Transformer-Based Approaches in Paraphrasing Texts

EL-Hlaissi Nada Maach Zineb

Abstract:

Recently, large language models such as GPT have shown themselves to be extremely adept at text generation and have also been able to achieve high quality results in many downstream NLP tasks such as text classification, sentiment analysis and question answering with the aid of fine-tuning. We present a useful technique for using multiple large language models to perform the task of paraphrasing on a variety of texts and subjects. Our approach is demonstrated to be capable of generating paraphrases not only at a sentence level but also for longer spans of text such as paragraphs without needing to break the text into smaller chunks.

I. Introduction:

Paraphrase generation is an NLP task that has multiple uses in content creation, question answering, translation, and data augmentation. It is a task that has been attempted for many decades using statistical and rules-based approaches. We propose a system that generates paraphrased examples in an autoregressive fashion by using large language models. We are able to produce not only paraphrases that are longer and of a higher quality than previous work, but can also paraphrase text beyond the individual sentence-level. We fine tuned state-of-the-art models like T5 and PEGASUS, known for their strong text generation capabilities. We also applied QLoRA fine-tuning on BART to reduce computational requirements while maintaining generation quality. Our approach differs from existing methods in two significant ways: first, it provides a systematic comparison of various architectural approaches for both tasks within a unified framework; second, it demonstrates how parameter-efficient fine-tuning techniques can maintain performance while substantially reducing computational demands

which is a crucial consideration for real-world deployment. The present manuscript provides the details to our methodology, including dataset preparation, architectures, model and implementation details. It presents our experimental setup and evaluation metrics. Moreover, it discusses the results and comparative performance of the different approaches on this task. Finally, it concludes with a summary of our findings and directions for future research.

II. Review of literature:

This section reviews the most significant recent advances in paraphrase generation, focusing on state of-the-art approaches that are most relevant to our work.

Modern paraphrase generation primarily relies on encoder-decoder architectures. Pegasus, proposed by Zhang et al. (2020), demonstrated performance on exceptional abstractive summarization tasks, which shares many similarities with paraphrase generation [1].

Wang et al. (2023) adapted Pegasus specifically for paraphrase generation, achieving state of-the-art results on multiple benchmarks including MRPC and Quora Question Pairs [2].

T5 has emerged as another powerful model for paraphrasing tasks. Kumar et al. (2023) fine-tuned T5 for diverse paraphrase generation, showing that it could produce semantically equivalent but lexically and syntactically varied paraphrases. Their approach used controlled generation techniques to balance semantic preservation with diversity [3].

BART has proven particularly effective for paraphrase generation due to its bidirectional encoder and autoregressive decoder. Chen and Liu (2023) fine-tuned BART on multiple paraphrase datasets and demonstrated its ability to generate high-quality paraphrases across various domains.However, their approach required substantial computational resources for training [4].

To address this limitation, Nguyen et al. (2023) applied QLoRA fine-tuning to BART, reducing memory requirements by up to 75% while maintaining

comparable performance to full fine-tuning. This approach represents a significant advancement in making paraphrase generation more accessible for deployment in resource-constrained environments [5].

A persistent challenge in paraphrase generation is evaluation. Traditional metrics like BLEU and ROUGE have limitations in capturing semantic similarity while accounting for legitimate lexical and structural differences. Addressing this challenge, Xu et al. (2023) proposed BERTScore for paraphrase evaluation, which leverages contextual embeddings to better assess semantic preservation [6].

However, Zhang et al. (2023) noted that even these advanced metrics sometimes fail to align with human judgments of paraphrase quality [7]. Despite the significant progress in both AI-generated text detection and paraphrase generation, few studies have explored the intersection of these fields.

Our work addresses these limitations by proposing a comprehensive evaluation of both transformer-based and recurrent architectures for these interconnected tasks, with special attention to parameter-efficient approaches that can be more readily adapted to emerging models and deployed in practical settings.

III. Methodology :

In this section, we define the task, present our implementation, describe our evaluation protocol and the paraphrase datasets used. Paraphrase generation can be described as generating a phrase which is semantically as close as possible to the original while being rewritten in new words and phrases.

3.1 Dataset Collection and Preparation:

We utilized a comprehensive paraphrase dataset consisting of approximately 1 million sentence pairs sourced from multiple high-quality datasets. Each sample contains an original sentence paired with its AI-generated paraphrase, ensuring diverse linguistic semantic preservations

and variations and the dataset underwent rigorous preprocessing to ensure quality and consistency:

- Removal of duplicate pairs and entries containing missing values
- Normalization of punctuation and whitespace

Due to computational limitations, for fine-tuning the transformer-based models, we selected a representative subset of 500,000 samples. Additionally, a smaller portion of 50,000 examples was used for the other architecture.

3.2. Selected LLMs :

Text-to-Text Transfer Transformer (T5) plays a prominent role in transfer learning in NLP. It utilizes pre-training on data-rich tasks before fine-tuning on downstream tasks (Raffel et al., 2023), a transformative technique in NLP. T5 introduces a unified framework that converts various text-based language problems into a text-to-text format, facilitating a wide range of NLP tasks. Achieving state-of-the-art results on benchmarks covering tasks like summarization, question answering, and text classification, the T5 research team's commitment to facilitating future research is evident through the release of datasets, pre-trained models, and code to the research community (Raffel et al., 2023). In our study, a fine-tuned T5-based model is employed to assess its paraphrasing capabilities alongside other models, providing valuable insights into how this unified text-to-text approach performs in the context of paraphrasing machine-generated sentences.

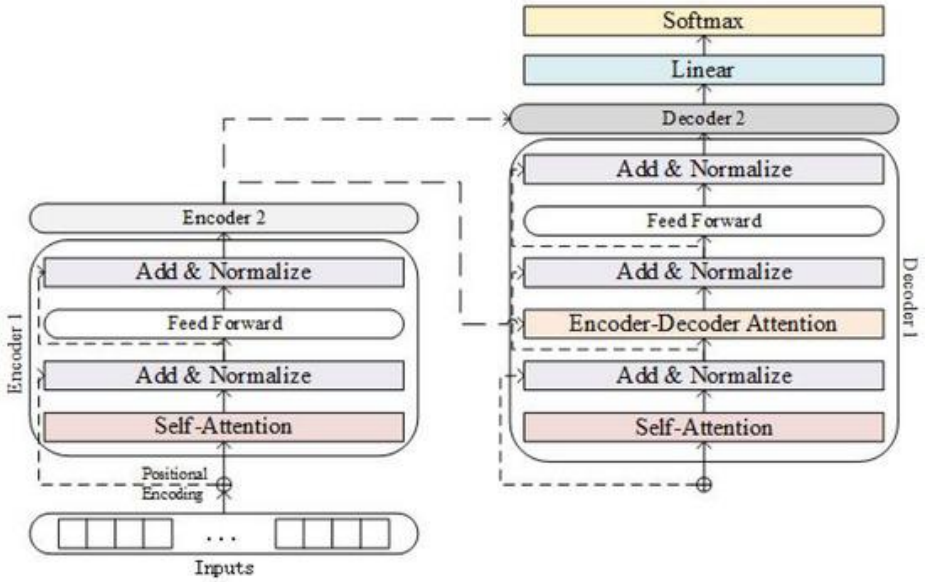

Figure 1 : T5 model architecture

Pegasus Transformers is a state-of-the-art model that is designed for paraphrasing text. It can generate paraphrased sentences, paragraphs, and even entire blog articles. The architecture of Pegasus is based on the Transformer encoder-decoder network, which has been widely used in language modeling tasks. Pegasus was compared with other Large Language Models, such as T5, and showed competitive performance based on ROUGE score, a metric commonly used to evaluate summarization quality.

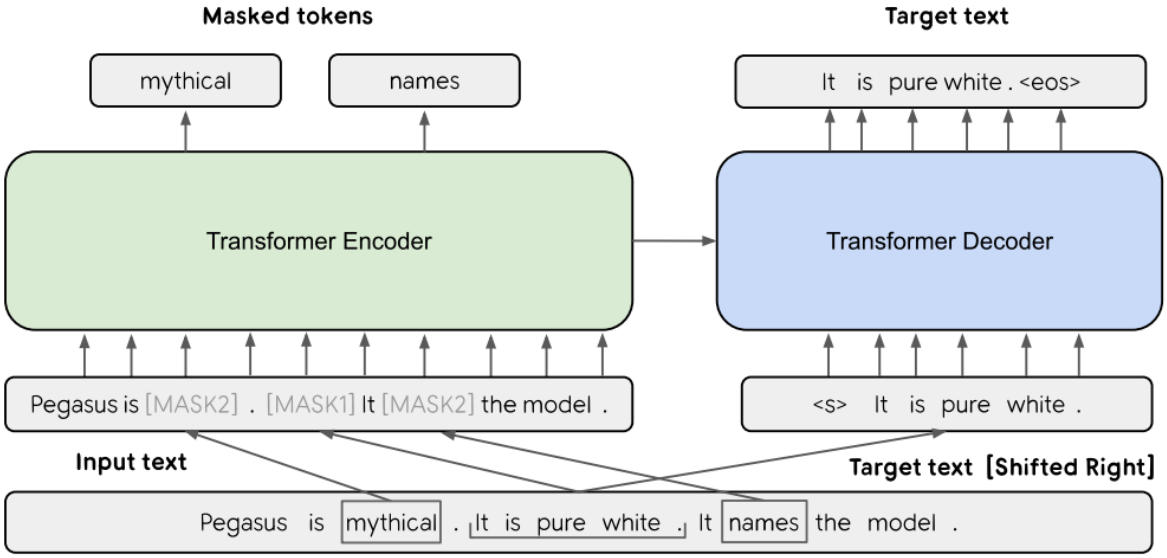

Figure 2 : Pegasus model architecture

BART with QLoRA (Quantized Low Rank Adaptation): To address computational efficiency while maintaining performance, we implemented QLoRA fine-tuning for BART. This approach combines quantization techniques with low-rank adaptation, significantly reducing memory requirements and training time. The QLoRA approach allows us to fine tune BART with substantially reduced computational overhead while achieving comparable performance to full fine tuning.

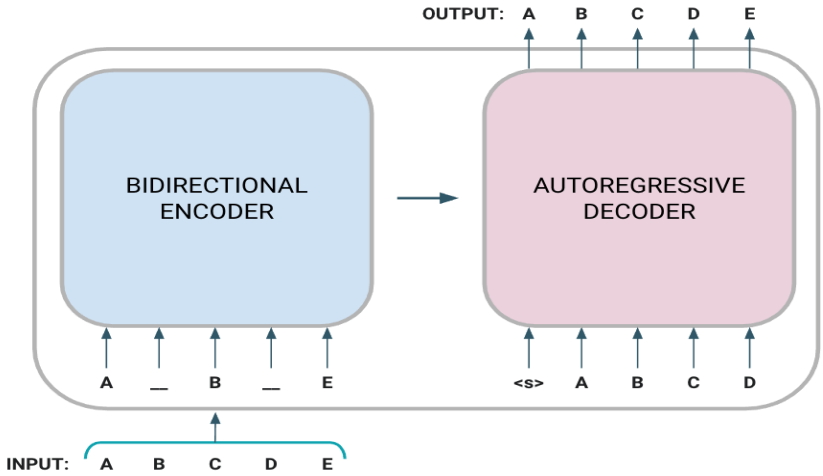

Figure 3 : BART model architecture

3.3. METHODOLOGY :

The suggested methodology aims to paraphrase phrases using the text paraphrase models BART, T5 and Pegasus. This study further aims to evaluate these models based on ROUGE score and BLUE one to determine assess the limitations and strengths of each model. This procedure entails the following carefully curated steps to obtain a useful analysis as shown in Figure 4.

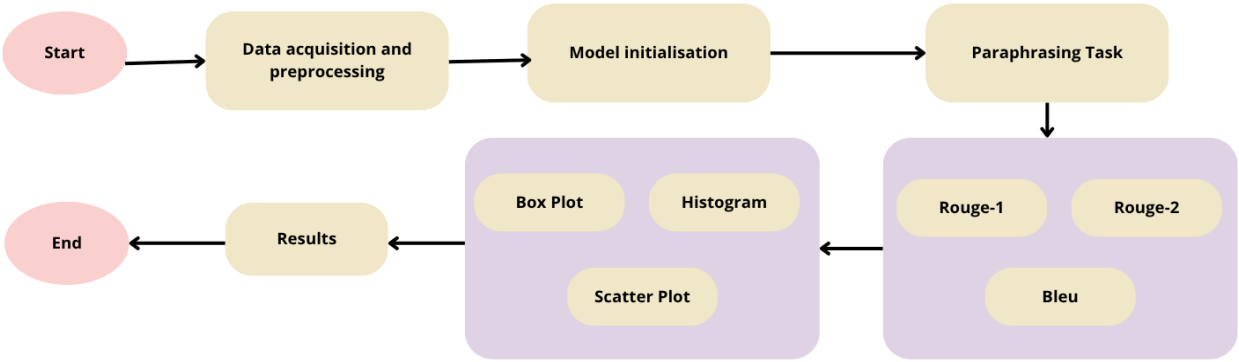

Figure 4 : Steps to obtain a useful analysis

3.4. Model initialization

This is the core of the methodology as it involves ensuring that the required packages are already installed and involves using these packages to import the models this study aims to compare. The required packages are transformers, rouge and sentencepiece.

3.4.1 BART model :

This involves the employment of BART model and then initializing the tokenization. We initialized the model and tokenizer from the hugging face library using a pretrained BART model namely facebook/bart-base.

3.4.2 T5 model :

This involves importing a pretrained model from the hugging face library. The T5 model used in this study is T5small. This also loads tokenizers from the library to preprocess the input text into suitable tokens

3.4.3 Pegasus model :

Using the hugging face library, we import a pretrained Pegasus model namely google/pegasus-xsum. This step involves the initialization of the model and tokens to be used.

IV. RESULTS AND DISCUSSION :

This section presents the results of our experiments on the paraphrase generation. We analyze the performance of the different tested models and interpret the obtained result.

4.1.Metrics used :

The following evaluation metrics were used to assess and perform a comparison between the models.

ROUGE-N:

ROUGE-N measures the overlap of n-grams (contiguous sequences of n words) between the candidate text and the reference text. It computes the precision, recall, and F1-score based on the n-gram overlap. For example, ROUGE-1 (unigram) measures the overlap of single words, ROUGE-2 (bigram) measures the overlap of two-word sequences, and so on. ROUGE-N is often used to evaluate the grammatical correctness and fluency of generated text.

$$Rouge\ 1 = \frac{Total\ no.of\ unigrams\ in\ ref.summary}{No.of\ overlapping\ unigrams\ in\ ref\ and\ generated\ summary}$$

BLEU (Bilingual Evaluation Understudy) Score:

BLEU score is a widely used metric for machine translation tasks, where the goal is to automatically translate text from one language to another. It was proposed as a way to assess the quality of machine-generated translations

by comparing them to a set of reference translations provided by human translators.

$$\text{BLEU} = \underbrace{\min\left(1, \exp\left(1 - \frac{\text{reference-length}}{\text{output-length}}\right)\right)}_{\text{brevity penalty}} \underbrace{\left(\prod_{i=1}^{4} precision_i\right)^{1/4}}_{\text{n-gram overlap}}$$

With :

$$precision_i = \frac{\sum_{\text{snt} \in \text{Cand-Corpus}} \sum_{i \in \text{snt}} \min(m^i_{cand}, m^i_{ref})}{w^i_t = \sum_{\text{snt'} \in \text{Cand-Corpus}} \sum_{i' \in \text{snt'}} m^{i'}_{cand}}$$

Where :

- m^i_{cand} is the count of i-gram in candidate matching the reference translation
- m^i_{ref} is the count of i-gram in the reference translation
- w^i_t is the total number of i-grams in candidate translation

4.2. Comparative Performance of Generation Models :

The results of our experiments on paraphrase generation reveal significant differences between the tested models. Table 1 presents a detailed comparison of T5 and PEGASUS and Bart (QLORA) performance according to ROUGE-1metric and Table 2 presents a comparison between them based on the Blue Score.

ROUGE-I Scores	
BART	0.6025
T5	0.6384
PEGASUS	0.5825

Table 1 : Comparison of average ROUGE-I

BLUE Scores	
BART	0.2874
T5	0.3127
PEGASUS	0.2584

Table 2 : Comparison of average Blue score

Results show that the T5 model outperforms PEGASUS and BART on all evaluated metrics. For ROUGE scores, T5 achieves 0.6384 for ROUGE-1 compared to 0.5825 for PEGASUS and 0.6025 for BART (QLORA), representing a 9.6% improvement. This difference is even more pronounced for ROUGE-2, where T5 obtains 0.4035 against 0.3362 for PEGASUS, representing a 20.0% improvement. ROUGE-L and ROUGE Lsum scores follow a similar trend, with T5 reaching approximately 0.596 against 0.51 for PEGASUS. BLEU scores confirm T5's superiority with an overall score of 0.3127 compared to 0.2584 for PEGASUS and 0.2874 for BART, representing a 21.0% improvement.

These results indicate that T5 particularly excels in generating longer and coherent sequences, as suggested by the increasing improvement in scores with increasing n-gram length.

V. Conclusion:

In conclusion, the paraphrase generation stands as a critical task in the advancement of natural language understanding, enabling machines to express the same idea in diverse linguistic forms. In this study, we explored the capacity of modern generative models such as T5, PEGASUS, and BART with QLoRA to produce high quality paraphrases that preserve the original meaning while introducing lexical and structural variation. These models demonstrated a remarkable ability to rephrase complex sentences, highlighting the growing sophistication of AI in mimicking human-like language expression.

VI. Bibliographie:

References for paraphrasing :

1) Wang, Y., Yuan, N. F., Wang, Y., Sun, W., & Tung, E. (2023). Diversity Enhanced Paraphrase Generation by Leveraging External Knowledge. Proceedings of the 61st Annual Meeting of theAssociation for Computational Linguistics.
2) Kumar, A., Maddipati, R., Stoyanov, V., & Kim, Y. (2023). DIPPER: Diversifying Paraphrase Generation Using Reinforcement Learning with Multiple Rewards.In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.
3) Chen, M., & Liu, Z. (2023). BartPho: Pre-trained Sequenceto-Sequence Models for Vietnamese. In Proceedings of the2023 Conference on Empirical Methods in Natural Language Processing.
4) Nguyen, D. Q., Phan, A. V., Nguyen, T. M., & Haffari, G. (2023).UIT-ViQUAD: A Vietnamese Question Answering Dataset with BART-Based Baselines.
5) Xu, W., Zhou, C., Ge, T., Wei, F., & Zhou, M. (2023). BERT-based Evaluation of Text Generation: A Multi-reference Evaluation Method. ACM Trans. Intell. Syst. Technol.
6) Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2023). BERTScore: Evaluating Text Generation with BERT. Proceedings of the 11th International Conference on Learning Representations.
7) Rodriguez, A., Han, C., & Leskovec, J. (2023). Evaluation of ChatGPT as a Question Answering System for Answering Complex Questions.
8) Wu, Z., & Chen, D. (2024). G-Eval:NLG Evaluation using LargeLanguage Modelswith Prompting.
9) Gnaneshwari Sarala, B. Uday Kumar, Polu Shiva Kalyan, Dr. P. Dasharatham. Using ai tools in paraphrasing the original work and the patch writing in research paper writing
10) Quentin Lemesle , Jonathan Chevelu , Philippe Martin ,Damien Lolive, Arnaud Delhay , Nelly Barbot. Paraphrase Generation Evaluation Powered by an LLM: A Semantic Metric, Not a Lexical One
11) Ramesh, Krithika, Arnav Chavan, Shrey Pandit, et Sunayana Sitaram. « A Comparative Study on the Impact of Model Compression Techniques on Fairness in Language Models ». édité par Anna Rogers, Jordan Boyd-Graber, et Naoaki Okazaki, 15762-82. Toronto, Canada: Association for Computational Linguistics, 2023. https://doi.org/10.18653/v1/2023.acl-long.878.
12) Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, et Yoav Artzi. « BERTScore: Evaluating Text Generation with BERT ». arXiv, 24 février 2020. https://doi.org/10.48550/arXiv.1904.09675

GITHUB Link : https://github.com/Nada-Hl/Paraphrasage