The Comparison and Analysis of Different Large Language Models in Text Summarization Task

Shengzhi Chen
School of Big Data & Software Engineering, Chongqing University, Chongqing, China

Keywords: Text Summarization, Large Language Models, CNN/Daily-Mail Dataset, ROUGE.

Abstract: This study primarily focuses on the evaluation of various large language models' performance in text summarization task, especially as their significance is increasingly apparent in applications like news media, academic research, and business intelligence. The main objective is to evaluate the performance of different models through both quantitative and qualitative methods. Specifically, author selected the test set from the CNN/Daily-mail dataset and used Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and Bidirectional Encoder Representation from Transformers (BERT) Score as evaluation metrics. After fine-tuning the parameters for each model, author conducted a detailed analysis of their predictive performance, including scoring and evaluation using Generative Pre-trained Transformer (GPT)-4. Conducted on the CNN/Daily-mail dataset, the experimental results show that, without any constraints, the summaries generated by GPT-3.5 perform best in terms of accuracy and completeness but are slightly lacking in conciseness. Summaries generated by Pre-training with Extracted Gap-sentences for Abstractive Summarization (Pegasus)-large are relatively shorter and mostly accurate but occasionally include redundant information. Fine-tuned Language Net Text-To-Text Transfer Transformer (Flan-T5) models produce more concise summaries but fall short in terms of accuracy and completeness. The outcomes of this research not only enrich the empirical understanding of text summarization but also offer directives for those employing large language models in this task.

1 INTRODUCTION

Artificial intelligence has developed to address many Natural Language Processing (NLP) tasks. However, some tasks like text summarization face performance challenges. Text summarization aims to provide accurate, concise overviews of lengthy documents, focusing on essential details while preserving context (Mridha et al 2021). The vast amount of textual data originates from web resources, news articles, books, scientific papers, and other documents. With the exponential increase in textual content on the Internet and other archives, users invest significant time and effort in searching, reading, and understanding information (El-Kassas et al 2021). Text summarization is a solution, enabling individuals to quickly extract key details from massive datasets, streamline content and present information efficiently. Large language models are an effective solution. This paper will provide a comprehensive overview and practical references on the performance of existing large language models in text summarization.

Text summarization solutions have evolved sequentially into four types: rule-based methods, statistical learning-based methods, deep learning-based methods, and large language model-based methods. Rule-based methods use artificially designed features for important and relevant text organization. They are simple and easy to implement but often result in low-quality summaries due to a lack of understanding of crucial information. Representative works include H.P. Luhn's method, which extracts essential sentences using word frequency and sentence position (Luhn 1958), and H.P. Edmundson's method, which calculates sentence importance using word frequency, sentence position, theme words, and cue words (Edmundson 1969). Statistical learning-based methods build statistical models between text and summarization using statistical features. They can use lexical-level statistical information but are unable to capture macro-level text information and deeper features. Notable works include Kupiec et al.'s Maximum Entropy Model-based approach (MEM), which evaluates sentence importance using sentence features and manual annotations (Kupiec et al 1995),

Chen, S.
The Comparison and Analysis of Different Large Language Models in Text Summarization Task.
DOI: 10.5220/0012799300003885

In *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, pages 417-421
ISBN: 978-989-758-705-4

and Conroy and O'Leary's Hidden Markov Model-based method (HMM), which selects the most appropriate sentences based on features and posterior probabilities and then generates summaries based on transition probabilities between sentences (Conroy et al 2001). Deep learning-based methods use deep neural networks to learn potential nonlinear mapping relationships from text to abstract. They can effectively represent text and generate summaries with deeper features, but these methods require large labeled datasets and may suffer from generation errors. Prominent works include Rush et al.'s neural network model employing a local attention mechanism (Rush et al 2015) and Nallapati et al.'s model uses sequence-to-sequence recurrent neural network (RNN) to tackle key challenges in abstractive summarization and achieves state-of-the-art performance on two corpora (Nallapati et al 2016). Large language model-based methods use pre-trained language models for text summarization, which can make full use of unlabeled data and general language knowledge to reduce generation errors. Representative works include Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformer (GPT)-4, Fine-tuned Language Net Text-To-Text Transfer Transformer (Flan-T5), and Pre-training with Extracted Gap-sentences for Abstractive Summarization (Pegasus). GPT-4, developed by OpenAI, is a large multimodal model founded on the Transformer architecture, achieving human-level performance in various tasks (OpenAI 2023). Pegasus, introduced by Zhang et al., predicts deleted or masked sentences using the remaining ones during pre-training to enhance summarization performance (Zhang et al 2020).

The main objective of this study is to compare the performance of various large language models in the task of text summarization. Specifically, first, author selected the test set from the CNN/Daily-mail dataset and chose Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and BERT Score as evaluation metrics. Second, author set special configurations for the parameters of different models and then calculated various metrics. Third, author analyzed and compared the predictive performance of different models represented by different evaluation metrics on this specific dataset. In addition, the author also used the GPT-4 model to score summaries generated by different models. The experimental results demonstrate that, under unconstrained conditions, summaries generated by GPT-3.5 are the most accurate and comprehensive but are not succinct. Pegasus-large can generate shorter and fairly accurate summaries, although it occasionally produces extraneous information. The Flan-T5 series of models are more concise but lack accuracy and completeness. This study can offer empirical insights and provide guiding references for the application and research of various large language models in text summarization task.

2 METHODOLOGY

2.1 Dataset Description and Preprocessing

The CNN/Daily-mail dataset serves as a popular standard in the field of text summarization, primarily containing news articles from Cable News Network (CNN) and Daily-mail along with their corresponding highlighted summaries (Hermann et al 2015). Each article in the dataset comes with a highlighted summary intended to capture the core content and message of the article, aiding researchers in evaluating the performance of automated text summarization algorithms. Author used the test set from version 3.0.0 of the CNN/Daily-mail dataset, extracting articles and their corresponding highlighted summaries for performance evaluation.

2.2 Proposed Approach

The primary aim of this study is to assess the performance of various large language models in text summarization task. To achieve this objective, author selected a range of distinctive large language models, such as GPT-3.5, Flan-T5, and Pegasus, to compare the performance differences and application environments that different features bring. Additionally, the process of generating article summaries is automated; however, some models, like GPT-3.5, currently require human guidance in the form of a specific prompt to complete the text summarization task. On the other hand, certain models like Flan-T5 and Pegasus can perform specific tasks through settings. Finally, the system calculates various metrics based on the generated summaries and the highlighted summaries in the original dataset. This is complemented by other evaluation methods, such as the GPT-4 scoring used in this study, to comprehensively reflect the performance of different language models. The process is shown in the Fig. 1.

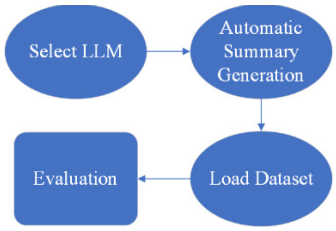

Figure 1: Flow Chart Process (Original).

2.2.1 Basic Architecture: Transformer

Transformer is currently the most popular deep learning model architecture in the NLP field, consisting of an encoder and a decoder, as shown in the Fig. 2. The self-attention mechanism stands out as the model's core feature, facilitating distinct attention allocation for each part of the input data. This enables language models to capture long-distance dependencies in the text and produce coherent and relevant text output.

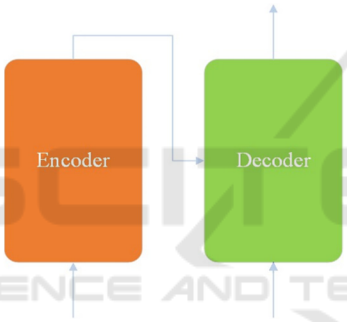

Figure 2: Transformer Architecture (Original).

2.2.2 Large Language Models

GPT-3.5 is a large language model introduced by OpenAI. It inherits the features of GPT-3, eliminating the need for fine-tuning for specific tasks, and employs Reinforcement Learning from Human Feedback (RLHF) to continually fine-tune the pre-trained language model, enhancing its text generation and understanding capabilities. Additionally, GPT-3.5 can handle multiple tasks; it has zero-shot learning abilities, allowing it to execute tasks without explicit task examples, and it also supports few-shot learning, guiding its output based on provided examples. GPT-3.5 is based on Transformer's Decoder model. In this study, author first uses OpenAI's official API and sets an appropriate prompt, then loads the dataset into the model to obtain text summaries, and finally evaluates various metrics.

Flan-T5 is a model proposed by Google that is optimized based on T5. T5 treats all NLP tasks as text-to-text conversion problems, allowing it to adapt to various NLP tasks without the need for task-specific models. Flan-T5, building on T5, has been fine-tuned for large-scale tasks, endowing the model with strong generalization capabilities and ultimately achieving "One Model for ALL Tasks". Flan-T5 is based on Transformer's Encoder-Decoder model. In this study, author first deploys the model locally and sets it for the text summarization task, then loads the test set into the model to obtain text summaries, and finally evaluates various metrics.

Pegasus is a language model specifically designed for generating abstractive summaries, proposed by Google. During its pre-training process, Pegasus randomly masks sentences from the document and asks the model to recover these masked sentences, enabling it to capture the core information of the text. Additionally, Pegasus can be fine-tuned for various summarization tasks and requires only 1,000 samples to achieve performance comparable to human-generated summaries. Pegasus is based on Transformer's Encoder-Decoder model. In this study, author first deploys the model locally, then loads the test set into the model to obtain text summaries, and finally evaluates various metrics.

2.3 Implementation Details

In the experiments conducted for this study, author uses Python version 3.10. The computational process was accelerated on a machine equipped with an NVIDIA RTX 4090 GPU and executed using an i9-13980HX CPU. When invoking the GPT API, author finalized the prompt as 'Summarize the text: '. To ensure diversity in the summaries generated by various models, the temperature parameter is set to 1. To have the model consider the entire text content, the top-p parameter is set to 1. For the evaluation of the ROUGE metrics, author opts to perform stemming before calculation. Additionally, no constraints are imposed on the maximum length of generated summaries, allowing the models to produce comprehensive and general outputs.

3 RESULTS AND DISCUSSION

This chapter analyzes the impact of model size on the performance of text summarization task and discusses the performance of different language models, as well as the characteristics and differences in the summaries they generate.

As illustrated in Fig. 3, based on the ROUGE metrics, both Flan-T5-base and Flan-T5-large

outperformed Flan-T5-small across all ROUGE metrics. Particularly in the case of ROUGE-2 and ROUGE-L, these two models demonstrated a stronger ability to capture multi-word phrases and long sentence structures in sentences. Additionally, Flan-T5-large, likely due to its superior ability to semantically understand text, generates summaries with more flexible wording and phrasing, which resulted in slightly lower scores in ROUGE-1 and ROUGE-2 compared to Flan-T5-base. This indicates that Flan-T5-large is better at preserving important information from the input text, while Flan-T5-base generates summaries that are most similar to the reference summaries. Flan-T5-small scored the lowest on all ROUGE metrics, indicating a significant difference between its generated summaries and the reference summaries.

Using the BERT Score as a gauge, the performance of the three models is fairly consistent, but Flan-T5-large still came out on top, achieving the highest BERT Score of 0.8763, suggesting that it can generate summaries most semantically similar to the reference summaries. Flan-T5-base followed closely with a BERT Score of 0.8745, indicating that it too can generate semantically relevant summaries. Flan-T5-small scored the lowest on BERT Score, registering at 0.8715, suggesting some semantic differences between its generated summaries and the reference summaries.

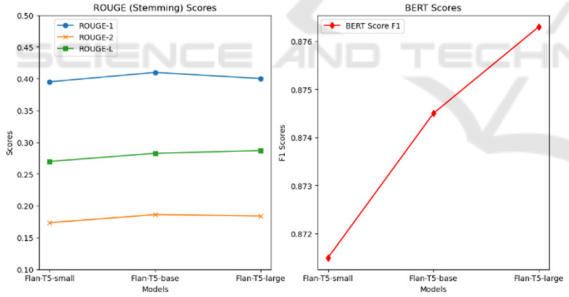

Figure 3: ROUGE and BERT Score by Model Size (Original).

As shown in Table 1, from the perspective of the ROUGE metrics, Flan-T5-large scored the highest on all ROUGE metrics, indicating that it can generate summaries most similar to the reference summaries. GPT-3.5 slightly outperformed Pegasus-large in ROUGE-1 and ROUGE-L, but fell behind in ROUGE-2. This suggests that GPT-3.5 and Pegasus-large have different focuses when it comes to retaining important information from the input text; the former emphasizes words and sentences, while the latter focuses more on phrases.

Looking at the BERT Score metric, Flan-T5-large achieved the highest score, reaching 0.8763, indicating that it can generate summaries most semantically similar to the reference summaries. GPT-3.5 scored the second highest in BERT Score, with a score of 0.8719, suggesting that it can also generate semantically relevant summaries. Pegasus-large scored the lowest in BERT Score, at 0.8630, indicating that the summaries it generates have some semantic differences compared to the reference summaries.

Table 1: ROUGE and BERT Score Across Models.

Model	ROUGE-1	ROUGE-2	ROUGE-L	F1 Score
GPT-3.5	0.3597	0.1412	0.2294	0.8719
Flan-T5-large	0.4003	0.1840	0.2869	0.8763
Pegasus-large	0.3504	0.1447	0.2301	0.8630

As shown in Fig. 4 and Table II, GPT-3.5 performs the best in retaining factual details from the original text, providing the most complete summaries that nearly cover all important information from the source material. The summaries it generates are well-structured and easy to understand, although they sometimes include some extra details. The Flan-T5 series of models usually produce relatively concise summaries, but occasionally introduce inaccurate information. They are not as comprehensive as GPT-3.5 and often miss out on key information, making their summaries harder to read. Among them, the small model performs the worst in terms of accuracy. Pegasus-large can generate concise summaries but still sacrifices some level of accuracy and completeness.

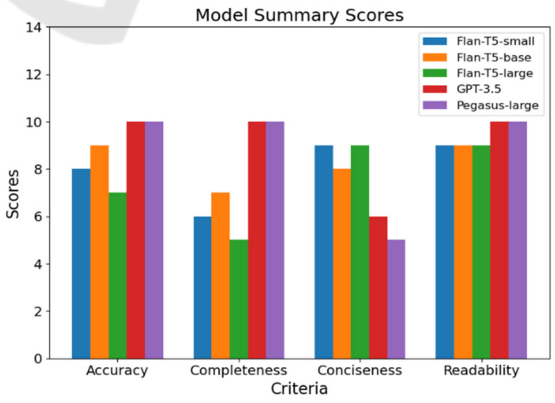

Figure 4: GPT-4's Model Ratings (Picture credit: Original).

Table 2: GPT-4's Model Evaluation.

Model	Score	Opinion
Flan-T5-small	5.8	Has significant issues in accuracy and completeness, lacking key information.
Flan-T5-base	6.5	Better in accuracy compared to the small model, but lacking in completeness.
Flan-T5-large	6.8	Good in accuracy but lacking in completeness.
GPT-3.5	8.9	Excellent in accuracy, coherence, and completeness.
Pegasus-large	7.2	Good in accuracy and coherence, but occasionally adds extra information.

4 CONCLUSION

This study compares the performance of different large language models in text summarization tasks. Extensive experiments are conducted to evaluate these models. The results show that the size of the model has a significant impact on its performance in text summarization. Generally, larger models tend to perform better. On the ROUGE and BERT Score metrics, Flan-T5-large performed the best, indicating that it can generate summaries most similar and semantically relevant to the reference summaries. GPT-3.5 performed second best, suggesting that it can also generate fairly similar and semantically relevant summaries. Pegasus-large performed the worst, indicating a significant deviation from the reference summaries. This may be related to the training methods and datasets used by the models. According to the evaluation of GPT-4, GPT-3.5 performs excellently in terms of accuracy, completeness, and readability, with only minor compromises in conciseness. The text summarization performance of Pegasus-large is average, and Flan-T5 models perform relatively poorly, especially in terms of accuracy and completeness. Therefore, the GPT-3.5 model can be chosen when the most accurate and complete summary is needed while Pegasus-large and Flan-T5-base could be suitable options when a shorter summary is needed but key information should be retained. In addition, Flan-T5-large or Flan-T5-small can be chosen when the shortest summary is desired at the expense of some information completeness. In the future, other mainstream large language models such as GPT-4, Llama 2, etc., will be considered as research objectives for the next stage. Additionally, further linguistic analysis of the text features of different models will be considered.

REFERENCES

Mridha, F. Muhammad, et al, “A survey of automatic text summarization: Progress, process, and challenges,” IEEE Access, vol. 9 2021, pp. 156043-156070

El-Kassas, S. Wafaa, et al, “Automatic text summarization: A comprehensive survey,” Expert systems with applications, vol. 165, 2021, pp. 113679.

Luhn, P. Hans, “The automatic creation of literature abstracts,” IBM Journal of research and development, vol. 2, 1958, pp. 159-165

Edmundson, P. Harold, “New methods in automatic extracting,” Journal of the ACM (JACM), vol. 16, 1969, pp. 264-285

Kupiec, Julian, P. Jan, F. Chen, “A trainable document summarizer,” Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, 1995

Conroy, M. John, P. Dianne, O'leary, “Text summarization via hidden markov models,” Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001

Rush, M. Alexander, C. Sumit, J. Weston, “A neural attention model for abstractive sentence summarization,” 2015, unpublished

Nallapati, Ramesh, et al, “Abstractive text summarization using sequence-to-sequence rnns and beyond,” 2016, unpublished

OpenAI, “GPT-4 technical report,” 2023, unpublished

J. Q. Zhang, et al, “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization,” International Conference on Machine Learning, PMLR, 2020

Hermann, M. Karl, et al, “Teaching machines to read and comprehend,” Advances in neural information processing systems, vol. 28, 2015