Novel Summarization

Calvin Lim
School of Computer Science Bina
Nusantara University Jakarta, Indonesia
calvin.lim001@binus.ac.id

Michael Vic Chow
School of Computer Science Bina
Nusantara University Jakarta, Indonesia
michael.chow@binus.ac.id

Kenichi Halim
School of Computer Science Bina
Nusantara University Jakarta, Indonesia
kenichi.halim@binus.ac.id

Abstract

This research explores the utilization of the Bidirectional Encoder Representations from Transformers (BERT) algorithm for novel text summarization. The study integrates two state-of-the-art Natural Language Processing (NLP) models, BERTSum and PEGASUS, aiming to leverage their unique strengths for enhanced text summarization performance. The models are trained and tested on diverse datasets, including the CNN-DailyMail dataset and a novel dataset derived from GPT-3.5. The performance of the models is evaluated using the ROUGE metrics, providing a quantitative measure of the models' summarization performance. The study suggests that the integration of BERTSum and PEGASUS could lead to improved efficiency and accuracy in digesting large amounts of text.

Keywords

Text Summarization, BERT, BERTSum, PEGASUS, Natural Language Processing, ROUGE Metrics, CNN-DailyMail Dataset, GPT-3.5, Machine Learning.

I. Introduction

In the face of an ever-growing digital landscape, the sheer volume of text data, coupled with escalating work demands and dwindling leisure time, poses a significant challenge for adult readers seeking to efficiently consume and comprehend necessary information. Fortunately, the rapid advancements in Artificial Intelligence (AI) offer a promising solution - text summarization. This technique condenses a text, preserving its crucial content, and has seen remarkable progress with the evolution of natural language processing. However, despite its ability to distill information, there is a persistent demand for more precise and efficient summarization methods.

This paper introduces a project that seeks to utilize the Bidirectional Encoder Representations from Transformers (BERT) algorithm for novel summarization. BERT has demonstrated exceptional capabilities in various natural language processing tasks, such as text classification and question answering. Yet, its potential in the realm of text summarization remains largely untapped. Our focus lies in fine-tuning the BERT model for text summarization, using a novel as a base, and assessing its performance. By capitalizing on the strengths of the BERT model, the system will comprehend the context and subtleties of a text, identifying the most crucial information to incorporate in the summary. This could prove beneficial for individuals who need to swiftly, accurately, and efficiently process large volumes of text, such as journalists, writers, and general readers.

In this study, we evaluate the performance of BERTSum and PEGASUS in novel summarization, aiming for enhanced efficiency and accuracy. The goal is to reduce the time a reader spends on a novel while preserving the essential information. Finally, we explore potential avenues for future research and conclude our findings.

II. Litterature Review

In the field of text summarization, the BERT model, introduced by Devlin et al. [1], has been a significant advancement. Abdel-Salam and Rafea [2] utilized this model in their performance study on extractive text summarization, providing a foundation for understanding its capabilities. This was further expanded by Liu [3], who delved into the fine-tuning of BERT for extractive summarization. The application of these models extends to specific language contexts, as demonstrated by Lucky and Suhartono [4] in their investigation of BERT for Indonesian abstractive text summarization.

The transformer model, proposed by Vaswani et al. [5], forms the backbone of many modern models, including BERT. The field has also seen the introduction of novel pre-training techniques, such as BART by Lewis et al. [6] and PEGASUS by Zhang et al. [7], which have expanded the scope of pre-training in natural language processing tasks. Tsvigun et al. [8] introduced the concept of active learning in text summarization, providing a potential avenue for model improvement.

Sharma et al, [9] provided a comprehensive review of these automatic text summarization methods, offering a broad context for the current study. Finally, the evaluation of these models is often conducted using the ROUGE package, introduced by Lin [10], providing a standard metric for performance evaluation. By understanding the frequency and importance of words in the text, we can calculate the similarities between documents and improve the performance of text summarization models.

III. Methodology

A. Models

The model employed in this research is a sophisticated integration of two state-of-the-art Natural Language Processing (NLP) models, BERTSum and PEGASUS. This composite model is designed to leverage the unique strengths of both constituent models, thereby enhancing its overall performance in text summarization tasks.

BERTSum, a variant of the Bidirectional Encoder Representations from Transformers (BERT), has demonstrated impressive performance in various NLP tasks, including text classification and question answering. It has been fine-tuned specifically for extractive summarization tasks, where the model identifies and ranks sentences based on their importance and retains the original wording. Conversely, PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization Sequence-to-sequence) is a model designed for abstractive summarization. It interprets information from the source text and generates summaries in its own words, creating a coherent, concise summary that retains the context and meaning of the original text while using different phrasing.

The integration of BERTSum and PEGASUS in our model aims to harness the extractive precision of BERTSum and the abstractive capabilities of PEGASUS. This combination allows the model to understand the context and nuances of a text, identify the most important information, and generate a summary that is both concise and representative of the original content. The model is trained and tested on a diverse range of datasets, including the CNN-DailyMail dataset and a novel dataset derived from GPT-3.5. This ensures a comprehensive evaluation of its summarization capabilities across different types of text.

The performance of the model is evaluated using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics. These metrics compare the summaries generated by the model with reference summaries and calculate the overlap of n-grams, word sequences, and word pairs, providing a quantitative measure of the model's summarization performance.

This hybrid model, with its integration of BERTSum and PEGASUS, represents a novel approach in the field of text summarization, promising improved efficiency and accuracy in digesting large amounts of text.

B. Dataset & Pre-processing

So because we wanted to make the models more accurate to accomplishing the task we set, we intended to further fine-tunes the models. Due to copyright and ethical issues we use CNN-DailyMail datasets. Fine-tuning of each model requires a different method to execute, by following each of the laid out instructions and training the model with each passing time, we expect it to increase the performance.

Standard preprocessing procedure for text based data used in NLP machine learning:

- Lower casing the corpus.
- Removing the punctuation.
- Removing the stopwords.
- Tokenizing the corpus.
- Stemming and Lemmatization.
- Word embeddings using CountVectorizer and TF-IDF.

Dataset for Testing In addition to the CNN-DailyMail datasets, we will also include a new dataset derived from GPT-3.5. This dataset will consist of carefully selected diverse novel clips that resembled real novels generated by the GPT-3.5 model. We will apply both extractive and abstractive summarization methods to these novel clips and compare the results with human summarization aid by GPT-3.5. This will allow us to evaluate the performance of our models in a novel context and assess their ability to generate coherent and meaningful summaries from diverse nature and complex narrative styles of novels.

C. Model Testing

To test the performance of our models, we will use both the CNN-DailyMail datasets and the new dataset derived from GPT-3.5. The models will be tasked with generating both extractive and abstractive summaries from these datasets. For the extractive summaries, the models will identify and rank sentences based on their importance and retain the original wording. For the abstractive summaries, the models will interpret information from the source text and generate summaries in their own words. This approach aims to create a coherent, concise summary that retains the context and meaning of the original text while using different phrasing.

D. Evaluation using ROUGE

The performance of the models will be evaluated using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics. ROUGE is a set of metrics used for evaluating automatic summarization and machine translation. It includes measures such as ROUGE-N (precision, recall, and F-score based on n-gram), ROUGE-L (longest common subsequence), and ROGUE-S (skip-bigram). These metrics compare the generated summaries with reference summaries (usually created by humans) and calculate the overlap of n-grams, word sequences, and word pairs, respectively.

IV. Results

In this section, we evaluated the performance of three summarization models, BERTSum, PEGASUS and BERTSum combined with PEGASUS (PEGASUS using BERTSum summay), using two different datasets: the CNN/DailyMail (CNN/DM) dataset, and a new dataset derived from GPT-3.5 prompts. The quality of the summaries generated by the models was assessed using ROUGE scores, which are a common metric for evaluating text summarization models.

The BERTSum model was tested on both the CNN/DM dataset and the new dataset derived from GPT-3.5 prompts. The results of these tests were presented in the Methodology chapter. The PEGASUS model was tested on all three datasets. On the CNN/DM dataset, the PEGASUS model achieved ROUGE scores that were competitive with those of the BERTSum model. When tested on the BERTSum dataset, the PEGASUS model's performance was slightly lower, but still within a reasonable range. The PEGASUS model was also tested on the new dataset derived from GPT-3.5 prompts. The results of these tests were particularly interesting, as they showed a significant improvement in ROUGE scores compared to the other datasets. This suggests that the GPT-3.5 prompts may be a valuable resource for improving the performance of text summarization models.

Finally, we tested a combination of the PEGASUS and BERTSum models on the GPT-3.5 dataset. The results of these tests were also promising, with ROUGE scores that were competitive with those of the individual models.

TABLE I. CNN/DailyMail

Models	***ROUGE-1***	***ROUGE-2***	***ROUGE-L***
BERTSum	0.53221	0.24775	0.48725

Models	***ROUGE-1***	***ROUGE-2***	***ROUGE-L***
PEGASUS (250)	0.40879	0.18319	0.38258
BERTSum + PEGASUS (250)	0.36081	0.15256	0.33685

TABLE II. GPT 3.5 PROMPTS

Models	***ROUGE-1***	***ROUGE-2***	***ROUGE-L***
BERTSum	0.64369	0.46940	0.60165
PEGASUS	0.43875	0.30175	0.41227
BERTSum + PEGASUS	0.42693	0.28228	0.39733

V. Conclusion

The findings we learned in this study provide a promising outlook on the capabilities of AI, specifically the BERTSum and PEGASUS model, in the field of text summarization. The performance of the models on various datasets, particularly the new dataset derived from GPT-3.5 prompts, was overall satisfactory, demonstrating the potential of these models in efficiently and accurately summarizing large volumes of text. The study also highlighted the value of fine-tuning pre-trained models for specific tasks, performance of the models on the GPT-3.5 dataset. The insights gained from this research underscore the vast potential of AI applications in the ever-growing digital content landscape. As we move forward, the possibilities for AI to revolutionize how we process and understand information seem limitless.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” May 2019. Available: https://arxiv.org/pdf/1810.04805.pdf

[2] S. Abdel-Salam and A. Rafea, “Performance Study on Extractive Text Summarization Using BERT Models,” Information, vol. 13, no. 2, p. 67, Jan. 2022, doi: 10.3390/info13020067.

[3] Y. Liu, “Fine-tune BERT for Extractive Summarization.” 2018, Available: https://arxiv.org/pdf/1903.10318.pdf

[4] H. Lucky and D. Suhartono, “Investigation of Pre-Trained Bidirectional Encoder Representations from Transformers Checkpoints for Indonesian Abstractive Text Summarization,” *Journal of Information and Communication Technology*, vol. 21, no. No.1, pp. 71–94, Nov. 2021, doi: https://doi.org/10.32890/jict2022.21.1.4.

[5] A. Vaswani *et al.*, “Attention Is All You Need,” *arXiv.org*, 2017. https://arxiv.org/abs/1706.03762

[6] M. Lewis *et al.*, “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,” Association for Computational Linguistics, 2020. Available: https://aclanthology.org/2020.acl-main.703.pdf

[7] J. Q. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,” ICML'20: Proceedings of the 37th International Conference on Machine Learning, vol. 119, p. 11328–11339, Jul. 2020, doi: 10.48550/arXiv.1912.08777.

[8] A. Tsvigun *et al.*, “Active Learning for Abstractive Text Summarization,” EMNLP-2022 Findings, Jan. 2023, doi: 10.48550/arXiv.2301.03252.

[9] G. Sharma and D. Sharma, “Automatic Text Summarization Methods: A Comprehensive Review,” *SN Computer Science*, vol. 4, no. 1, Oct. 2022, doi: https://doi.org/10.1007/s42979-022-01446-w.

[10] C.-Y. Lin, “ROUGE: a package for automatic evaluation of summaries,” Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), 2004.