.

Research Proposal of Long-form Text Generation of Large Language Models

Jinlong Liu

Abstract

I Research Topic

Natural Language Processing (NLP) has made significant progress in recent years, with the development of large language models such as GPT-3, BERT, and T5. These models have shown impressive capabilities in tasks such as language modeling, text generation, and machine translation, among others. However, one common issue with these models is their ability to generate long-form text, such as writing stories, essays, or articles. It is a chanllenging task for models to garentee the coherence and meaningfulness of the generated text. This research proposal aims to address this issue by developing a deep learning model or using algorithms to improve the abilities of the model to generate coherent and meaningful text that resembles human-generated text.

Research nowadays is focus on the score function or decoder algorithms, such as top-k, Beam search, nucleus sampling, and so on.But I want to propose a potential solution to this issue by using Graph Neural Networks (GNNs) to model the relationships between words in a text and generate new text. The motivation behind using GNNs is to capture the complex dependencies and interactions among words in a text, which can help the model generate more coherent and contextually relevant text. The proposed model will be based on the newest model such as T5, and will be evaluated on text generation tasks in specific domains, such as writing stories.

II Review of the literature

In recent years, there have been a number of studies on text generation problems. This kind of problem is focused on the abilities of LLMs could focus on the coherence, interesting and relatedness of the generated text, which is a crucial part of LLMs. This ability need to structure a sequence of events, which are factual, fictional or a mixture of both, then create a coherent and detailed convey picture of the story [1].

To specific solve this kind of issue, using a better score function or better decoding algorithms has already become a common solution [2]. Additionally, There are several algorithms has been proposed in last year, such as Best-k Search, it is based on best-first search by adding parallel exploration, heap pruning and temporal decay part to enhance decoding performance. [3], another decoding method Minimum Bayes Risk Decoding(MERD) selects the least expected loss for a probabilistic model follow the result of utility or reward function. [4], then Penalty Decoding is a method to reduce the burden of penalty selection by addressing the overly short sentences caused by excessively penalties [5].

Furthermore, an improved method for storytelling Detailed Outline Control(DOC) is mentioned in this domain, it combines breadth-first expansion, event candidate generation, filtering and reranking part to enhance the performance of storytelling ability of LLMs [6]

The main

III Research objectives

IV Research strategy

V Schedule and budget

The Timeline for this research topic shown as Table 1

Time	Activities
First Term	Literature Review Understand the theory part in SeqGAN Progress report
Second Term	Modify SeqGAN Combine it with other Framework Progress report
Third Term	Training model then evaluate it Completion draft of paper Progress report
Fourth Term	Submission papers in optimized edition Progress report
Fifth Term	Keeping optimize developed framework Progress report
Sixth Term	Training model then evaluate it Compare ot with fromer data Progress report
Seventh Term	Completion draft of paper Submission papers in optimized edition
Eighth Term	Completion of thesis Submission of the thesis

Table 1: Schedule for Ph.D.

For budget part, the main cost will be model training, the cost mainly in building powerful server or renting a powerful cloud server.

References

[1] Y. Du and L. Chilton, "StoryWars: A dataset and instruction tuning baselines for collaborative story understanding and generation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 3044–3062. [Online]. Available: https://aclanthology.org/2023.acl-long.171

[2] A. Amini, R. Cotterell, J. Hewitt, L. Malagutti, C. Meister, and T. Pimentel, "Generating text from language models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, Y.-N. V. Chen, M. Margot, and S. Reddy, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 27–31. [Online]. Available: https://aclanthology.org/2023.acl-tutorials.4

[3] J. Xu, C. Xiong, S. Savarese, and Y. Zhou, "Best-k search algorithm for neural text generation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 12 385–12 401. [Online]. Available: https://aclanthology.org/2023.acl-long.692

[4] M. Suzgun, L. Melas-Kyriazi, and D. Jurafsky, "Follow the wisdom of the crowd: Effective text generation via minimum Bayes risk decoding," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 4265–4293. [Online]. Available: https://aclanthology.org/2023.findings-acl.262

[5] W. Zhu, H. Hao, and R. Wang, "Penalty decoding: Well suppress the self-reinforcement effect in open-ended text generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1218–1228. [Online]. Available: https://aclanthology.org/2023.emnlp-main.78

[6] K. Yang, D. Klein, N. Peng, and Y. Tian, "DOC: Improving long story coherence with detailed outline control," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 3378–3465. [Online]. Available: https://aclanthology.org/2023.acl-long.190