International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0208
e ISSN: 2584-2854
Volume: 02
Issue: 05 May 2024
Page No: 1536-1540

Textual Compression Using Lamini-LM

Sanket Dudhmande[1], Shivam Golliwar[2], Ameya Bhagwat[3], Ram Ghiya[4], Archana Bhade[5]
[1,2,3,4]Department of Information Technology, GCE Amravati, Maharashtra, India.
[5]Assistant Professor- Department of Information Technology GCE Amravati, Maharashtra, India.
Email Id: *dudhmandes@gmail.com[1], shivam.golliwar@gmail.com[2], ameyabhagwat360@gmail.com[3], ramdghiya@gmail.com[4], hod.it@gcoea.ac.in[5]*

Abstract
In the age of digital overload, the ability to efficiently summarize large amount of text has become a critical capability. This project focuses on the development of a document summarization system that uses the power of the LaMini-LM model, a member of the Langchain family of large language models (LLMs). By using the natural language processing (NLP) techniques through the Langchain LLM technology it creates a robust and practical solution for automatically generating concise and informative summaries of text documents. The methodological approach of this project combines the strengths of NLP techniques, including the transformer architecture, the T5 model, and specialized components such as the T5 tokenizer and the Pipeline API. This strategic integration of technologies allows to create a comprehensive and efficient document summarization system that can effectively process and summarize a diverse range of text documents. The successful development and implementation of the document summarization system using the LaMini-LM model represents a significant advancement in the field of natural language processing.
Keywords: *Lamini-LM; Natural Language Processing (NLP); Pipeline; Transformer Architecture; T5 Model.*

1. Introduction

In today's world with the explosion of textual data the ability to efficiently process and summarize huge amount of text has become important. Thus, the need for tools that can extract the key data and present it in a precise and meaningful manner has become paramount [12]. Document summarization, the task of generating a smaller version of a text while preserving its essential content, is a vital capability that can greatly enhance productivity, decision-making, and knowledge management. To address the challenges of document summarization, this research project focuses on the utilization of the LaMini-LM model, a member of the Langchain family of large language models (LLMs). The Langchain LLM technology focuses on maintaining long-term contextual dependencies and efficient architecture, offers a promising approach to tackle the complexities of text summarization. The methodological approach combines the strengths of natural language processing techniques, including the transformer architecture, the T5 model and specialized components such as the T5 tokenizer and the Pipeline. This integration of latest technologies allows to create a comprehensive and efficient document summarization solution that can effectively process and summarize a diverse range of text documents [13].

1.1. Transformer

Transformers are a type of deep learning model that rely on a unique mechanism called self-attention to capture the contextual relationships within input text. This self-attention mechanism allows transformers to understand the meaning and semantics of language in a more sophisticated manner compared to traditional models, making them highly effective.[2] The transformer architecture is characterized by its multi-headed attention layers, feed-forward neural networks, and residual connections. These components work together to create a powerful and flexible model that can process and generate text with high accuracy. The attention mechanism, in particular, enables transformers to focus on the most relevant parts of the input when producing output, which is crucial

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0208
e ISSN: 2584-2854
Volume: 02
Issue: 05 May 2024
Page No: 1536-1540

for tasks like document summarization where the model needs to identify and synthesize the key information Figure 1. [2]

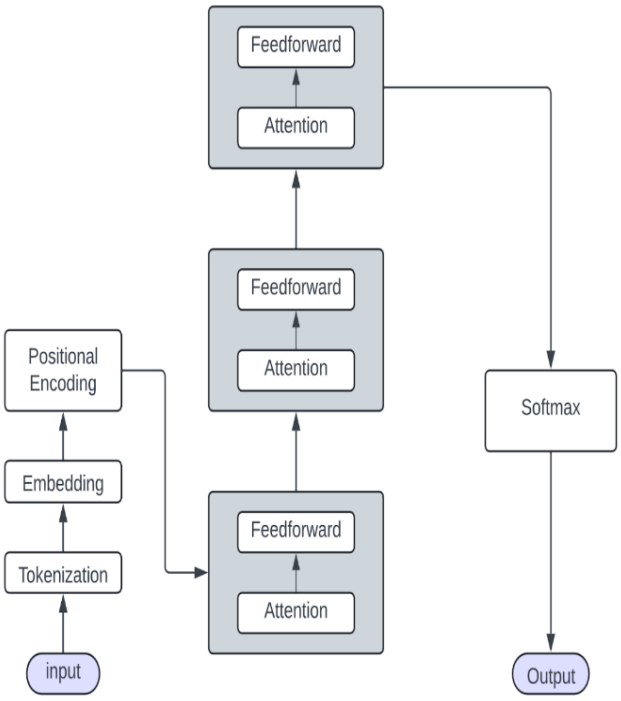

Figure 1 Transformer Model

1.2. Lamini-LM

LaMini-LM is a collection of language models created by extracting knowledge from instruction-tuned Large Language Models (LLMs) into smaller ones. The process involves the development of a large instruction dataset consisting of 2.58 million examples, covering diverse topics sourced from existing datasets and generated instructions. This dataset is augmented using techniques like Example-Guided Instruction Generation and Topic-Guided Instruction Generation. Responses for each instruction are generated using GPT-3.5-turbo, resulting in the LaMini instruction dataset.[1] After creating the dataset, multiple smaller language models with varying sizes and architectures are fine-tuned using it. The performance of these models is evaluated across various Natural Language Processing (NLP) benchmarks and through human assessment Figure 2. [3] LaMini-LM can be easily downloaded onto one's system and run on a CPU, offering accessibility and convenience compared to other models that often require heavy computational resources and specific hardware for deployment.

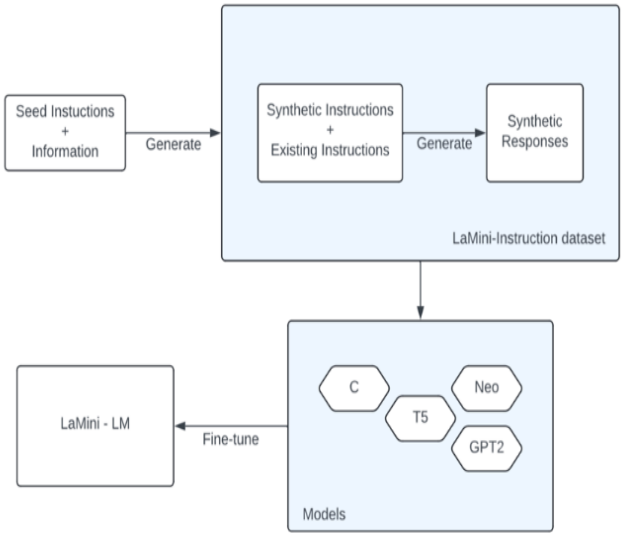

Figure 2 Overview of Lamini – LM [1]

2. Libraries Utilized

2.1. Langchain

The Langchain family of large language models (LLMs) represents a novel approach to developing highly capable and efficient natural language processing (NLP) models [11]. These Langchain LLMs, of which LaMini-LM is a member, are characterized by their multi-layer transformer-based architecture that allows them to capture complex contextual information from text data. A defining feature of Langchain LLMs is their ability to maintain a 'long chain' of contextual dependencies, enabling them to comprehend and generate text with a deeper understanding of the underlying semantics. In the context of the document summarization project, it is employed to generate concise and informative summaries of input text documents. The Langchain LLM technology underpinning LaMini-LM allows the model to understand the overall context and structure of the documents, enabling it to extract the most salient information and synthesize it into coherent summaries. [4] The `Recursive Character Text Splitter` from the `langchain.text_splitter` module is designed to break down longer text documents into smaller, more manageable chunks, which is crucial when

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0208
e ISSN: 2584-2854
Volume: 02
Issue: 05 May 2024
Page No: 1536-1540

working with large or complex input documents. By splitting the text recursively, the splitter can preserve the overall structure and context of the document, ensuring that the summarization process can effectively capture the key information. [4] To obtain the text content for the summarization task, it utilizes two document loaders from the `langchain.document_loaders` module: `PyPDFLoader` and `DirectoryLoader`. The `PyPDFLoader` is specifically designed to extract text from PDF documents, allowing the system to process a wide range of input formats. The `DirectoryLoader`, on the other hand, enables the loading of multiple documents from a directory, providing a convenient way to handle a collection of text-based files. The core of the document summarization process is handled by the `load_summarize_chain` function from the `langchain.chains.summarize` module. This function loads a pre-trained summarization model, which is then used to generate concise and informative summaries of the input documents [14].

2.2. T5 Tokenizer

To effectively process and utilize the T5 model within the document summarization project, a specialized tokenizer is employed. The T5 Tokenizer is responsible for converting the input text into a format that the T5 model can understand. This tokenizer breaks down the input text into a sequence of tokens, which can represent individual words or sub-word units. The tokenization process ensures that the input data is properly formatted and can be efficiently processed by the T5 model, laying the groundwork for the summarization task. [8] One of the key capabilities of the T5 model that makes it well-suited for the document summarization project is its ability to perform conditional generation. Conditional generation refers to the task of generating output text based on a given input. In the context of the document summarization, this means that the T5 model can take a longer input document and generate a concise summary as the output. This conditional generation capability is crucial for the success of the document summarization project using the LaMini-LM model, as it allows the system to effectively extract the key information from the input and present it in a compact form [9]. To streamline the integration and usage of the T5 model, tokenizer, and conditional generation capabilities, the project likely employs the Pipeline API provided by the Hugging Face Transformers library. The Pipeline abstraction allows for the seamless integration of these various components into a cohesive workflow, making it straightforward to apply the document summarization functionality to the input documents. By encapsulating the complex underlying processes, the Pipeline approach simplifies the usage of the T5 model and its associated components, enabling the researchers to focus on the overall effectiveness of the document summarization system. [5]

3. Overview of Implementation

In the context of our text summarization project, the synthesis step is where the generated summary output is based on the processed and transformed information from the input text in Figure 3.

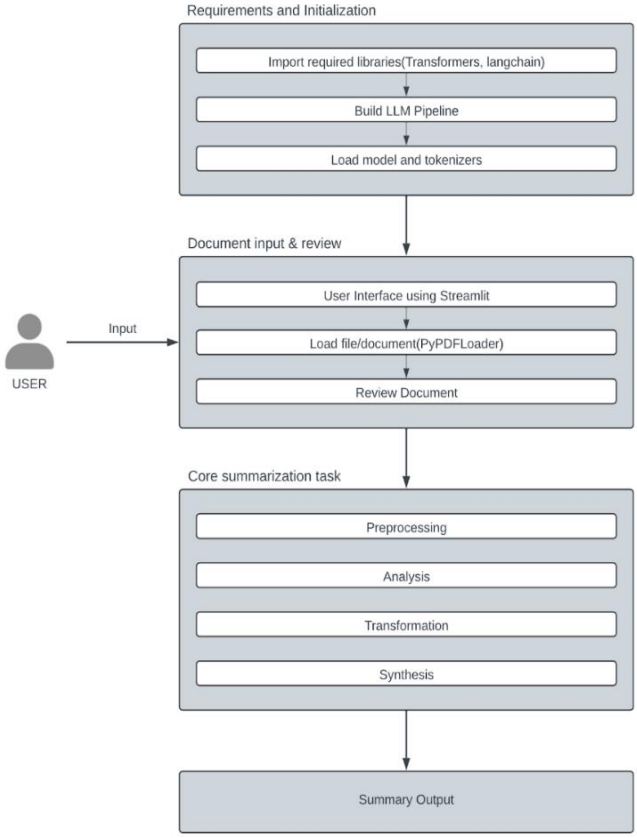

Figure 3 Implementation Steps

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0208
e ISSN: 2584-2854
Volume: 02
Issue: 05 May 2024
Page No: 1536-1540

This step involves the aggregation and condensation of key points, ideas, and information extracted during the analysis and transformation stages. During synthesis, the model leverages various techniques such as sentence compression, abstraction, and relevance ranking to generate a concise and coherent summary that captures the essential elements of the original text. Additionally, the synthesis step may involve the application of natural language generation (NLG) techniques to ensure that the summary output is fluent and coherent. This may include restructuring sentences, paraphrasing, and ensuring grammatical correctness to enhance readability and comprehension [10].

4. Evaluation

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score is a metric used to evaluate the quality of summaries and machine-generated text by comparing them to reference summaries. It measures the overlap between the n-grams (sequences of words) in the generated text and those in the reference summaries. A higher ROUGE score indicates a better alignment between the generated text and the reference summaries, reflecting improved summarization performance [15]. [6] We tested the efficacy of the model by comparing the summaries generated by the model with sample summaries available. The model produced sufficient score and brief summaries. "ROUGE-1" assesses the match of single-word sequences between the generated summary and the reference. "ROUGE-2" evaluates the agreement of two-word sequences between the generated summary and the reference. "ROUGE-L" gauges the longest common subsequence between the generated summary and the reference, accounting for the longest shared sequence of words

Table 1 Obtained ROGUE Score

ROGUE	Score
ROGUE - 1	0.300
ROGUE - 2	0.185
ROGUE - L	0.222
ROGUE -Lsum	0.222

"ROUGE-LSUM" assesses the longest common subsequence between the generated summary and the reference summary, focusing on the longest matching sequence of words in Table 1. [7]

Acknowledgement

Prof. A.W. Bhade, Head of Department of Information Technology, Government College of Engineering, Amravati, has been a tremendous source of support and guidance to the authors throughout this project.

Conclusion

The successful development and implementation of the document summarization system using the LaMini-LM model represents a significant advancement in the field of natural language processing. By leveraging the power of the Langchain LLM technology, the researchers have created a highly capable and efficient tool that can effectively extract the key information from text documents and present it in a concise and informative manner. The core of the document summarization system, the LaMini-LM model, is a member of the Langchain family of large language models. The Langchain LLM technology, with its focus on maintaining long-term contextual dependencies and efficient architecture, has enabled the LaMini-LM model to excel in the task of document summarization. The ability to capture the nuanced meaning and structure of the input text, combined with the model's compact size and scalability, make LaMini-LM a highly versatile and practical solution for real-world applications. The successful implementation of the document summarization project was made possible by the seamless integration of various key components, including the T5 model, the transformer architecture, the T5 tokenizer, and the Pipeline API. The ability to generate concise summaries of text documents can enhance decision-making, and knowledge management in a various domain such as business, academia, and government. The LaMini-LM-based document summarization system can lead to significant improvements in productivity and information processing making it a valuable asset in today's data-driven world.

References

[1]. LaMini-LM: A Diverse Herd of Distilled

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0208
e ISSN: 2584-2854
Volume: 02
Issue: 05 May 2024
Page No: 1536-1540

Models from Large-Scale Instructions Minghao Wu1, Abdul Waheed1Chiyu Zhang, Muhammad Abdul-Mageed, Alham Fikri Aji

[2]. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu .

[3]. Attention Is All You Need :Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

[4]. Asyrofi, R., Dewi, M. R., Lutfhi, M. I., & Wibowo, P. (2023). Systematic Literature Review: Langchain Proposed. In 2023 International Electronics Symposium (IES). IEEE.DOI:10.1109/IES59143.2023.10242497.

[5]. Sudharson, D., Thrisha Vaishnavi, K. S., Hariprakassh, S., Abiram, B., Saranya, K., & Tanwar, P. (2023). An Abstractive Summarization and Conversation Bot using T5 and its Variants. In 2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT) (pp. 1-8). IEEE. DOI:10.1109/ICAICCIT60255.2023.10465740.

[6]. ROUGE: A Package for Automatic Evaluation of Summaries Chin-Yew Lin Information Sciences Institute University of Southern California 4676 Admiralty Way Marina del Rey, CA 90292 cyl@isi.edu

[7]. Irina Radeva, Ivan Popchev, Lyubka Doukovska, Miroslava Dimitrova. "Web Application for Retrieval-Augmented Generation: Implementation and Testing", Electronics, 2024

[8]. Semantic Feature Verification in FLAN-T5,Apr 2023,LicenseCC BY 4.0 Siddharth SureshSiddharth SureshKushin MukherjeeTimothy T RogersTimothy T Rogers

[9]. Awasthi, K. Gupta, P. S. Bhogal, S. S. Anand, and P. K. Soni, "Natural Language Processing (NLP) based Text Summarization - A Survey," in Proceedings of the IEEE International Conference on Information and Communication Technology (ICICT), 20-22 January 2021, DOI: 10.1109/ICICT50816.2021.9358703.

[10]. R. Rahul, S. Adhikari, and M. Monika, "NLP based Machine Learning Approaches for Text Summarization," in Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC 2020), ISBN: 978-1-7281-4889-2, IEEE Xplore Part Number: CFP20K25-ART.

[11]. M. BM, and P. AL, "SUMMARIZING TEXTS USING NLP BASED APPROACHES," International Research Journal of Modernization in Engineering Technology and Science, vol. 05, no. 08, August 2023, e-ISSN: 2582-5208.

[12]. See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL).

[13]. Zhanlin Ji, Ivan Ganchev, Jiawen Jiang, Haiyang Zhang, Chenxu Dai, Qingjuan Zhao & Hao Feng, “Enhancements of Attention-Based Bidirectional LSTM for Hybrid Automatic Text Summarization”, 2021, pp. 3110143

[14]. TextRank: Bringing Order into Texts Rada Mihalcea and Paul Tarau Department of Computer Science University of North Texas rada,tarau

[15]. Ofir Press and Lior Wolf. Using the output embedding to improve language models. arXiv preprint arXiv:1608.05859, 2016