Comparing Seq2Seq, BERT, and SimpleT5 Models for News Article Summarization: A Comprehensive Analysis

Himaja Kinthada, Siddhant Soam, Siddharth Pradhan, Tanmay Agarwal
New Jersey Institute of Technology

Abstract

This paper focuses on the intricate task of headline generation from news articles, aiming to distill crucial details into concise, informative and engaging summaries. Our analysis encompasses typical straightforward rule-based systems to cutting-edge deep learning methodologies, including transformer-based models like T5 and Bidirectional Encoder Representations from Transformers (Devlin et al., 2019). The core objective of our paper is to construct a Sequence-to-Sequence encoder-decoder model from the ground up and conduct a comparative analysis of these three specified models on the News Summary dataset. It is believed that an abstractive method of text summarization is a powerful way of summarizing texts, aiming to shed light on their performance and nuances. Through extensive analysis and comparison, we seek to identify the strengths and weaknesses of each model, ultimately contributing insights into the advancements and challenges within the realm of news article summarization using state-of-the-art deep learning techniques.

1 Introduction

There are two main types of text summarization: Extractive and Abstractive. The term "extractive summarization" describes the process of selecting key phrases or sentences from the source text. Method-based on extraction to construct the Summary, pick specific sentences or paragraphs from the original content. There are numerous algorithms that can automate this operation, but it can also be carried out manually. Similarity of sentences in the input material; sentence weight will choose phrases with more authority. “Abstractive summarization” is when the entire text/content is rewritten in a few words or lines, possibly including some new terms. New phrases created using abstractive techniques convey the original text’s meaning. While often more challenging than extraction-based summarization, this method can produce a summary that seems more authentic.

We investigate the applicability of BERT in text summarization within a comprehensive framework that includes both extractive and abstractive modeling approaches as proposed by (Liu and Lapata, 2019). The author worked on both extractive and abstractive summarization to showcase how BERT can be usefully applied in the text summarization. BERT combines both word and sentence representations in a single very large Transformer (Vaswani et al., 2023); it is pre-trained on vast amounts of text, with an unsupervised objective of masked language modeling and next-sentence prediction and can be fine-tuned with various task-specific objectives.

2 Background & Related Work

2.1 Pre-trained Models

Pretrained language models are usually used to enhance performance in language understanding tasks. Very recently, there have been attempts to apply pretrained models to various generation problems (Edunov et al., 2019); (Rothe et al., 2019). When fine-tuning for a specific task, unlike ELMo whose parameters are usually fixed, parameters in BERT are jointly fine-tuned with additional task specific parameters. The model used for the summarization is built through a transfer learning process on a T5 model. The choice of T5 is justified by the fact it provides a better generalization compared to BERT.

2.2 Abstractive Summarization

Abstractive summarization involves neural techniques that view the task as a sequence-to-sequence problem. In this framework, an encoder transforms a sequence of tokens in the source document x = [x1, ..., xn] into continuous representations z = [z1, ..., zn]. Subsequently, a decoder generates the target summary y = [y1, ..., ym] token by token, following an auto-regressive method. This approach models the conditional probability as p(y1, ..., ym|x1, ..., xn). Initial implementations by (Rush et al., 2015) and (Nallapati et al., 2016) introduced the neural encoder-decoder architecture to text summarization.

Further advancements in this model were made by (See et al., 2017), enhancing it with a pointer generator network (PTGEN) enabling word replication from the source text and a coverage mechanism (COV) to track summarized words. (Celikyilmaz et al., 2018)) proposed a system employing multiple agents (encoders) and a hierarchical attention mechanism for decoding, culminating in their Deep Communicating Agents (DCA) model trained end-to-end through reinforcement learning. Similarly, (Paulus et al., 2017) introduced a deep reinforced model (DRM) addressing coverage issues via an intra-attention mechanism where the decoder focuses on previously generated words. (Gehrmann et al., 2018) took a bottom-up approach (BOTTOM UP), where a content selector determines which phrases from the source document should constitute the summary, employing a copy mechanism solely for preselected phrases during decoding. Lastly, (Narayan et al., 2018) proposed an abstractive model specialized for extreme summarization, utilizing convolutional neural networks conditioned on topic distributions (TCONVS2S).

2.3 Sequence-to-Sequence Encoder-Decoder Model

Within our exploration of advanced Natural Language Processing (NLP) models for news article summarization, the Sequence-to-Sequence (Seq2Seq) model emerges as a foundational architecture. This model, rooted in the encoder-decoder paradigm, is designed to seamlessly handle sequential data. In a nutshell, the encoder component processes an input sequence, such as a 100-token representation of news text, capturing its contextual information. Subsequently, the decoder generates an output sequence, flexibly accommodating the dynamic nature of the target summary by accepting a sequence of integers representing words with varying lengths.

3 Experimental Setup

3.1 Dataset

This paper has a dataset which is sourced from multiple reputable news sources, amalgamating articles from Inshorts, Hindu, Indian Times, and Guardian, spanning between February and August 2017. This comprehensive dataset encompasses a total of 102,274 instances, consisting of headlines and summary text which we have used as the target text and source text respectively. There are other features available which comprise essential information such as news summaries, headlines, author names, publication dates, and complete news text. The dataset is bifurcated into two files: 'news_summary.csv' and 'news_summary_more.csv', with each file containing distinct columns.

The 'news_summary_more.csv' file consists of crucial details such as author names, publication dates, headlines, read_more links for headlines, as well as summary and complete news text columns. Conversely, the 'news_summary_more.csv' file contains column headlines and the summarized text. This dataset stands as a valuable resource for NLP practitioners, researchers, and enthusiasts due to its diverse range of articles, providing an opportunity for various text-related tasks such as summarization, headline generation, and text understanding, owing to its rich and multi-dimensional content.[1]

3.2 Data Preprocessing

This paper delves into comprehensive data preprocessing pipeline that significantly impacts the quality and effectiveness of natural language processing tasks. We have meticulously employed a series of techniques to prepare and refine the dataset for subsequent analysis and modeling. Our preprocessing methodology included several crucial steps: data concatenation, text stripping through regular expressions, SpaCy-based cleaning, and filtering data based on word count. Moreover, we selectively choose short text and summary pairs, recognizing the importance of concise and coherent data for efficient model training. To enhance the data representation, special tokens were strategically added. We have meticulously curated the dataset, dividing it into a training set of 70,000 instances, a validation set containing 10,000 instances, and a test set comprising 1,000 instances, allowing for comprehensive model evaluation.

Furthermore, we delved into the intricate aspects of tokenization, identifying rare words and calculating the vocabulary size to better understand the dataset's linguistic diversity. To ensure data integrity, they removed headlines containing only '<START>' and '<END>' tokens, streamlining the dataset for optimal quality. Additionally, we prepared label data by transforming words into integers, a critical step in facilitating model learning.

[1] https://github.com/SiddhantSoam/Text_Summarization

Equally important was the effort to standardize sequence lengths across the dataset, ensuring uniformity in the input data structure, which is crucial for model convergence and training efficiency. Overall, these rigorous preprocessing techniques significantly contributed to the dataset's refinement, laying a robust foundation for subsequent NLP model training and evaluation in the study.

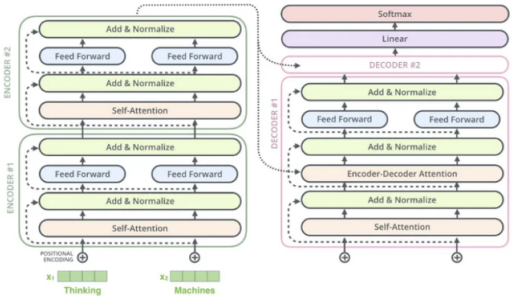

Figure 1: T5 Architecture

3.3 Simple T5 Model

The Transformer-based Text-to-Text Transformer (T5) model, known for its versatility in natural language processing tasks, is an encoder-decoder architecture designed by Google. With its text-to-text framework, T5 interprets a multitude of tasks as text-to-text problems, allowing for seamless adaptation to various tasks via a single unified architecture. The base version of T5 is equipped with 220 million parameters, enabling it to handle a wide array of language-based tasks with exceptional performance. Its ability to accommodate different input-output formats by manipulating text makes it a robust choice for various NLP applications Refer to (Raffel et al., 2019) for more details.

In this paper, we employed the T5 base model for specific tasks and conducted further training to fine-tune its performance. Utilizing a data frame structure with two columns labeled as "source_text" and "target_text," we fine-tuned the pre-trained T5 model by conducting additional training on a dataset containing 10,000 rows over five epochs. Throughout this fine-tuning process, the training loss achieved was 0.6599, while the validation loss measured 1.3475, reflecting the model's learning progress. To generate concise summaries, the addition of the "summarize: " keyword at the beginning of the text was necessary. For expedited execution time, we generated headlines for 100 articles, noting that the entire notebook execution duration using a GPU P100 took approximately 43 minutes, underscoring the computational efficiency of the model.

3.4 BERT Model

The BERT (Bidirectional Encoder Representations from Transformers) small model is a pivotal innovation in natural language processing, characterized by its bidirectional nature, which enables it to understand the context of words in a sentence by considering both preceding and succeeding words. With 110 million parameters distributed across 12 transformer layers, the model's architecture is defined by a hidden size of 768 and 12 attention heads. Pretrained on a substantial corpus, BERT has showcased exceptional capabilities in understanding language nuances and performing various NLP tasks effectively.

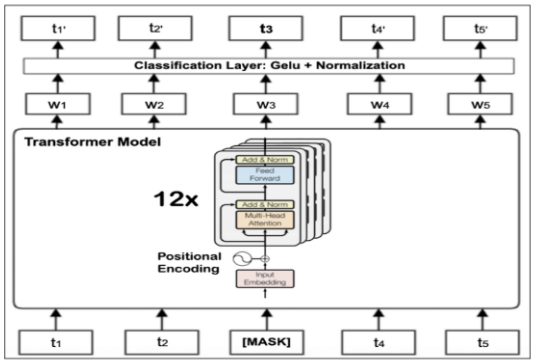

Figure 2: BERT architecture

In this paper, we leveraged the pretrained BERT small model from the paper (Liu and Lapata, 2019). This implementation involved fine-tuning the encoder twice, first on an extractive summarization task and subsequently on an abstractive summarization task, culminating in the nomenclature "BertSumExtAbs." The rationale behind this approach lies in the paper's assertion that this sequential fine-tuning significantly augments the quality of abstractive summarization. It's noteworthy that the notebook's execution time, encompassing environment setup, model downloading, and the summarization generation task on a GPU P100, amounted to approximately one hour, showcasing the model's efficiency in handling text summarization tasks within a reasonable time frame.

3.5 Seq2Seq Model

Our Seq2Seq model is configured with an approximate total of 18 million parameters, highlighting

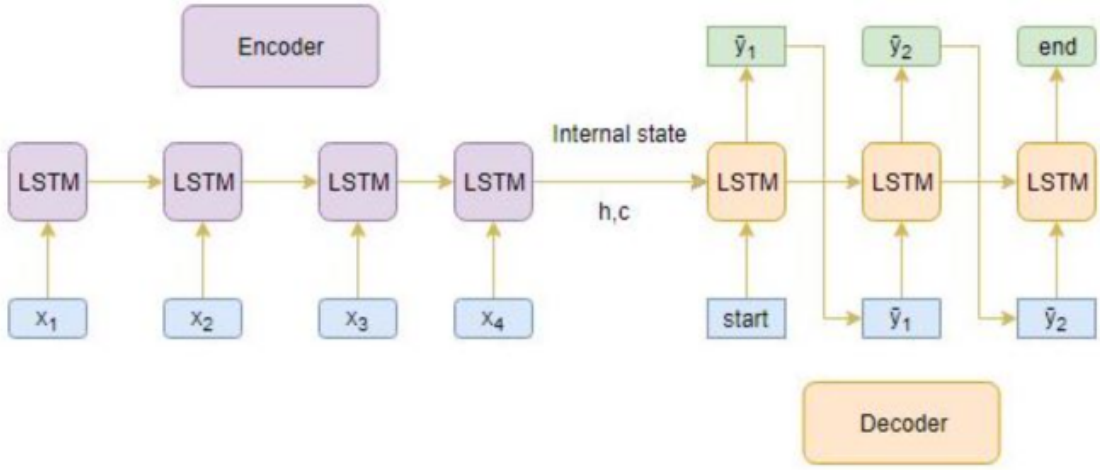

Figure 3: Seq2Seq architecture

its capacity for nuanced learning. In the training phase, we employed a robust dataset comprising 80,000 rows of news articles, subjecting the Seq2Seq model to an extensive 50-epoch training regimen with a batch size of 128. The recorded training loss of 3.20 and validation loss of 3.65 attest to the model's adaptability and effectiveness in generalizing to unseen data. Remarkably, the training process was executed efficiently, with a duration of approximately 3 hours on a GPU P100, underscoring the model's computational efficiency and practical utility in real-world applications.

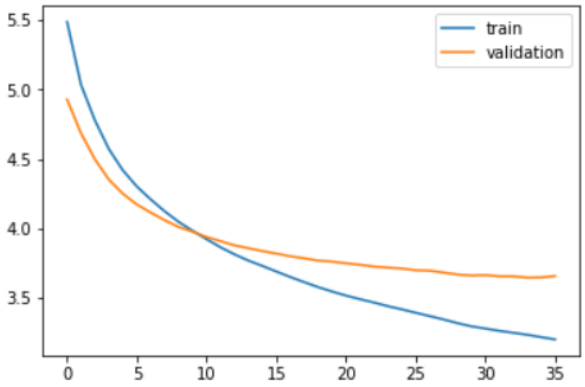

Figure 4: Training and Validation Loss for Seq2Seq Model

4 Results

We evaluated summarization quality automatically using ROUGE and BLEU scores. y automatically using ROUGE (Lin, 2004). We report unigram and bigram overlap (ROUGE-1 and ROUGE-2) as a means of assessing informativeness and the longest common sub sequence (ROUGE-L) and BLEU score as a means of assessing fluency.

In Table 1, we present a comparative analysis of the performances of T5, BERT, and Seq2Seq models across text summarization tasks, employing diverse evaluation metrics. Notably, our findings reveal substantial variations in the effectiveness of these models in generating succinct and informative summaries. The T5 model emerges as the top performer across all metrics, showcasing its superiority over the other models. Specifically, in terms of Rouge-1, T5 achieves a significantly higher score of 0.537 compared to BERT (0.213) and Seq2Seq (0.084). Similarly, for Rouge-2 and Rouge-L metrics, T5 consistently outperforms both BERT and Seq2Seq, demonstrating its capability to capture more comprehensive text abstractions and semantic correlations within the summaries. Moreover, T5 demonstrates notable proficiency in BLEU score, surpassing BERT and Seq2Seq with a score of 0.472, indicating its proficiency in generating summaries aligned with the reference texts. The results underscore the robustness and effectiveness of the T5 model in the context of text summarization tasks.

In contrast, BERT and Seq2Seq models exhibit relatively lower performance across all evaluation metrics, emphasizing the comparative limitations of these models in generating summaries that cap-

Metric	T5	BERT	Seq2Seq
Rouge-1	0.537	0.213	0.084
Rouge-2	0.139	0.043	0.007
Rouge-L	0.486	0.198	0.081
BLEU	0.472	0.091	0.100

Table 1: Summarizes our results on the processed news summary dataset which has "text" as input feature and "headlines" as target. The average ROUGE and BLEU scores obtained for T5, BERT and our Seq2Seq model are used for comparing the model performances.

ture the essence of the source texts. These findings suggest the promising potential of T5 in enhancing summarization capabilities, leveraging its versatile architecture and pretraining strategies, thereby advancing the state-of-the-art in generating coherent and concise summaries from lengthy textual inputs.

5 Limitations

The intricacies of the Seq2Seq architecture demand substantial computational resources, and the imposed limitations hindered our capacity to optimize the model's parameters adequately. Consequently, this restriction impacted the model's ability to capture intricate relationships within a few summaries of our data.

Furthermore, the adoption of the pre-trained BERT model, while promising in its ability to comprehend context and generate longer headlines, presented limitations in terms of evaluation metrics.

6 Conclusion

In this paper, we conducted a comprehensive comparison among three state-of-the-art models—BERT, T5, and Seq2Seq—focusing on their efficacy in the headline generation task. Our study emphasized the evaluation of these models using essential metrics such as ROUGE and BLEU scores, which are widely recognized in assessing the quality and coherence of generated summaries. Through meticulous experimentation and evaluation, we aimed to provide insights into the performance variations and strengths of these models in generating headlines.

Overall, our comparative study highlights the nuanced performance variations among BERT, T5, and Seq2Seq models in the context of headline generation. The extensive evaluation using established metrics provides valuable insights into their respective strengths and areas for improvement, contributing to the ongoing discourse on optimizing headline generation using state-of-the-art NLP models.

7 Acknowledgements

We acknowledge that this research forms a vital component of an academic endeavor. Our heartfelt gratitude extends to Dr. Du Mengnan, our esteemed professor at the New Jersey Institute of Technology (NJIT), whose invaluable guidance and unwavering support significantly contributed to the fruition of this research endeavor.

Furthermore, we extend our sincere appreciation to the Kaggle contributors for their valuable contributions in providing the dataset essential for our study. Their efforts in curating and sharing the dataset have been instrumental in the successful execution of this research. We are immensely grateful for their dedication and generosity.

References

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1662–1675, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4052–4059, Minneapolis, Minnesota. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of*

the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *CoRR*, abs/1705.04304.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2019. Leveraging pre-trained checkpoints for sequence generation tasks. *CoRR*, abs/1907.12461.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.