, J., Newton, G., Burgoyne, S. L., and Marquis, G. (2022). Geoscience language models and their intrinsic evaluation. *Applied Computing and Geosciences*, 14:100084.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heinz, I., and Roth, D. (2021). Recent advances in natural language processing via large pretrained language models: A survey.

Raimondo, S., C. T. Z. A. B. L. K. D. H. J. B. S. N. G. and Lawley, C. (2022). Datasets to support geoscience language models.

Seo, S., Cheon, H., Kim, H., and Hyun, D. (2022). Structural quality metrics to evaluate knowledge graphs.

Yang, H., Lin, Z., and Zhang, M. (2022). Rethinking knowledge graph evaluation under the open-world assumption.