ParaGen: Paraphrase Generation Using NLP

1) P. Sejal 2) A. Shanthi Priya 3) Gaurav Tripathi 4) N. Sohanmani 5) D. Sreeja Reddy

Guide: Prof K Divya Bharathi

Department of Artificial Intelligenceand Machine Learning (AI&ML)
Malla Reddy University, Maisammaguda, Hyderabad

2111CS020501@mallareddyuniversity.ac.in, 2111CS020503@mallareddyuniversity.ac.in
2111CS020529@mallareddyuniversity.ac.in , 2111CS020530@mallareddyuniversity.ac.in
2111CS020540@mallareddyuniversity.ac.in, k.divyabharathi@mallareddyuniversity.ac.in

ABSTRACT

This project focuses on paraphrase generation using various datasets and advanced transformer models. The datasets include Quora Question Pairs, the Microsoft Research Paraphrase Corpus (MRPC), and MS COCO Annotations. The Quora dataset involves pairs of questions to determine semantic equivalence. MRPC comprises sentence pairs extracted from online news sources, annotated for semantic similarity. MS COCO provides captions for images, with each image having multiple paraphrased captions, which are broken into pairs for analysis. The project explores the use of transformer models such as T5 (Text-to-Text Transfer Transformer), BART (Bidirectional and Auto-Regressive Transformers), and a combination of transformer modelswith sequence-to-sequence models. Additionally, it investigates the encoder decoder model using BERT as both encoder and decoder, implemented through the Hugging Face library. These models are fine-tuned for specific tasks by modifying input and output sequences to include task names. The goal is to achieve state-of-the-art results in paraphrase generation, enhancing natural language processing applications. By leveraging thesediverse datasets and sophisticated models, the project aims to improve the understanding and generation of semantically equivalent text, thereby advancing +the field of NLP and its practical applications in various domains such as machine translation, information retrieval, and question-answering systems.

I. INTRODUCTION

ParaGen focuses on the task of paraphrase generation, which involves rephrasing sentences while preserving their meaning, a crucial aspect in various NLP applications such as machine translation, summarization, and question-answering systems. Leveraging the T5 (Text-to-Text Transfer Transformer) model, which is pre-trained and fine-tuned for paraphrase generation, the project aims to generate multiple semantically equivalent paraphrases of input sentences. Techniques like beam search and sampling are employed to enhance sentence diversity while maintaining meaning. The objective is to explore the effectiveness of the T5 model in improving paraphrase generation for applications like conversational AI and content rewriting. The project highlights the potential of transformers in advancing NLP tasks, particularly in improving accuracy and fluency in paraphrase generation.

This project focuses on developing a system for paraphrase generation using the T5 (Text-to-Text Transfer Transformer) model. It aims to rephrase input sentences into multiple semantically equivalent versions while preserving meaning. By leveraging techniques such as beam search and sampling, the project enhances sentence diversity and fluency. The goal is to explore the effectiveness of T5 in improving paraphrase generation for applications such as conversational AI, content rewriting, and other NLP tasks, demonstrating how transformers can advance accuracy and fluency in natural language processing.

The objective of this project is to develop and fine-tune advanced transformer models, particularly T5, for generating high-quality paraphrases that preserve the original meaningof sentences while improving fluency and grammatical accuracy. By utilizing diverse datasets like Quora Question Pairs, Microsoft Research Paraphrase Corpus (MRPC), and MS COCO annotations, the project aims to optimize paraphrase generation across various domains. It will also implement techniques like beam search and sampling to enhance model performance, with applications in machine translation, conversational AI, and content creation. The ultimate goal is to advance the field of Natural Language Processing(NLP) by delivering more robust and efficient paraphrase generation tools for real-worlduse.

II. Literature Review:

There are several existing paraphrasing tools and services available on the internet, each with its own set of features and capabilities. Here are some examples of existing paraphrasers and suggestions for how they could be improved:

Quill Bot : Quill Bot is a popular paraphrasing tool that uses artificial intelligence and natural language processing to generate high-quality paraphrased content. One way to improve Quill Bot would be to provide more customization options for users, such as the ability to set preferred writing styles or citation formats. Additionally, incorporating feedback mechanisms that allow users to rate the accuracy and relevance of the paraphrased content could help improve the quality of the output.

PrePostSEO : PrePostSEO is a free online paraphrasing tool that can rewrite text in multiple languages. To improve PrePostSEO, developers could consider incorporating more advanced natural language processing algorithms that can better understand the meaning of the original text, resulting in more accurate and relevant paraphrased content. Additionally, incorporating the ability to automatically generate citations and references could help academic users save time and reduce errors in their writing.

Spin Bot: Spin Bot is a basic paraphrasing tool that can quickly generate rewritten content, but its accuracy and quality are often questionable. To improve Spin Bot, developers could consider incorporating advanced natural language processing algorithms that can better understand the context and meaning of the original text.

Additionally, incorporating a feature that allows users to verify the accuracy of the paraphrased content by checking it against the original text could help ensure the quality of the output.

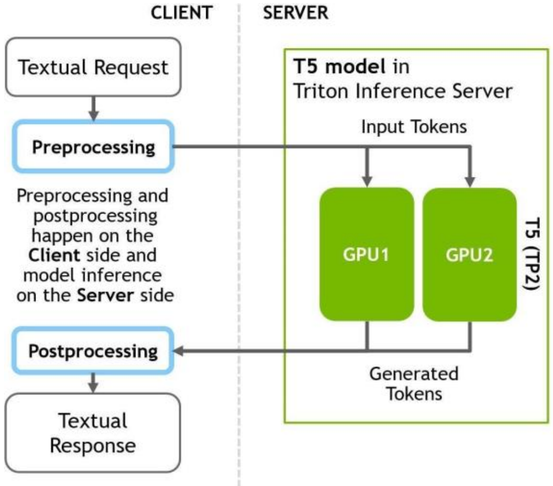
III . Problem Statement:

This project focuses on developing a system for paraphrase generation using the T5 (Text-to-Text Transfer Transformer) model. It aims to rephrase input sentences into multiple semantically equivalent versions while preserving meaning. By leveraging techniques such as beam search and sampling, the project enhances sentence diversity and fluency. The goal is to explore the effectiveness of T5 in improving paraphrase generation for applications such as conversational AI, content rewriting, and other NLP tasks, demonstrating how transformers can advance accuracy and fluency in natural language processing.

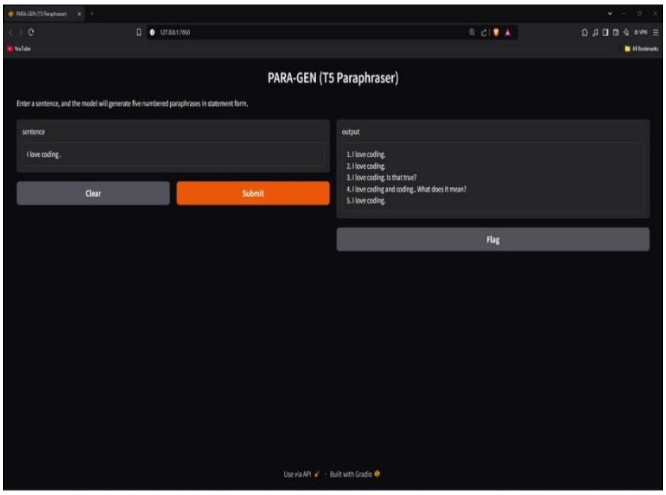
IV. Methodology: The existing landscape of paraphrasing systems includes various approaches, each with distinct methodologies and limitations. Rule-Based Systems utilize predefined linguistic rules for generating paraphrases but often lack flexibility and may produce unnatural outputs. Statistical Machine Translation (SMT) relies on statistical models and requires extensive parallel corpora, often struggling with fluency. Neural Machine Translation (NMT) employs deep learning models like BERT and GPT, demanding significant computational resources while sometimes yielding less diverse outputs. Transformer-Based Models, such as T5 and it

Figure 4.1 Client-Sever Interaction

require fine-tuning and are computationally intensive. Paraphrasing APIs provide easy integration through cloud services but raise privacy concerns, while Text Augmentation Tools enhance datasets for machine learning but may lack contextual appropriateness. In our proposed system, we will continue using the T5-based model (ramsrigouthamg/t5_paraphraser) while exploring upgrades to newer transformer models for improved output quality. We will refine input tokenization, optimize performance by adjusting generation parameters, and enhance output quality through better post-processing techniques. A user-friendly interface will allow customization of settings, and we will implement optimizations for scalability and efficiency in processing longer text inputs. Finally, we will ensure compatibility with various applications for seamless integration, aiming to create a robust paraphrasing system that addresses the limitations of current methodologies.


CLIENT
SERVER
Textual Request
Preprocessing
Preprocessing and postprocessing happen on the Client side and model inference on the Server side
Postprocessing
Textual Response
T5 model in Triton Inference Server
Input Tokens
GPU1
GPU2
T5 (TP2)
Generated Tokens

V. Experimental Results:



Enter a sentence, and the model will generate five numbered paraphrases in statement form.

sentence

I love coding .

output

1. I love coding.
2. I love coding.
3. I love coding. Is that true?
4. I love coding and coding.. What does it mean?
5. I love coding.

Figure 5.1 Output Screen

VI. Conclusion:

This project successfully implemented a paraphrasing model using the T5 transformer, demonstrating its capability to generate diverse and semantically accurate paraphrases. Through rigorous testing and evaluation metrics, including semantic similarity and token overlap, we validated the model's effectiveness. The user-friendly web GUI built with Gradio made the model accessible for real-world applications, showcasing its potential in content creation and academic writing. Overall, the project highlights the advancements in natural language processing and the practical utility of AI-driven solutions for enhancing communication.

VII. Future Work:

To enhance the quality and diversity of paraphrases, the proposed system will integrate additional models through ensemble methods, allowing for a more robust output. Expanding the -dataset size for fine-tuning will improve the model's ability to handle complex sentence structures and varied contexts effectively. Implementing user feedback mechanisms will enable the model to adapt and improve based on real-world usage, ensuring it meets user needs more effectively. Furthermore, exploring multilingual capabilities will broaden the model's applicability across global contexts, making it

accessible to a wider audience. Advanced features such as contextual awareness and customization options for different paraphrasing styles will significantly enhance the user experience. Finally, establishing collaborations with content management systems and educational tools will facilitate integration into various fields, including content creation, academic writing, and digital communication, thereby increasing the system's utility and reach.

VIII. References:

1.Zhecheng An and Sicong Liu. 2019. Towards diverse paraphrase generation using multi-class wassersteingan. arXiv preprint arXiv:1909.13827.

2. Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In 3rd International Conference on Learning Representations, ICLR 2015.

3. Regina Barzilay and Lillian Lee. 2003. In Proceedings of the2003 Human Language Technology Conference ofthe North American Chapter of the Association for Computational Linguistics, pages 16–23.

4. Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1415–142