ISSN No:-2456-2165

Evaluating Next Generation NLG Models for Graph to Text Generation with Fine-Tuning Vs Generalized Model for Longer Generated Texts

Prashant Kaushik
Department of Computer Science & Information Technology
Jaypee Institute of Information Technology, Noida

Abstract:- The paper investigates the feasibility of generative models for graph-to-text generation tasks, particularly in a zero-shot setting where no fine-tuning or additional training resources are utilized. The study evaluates the performance of GPT-3 and ChatGPT on graph-to-text datasets, comparing their results with those of fine-tuned language model (LLM) models like T5 and BART. The findings reveal that generative models, specifically GPT-3 and ChatGPT, exhibit the ability to produce fluent and coherent text, with notable BLEU scores of 11.07 and 11.18 on the AGENDA, & WebNLG datasets, respectively for longer texts. Despite this success, error analysis highlights challenges for actual product usage. In particular Generative models struggle with understanding semantic based relations among entities contexts, leading to the generation of text with hallucinations or irrelevant information. As part of the error analysis, the study employs BERT to detect machine-generated text, which are achieving high macro-F1 scores. The generated text by the generative models is made publicly available by various authors, contributing to the research community's understanding of the capabilities and limitations of such model in the context of graph-to-text generation tasks.

Keywords:- *LLMs, Large Language Models, Generative Models, Graph to Text, Text Generation, Bleu, Rogue.*

I. INTRODUCTION

Graph-to-text generation is a challenging natural language processing task that involves converting structured graph representations into coherent and human-readable textual descriptions. In this process, nodes in the graph represent entities, and edges denote relationships between these entities. The goal is to generate linguistically accurate and contextually relevant text that encapsulates the essential information encoded in the graph. This task is crucial in various domains, including data summarization, knowledge graph completion, and generating textual narratives from structured data sources. Recent advancements in leveraging large language models and attention mechanisms have shown promising results in improving the accuracy and fluency of generated text. Despite these strides, challenges persist, such as ensuring semantic understanding, handling ambiguity, and addressing issues like hallucinations where models generate information not explicitly present in the input graph. Ongoing research focuses on refining techniques to enhance the capabilities of graph-to-text generation models and make them more adept at capturing nuanced relationships and producing high-quality textual descriptions.

GPT Models are capable of generating longer text, but there are some limitations and strategies to consider like Token Limit some GPT model has a soft limit of about 4000 tokens (approximately 500 words). If the input text exceeds this limit, the response may get cut off1. Detailed prompts helps generate longer responses, you can start with detailed prompts that specify length and give GPT models an additional information to write on.

Dividing Text helps very long pieces of text, one approach is to divide the text into smaller fragments, retrieve the appropriate pieces according to various tasks, and then send them through an API calls. Continuation text response gets cut mid-sentence, prompts like ‘continue’, ‘expand’, or ‘go on’ to encourage it to generate more relevant texts.

Rewriting Longer variants of prompts can be asked for GPT models to rewrite its response using more words to get a longer response. These longer sentences are the focus metrics for product ready models.

Remember, the quality of the output depends on the quality of the input. Coherent and logical prompts tend to yield better results.

The evaluation of the models' proficiency in translating graph data into coherent text is conducted on the test sets of two main graph-to-text generation datasets: WebNLG (Gardent et al., 2017) and AGENDA (Koncel-Kedziorski et al., 2019). Employing the linearized sequence representation method introduced by Ribeiro et al. (2021a), where the graph is transformed into a text sequence (as depicted in Figure 1), we assess the generative models' performance comprehensively.To gauge their effectiveness, we conduct a thorough evaluation on each dataset, then employing machine translation metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and ROUGE (Lin, 2004) to evaluate the quality of the generated texts. The results indicate that the generation based models do not reach the level of quality achieved by state-of-the-art approaches in graph-to-text generation. To gain insights into

the patterns of mistakes made by the generative models, error analysis is performed by comparing the generated texts with the reference texts.

In addition to this, we fine-tune a BERT model for the specific task of detecting machine-generated text. This analysis, along with the generated texts, is made publicly accessible on GitHub. This availability aims to facilitate future research, enabling a deeper understanding and analysis of machine-generated text, thereby contributing to advancements in trustworthy AI. The research community can leverage these resources for further exploration and improvements in the field of graph-to-text generation.

➢ *Our Contributions Listed in the below Mentioned Points:*

- Performance evaluation of GPT models using zero-shot setting.
- Ability to generate longer responses for product readiness.
- Coherence measure of fine-tuned vs generalized models.

II. RELATED WORK

The text generation from deep learning model has been a hot topic of research from last 2-3 years. Earlier it was basic Natural language generation now to long text generation using GPT models. These models have been in industrial usage now a days for various usage like contract generation, letter writing etc. Multilingual [1][2] text generation is tending now to cover more customer base by various models. But all these advantages are not without problems or short coming in generated text. Problems like hallucination, deviated context [2] etc is the main bottlenecks of the models.

Context bucketing of the generated response has been attempted by model for NLG [5][8]. These models use GAN with cascaded CNN to train together. This training is tuned such a way that the generated response gives one context for one set of inputs. Similar approach has been applied [9] for traffic prediction for telecom networks.

Object based tokens were used for [6][7] fast video classifications and summarizations. These tokens are also used in large text generations by GPTs [3][4]. These variations in number of tokens and its processing is improved and matured by larger GPT models [4][10]. These tokens have been arranged as graph to text based mapping as well as data to text based mapping [12][4]. Alongside the development of the models there has been work on evaluation metrics along with human evaluation. Bleurt, BLEU [11] etc. has been developed to measure the relevant summarization, which allow a large-scale automatic evaluation of huge mapping response from GPT models [13].

Along with high scale of evaluation there has been work on representation of intermediate mapping between input and outputs [14], which allow internal exploration of mapping and training processes of the NLG models.

Industrial use-case related evaluation has also been performed by various product companies [15] for various platform-based cases like recruitment test generation and its weightage on various sections of question set. Chat bot related training [15] [16] [17] is also being extensively tested by product-based research groups. These outcomes are getting in various directions like AI bots are useful to how bad they are.

Some survey also highlighted the impact and technologies used for these NLG bots [18] [19] [20] along with direction of development for human centric generative models. These survey and direction highlight human based metrics as well to evaluate the large text generation models for next level of training and development. Over all we can say that as of the current stage of GPT models a lot of work is to be done for making it usable and to serve humanity.

III. DATASETS

The evaluation of generative models is conducted using two extensively employed datasets, AGENDA and WebNLG, which have been prominent in recent research on graph-to-text datasets. These datasets are chosen for their prevalence and represent distinct domains: AGENDA focuses on the scholarly domain, pairing knowledge graphs with scientific paper abstracts, while WebNLG encompasses a more general domain, mapping RDF triples from the knowledge graph DBpedia to text.

For our experiments, we specifically concentrate on the test sets of AGENDA and WebNLG, eliminating the need for additional model training. AGENDA comprises instances featuring titles, entities, graphs, and abstracts of scientific papers. The graphs are automatically extracted from the SciIE information extraction system (Luan et al., 2018), with the title, entities, and graph utilized as input for the models.

WebNLG, on the other hand, serves as a benchmark for converting sets of RDF triples to text. The RDF triples are subgraphs of the DBpedia knowledge graph, and the corresponding texts describe these graphs succinctly in one or a few sentences. The WebNLG challenge has released multiple versions of this dataset since 2017, contributing to its ongoing significance in graph-to-text generation evaluations.

IV. EXPERIMENTS

As all GPT's requires a sequence of text as inputs, we convert the graph structure in a linear representation. This converts the text into head, relation, and tail entities. In the other datasets named AGENDA these attributes have been added to make a graph as linear structure.

Base model for comparison and contrast are T5 and BART using the above linear sequence as inputs and outs as generated texts. These models have been selected as base model for comparison as they are fine-tuned as well as generalized to some extent. The models to be evaluated are

ChatGPT (gpt-3.5) and GPT-3. The metrics used are employing machine translation metrics based on para such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004) to evaluate the quality of the generated texts. Minimum token used are 4000 for all test sets as input and the output the generated tokens are separated into various categories as shown in the results tables.

Table 1 Input Token is Fixed at 4000 Tokens (~500 words) with AGENDA Dataset

Model	BLEU	METEOR	RougeL	Number of output Tokens
T5	23	24		6000+
BART	23.5	25		6000+
GPT-3	10	14	26	7000+
ChatGPT3.5	11.2	15	26	7000+
GPT-3	10	16	28	11000+
ChatGPT3.5	11.2	15	28	11000+

Table 2 Input Token is Fixed at 4000 Tokens (~500 words) with WebNLG Dataset

Model	BLEU	METEOR	RougeL	Number of output Tokens
T5	53	44		6000+
BART	53.5	45		6000+
GPT-3	20	27	40	7000+
ChatGPT3.5	12.2	29	42	7000+
GPT-3	20	29	40	11000+
ChatGPT3.5	12.2	30	42	11000+

V. RESULTS

Four GPT models has been taken for experiments for the scope of this article. The number of input tokens are scoped to 4000 tokens. The output responses are considered over 6k tokens only. The table 1 and 2 shows the models and their responses metrics along with the score in the BLEU and METEOR based evaluations, which are very popular for measuring the context of the generated text. The large text generation models looks to be in bad shape for context and relevancy for input text based on BLEU based scores. METEOR scores are also low and can't promise the production use of these models as of now. But as for shorter text it looks promising for non-production use-cases.

➢ *Error Analysis of Experimentation*

Error analysis is a crucial process in experimentation and understanding, involving the systematic examination of errors that occur within a system or application. This multifaceted approach encompasses the identification, categorization, and thorough investigation of errors to discern their root causes. Through meticulous root cause analysis, developers and relevant stakeholders aim to understand the underlying issues leading to errors, whether they stem from coding errors, system architecture, or other unforeseen factors. The severity and impact of errors are assessed and based on that many responses were discarded

Based on the above methods for error analysis, we have removed some responses which are fully out of context and proved outlier. All the scores are calculated above are after the removal of error cases. As the outliers can significantly impact the score metrics. Thereby making the evaluation biased.

VI. CONCLUSION

The paper explored the generative models for evaluation and diversity purposes. We used two bench marked datasets AGENDA and WebNLG for generating the evaluation text from the LLMs. The adopted linearized graph representation approach, following the related prior research, was leveraged in our study. Utilizing the zero-shot capability of language models, it is incorporated prompts at the beginning of the input text for both GPT-3 and ChatGPT. A comprehensive evaluation, employing various metrics like Bleu and ROGUE was used. However, our findings indicate that generative models, despite their zero-shot capabilities, do not outperform previous models that have undergone training and fine-tuning on large datasets. This underscores the limitations of generative models in achieving state-of-the-art performance in graph-to-text generation tasks. Additionally, an error analysis of the text generated by the models revealed challenges in capturing relationships between entities, often resulting in the generation of unrelated information and hallucinations. To further scrutinize the machine-generated text, we employed fine-tuned BERT for a text classification task. BERT exhibited high F1 scores in distinguishing between machine-generated and human-written text. In conclusion, our study provides a thorough evaluation of generative models for graph-to-text generation. To advance this field, future work should concentrate on refining machine-generated text and mitigating hallucinations. This may involve further exploration of generative models and novel training techniques to improve their effectiveness in graph-to-text generation tasks.

FUTURE WORK

The future work in the evaluation of large language models could explore several avenues to enhance our understanding and optimize their performance. Some potential directions for future research include bias detection and its mitigation, domain specific evaluation, multi modal capabilities, user-centric evaluations and robustness to adversarial attacks. Future work in the evaluation of large language models should strive to address these challenges and contribute to the ongoing improvement of these powerful language generation systems across diverse applications and contexts. We in particular will work on almost all the aspects of discussed above as well the automation aspects of evaluation.

REFERENCES

[1]. Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. arXiv preprint arXiv:2303.12528

[2]. Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wen liang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multi task, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023.

[3]. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

[4]. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language under standing. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech nologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[5]. P. Kaushik, K. V. Kumar and P. Biswas, "Context Bucketed Text Responses using Generative Adversarial Neural Network in Android Application with Tens or Flow-Lite Framework," 2022 8th International Conference on Signal Processing and Communication (ICSC), Noida, India, 2022, pp. 324-328, doi: 10.1109/ICSC56524.2022.10009634.

[6]. P. Kaushik and V. Saxena, "Fast Video Classification based on unidirectional temporal differences based dynamic spatial selection with custom loss function and new class suggestion," 2023 International Conference on Disruptive Technologies (ICDT), Greater Noida, India, 2023, pp. 419-423, doi: 10.1109/ICDT57929.2023.10150644.

[7]. P. Kaushik and V. Saxena, "Video annotation & description using machine learning & deep learning: critical survey of methods" Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing August 2023Pages 722–735

[8]. Sunil prasad P Kaushik, "Context Aware GAN for sequence text generation in tensor-flow lite for android AI chat application" , International Journal of Scientific & Engineering Research Volume 12, Issue 9, 2021.

[9]. P. Kaushik, S. Singh and P. Yadav, "Traffic Prediction in Telecom Systems Using Deep Learning," 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2018, pp. 302-307, doi: 10.1109/ICRITO.2018.8748386.

[10]. Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021a. Investigating pretrained language models for graph-to-text genera tion. In Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI, pages 211–227, Online. Association for Computational Lin guistics.

[11]. Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text gener ation. In ACL.

[12]. Chao Zhao, Marilyn Walker, and Snigdha Chaturvedi. 2020. Bridging the structural gap between encoding and decoding for data-to-text generation. In Proceed ings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2481–2491, On line. Association for Computational Linguistics.

[13]. Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In Proceedings of the 2018 Con ference on Empirical Methods in Natural Language Processing, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

[14]. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Represen tations, ICLR'19.

[15]. Panagiotis Tsoutsanis and Aristotelis Tsoutsanis, "Evaluation of Large language model performance on the Multi-Specialty Recruitment Assessment (MSRA) exam" Computers in Biology and Medicine, volume 168 pages 1007794, 2024

[16]. Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. 2023. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. https:// github.com/nomic-ai/gpt4all. Accessed: 2023-06-03

[17]. Joshua Au Yeung, Zeljko Kraljevic, Akish Luintel, Alfred Balston, Esther Idowu, Richard J Dobson, and James T Teo. 2023. Ai chatbots not yet ready for clinical use. Frontiers in Digital Health, 5:60.

[18]. Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. arXiv preprint arXiv:2307.03109.
[19]. Xiang'Anthony' Chen, Jeff Burke, Ruofei Du, Matthew K Hong, Jennifer Jacobs, Philippe Laban, Dingzeyu Li, Nanyun Peng, Karl DD Willis, ChienSheng Wu, et al. 2023. Next steps for humancentered generative ai: A technical perspective. arXiv preprint arXiv:2306.15774
[20]. Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. arXiv preprint arXiv:2306.16092.