Domain-Specific Performance of Sequence-to-Sequence Models in Abstractive Text Summarization

Koushik Chanda
University of Florida

Rahul Mora
University of Florida

Sandeep Ragampudy
University of Florida

Rasagna Yarlagadda
University of Florida

Akhil Chatla
University of Florida

Abstract

Text summarization is a process that involves shortening long passages into concise summaries while preserving key information. With the advent of Sequence-to-Sequence (Seq2Seq) models, the ability to generate coherent and fluent summaries has significantly improved. However, controlling specific aspects of the output, such as word limit and the amount of paraphrasing, remains a challenge. In this paper, we evaluate Seq2Seq models with output control capabilities, enabling the generation of summaries that adhere to desired constraints. We used two models, Gemma V2 and Bart large CNN, and assessed their performances using ROUGE scores. We trained the model on two domains: News and Law, using relevant datasets from each field. We used Rogue N and Rogue L scores which are metrics used to evaluate the performance of automatic text summarization systems. Our analysis revealed that the Bart-Large-CNN model demonstrated a consistent performance across various domains, showing competitive results in both news and legal text summarization tasks.

1 Introduction

The massive growth of digital text data from sources like news articles, books, documents, and research papers has made manually reading and summarizing this content extremely difficult and impractical (El-Kassas et al., 2020). Text summarization has become a robust solution to address this challenge, enabling systems to generate concise summaries automatically, encapsulating the main concepts while avoiding redundancy(El-Kassas et al., 2020). Text summarization finds diverse applications like creating search snippets, condensing news headlines, facilitating legal abstracts, and summarizing complex biomedical texts (El-Kassas et al., 2020).

Within natural language processing (NLP), text summarization for large documents is a significant challenge (Syed et al., 2021). Text summarization provides a valuable alternative to manual summarization by generating accurate overviews while preserving essential information. Driven by neural networks and models like BERT, GPT, and BART, abstractive methods have recently made significant progress despite being more complex (Syed et al., 2021).

Sequence-to-Sequence (Seq2Seq) models, which are popularized for machine translation, have shown remarkable success (Yao and Koller, 2022) in text summarization due to their ability to capture the nuanced relationships between words in a sentence and generate fluent, coherent summaries.

Despite their success, Seq2Seq models face challenges in controlling specific aspects (Zhang et al., 2019) of the output, such as the word limit and the degree of paraphrasing. The ability to control these factors is crucial for various applications, including generating concise summaries for mobile displays, creating paraphrased content for plagiarism avoidance, and simplifying texts for readability enhancement (Zhang et al., 2019). Traditional Seq2Seq models lack explicit mechanisms to regulate these output characteristics, often leading to summaries that are either too verbose or too similar to the original text (Zhang et al., 2019).

The need for effective automated text summarization methods has become increasingly crucial. The internet has overwhelming volume of lengthy documents, making manual summarization efforts time-consuming (Karjule et al., 2023). By using abstractive text summarization techniques and large pre-trained language models, our approach generates concise yet informative summaries that capture the salient points from complex legal documents and news articles. By leveraging domain-specific data and transfer learning strategies, our model effectively extracts critical information while preserving the nuances and terminologies inherent to these specialized domains. This paper details the architecture, training process, and evaluation of our Seq2Seq summarization model on news and legal domains.

2 Related Work

The rise of generative models such as Bart-Large-CNN, Gemma, GPT significantly expanded the possibilities in abstractive summarization. This section discusses the research done in the strategic fine-tuning of LLMs and the broader implications for domain-specific text summarization.

2.1 Fine-Tuning LLMs for Domain Specific Applications:

The paper "Fine-tuning and Utilization Methods of Domain-specific LLMs" (Jeong, 2024) by Cheonsu Jeong provides a comprehensive analysis of the methodologies for fine-tuning Large Language Models (LLMs) for specific domains, with a focus on the financial sector. It details the process of dataset selection, model pre-training, and the importance of constructing domain-specific vocabularies, while also addressing the imperative of security and regulatory compliance. Practical implementations are explored through case studies in finance, such as stock prediction and sentiment analysis, demonstrating the potential and challenges of LLMs in real-world applications. The study not only highlights the advantages of fine-tuned LLMs in improving decision-making and operational efficiency in finance but acknowledges that the dataset used for fine-tuning potentially introduce bias or limit the generalizability of the model.

2.2 Text summarization using Various LLMs

The paper "Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models" by Lochan Basyal and Mihir Sanghvi (Basyal and Sanghvi, 2023) investigates the efficacy of various LLMs in the task of text summarization. Utilizing the CNN Daily Mail and XSum datasets, the study evaluates summaries using BLEU, ROUGE, and BERT Scores, with the text-davinci-003 model outperforming the others. The study reveals the potential of LLMs to generate concise summaries while retaining critical information. However, limitations include a focus on only two datasets which might not reflect the models' performance across more diverse text types and the computational resources required for such large models.

2.3 Enhancing Abstractive Text Summarization Through a Hybrid LSTM-CNN Framework

In the paper "Abstractive text summarization using LSTM-CNN based deep learning"(Song et al., 2019) a framework called ATSDL is discussed for abstractive text summarization (ATS) using a combination of Long Short-Term Memory (LSTM) (Yu et al., 2019) networks and Convolutional Neural Networks (CNN) (Xia et al., 2020). ATS is challenging due to the need to create coherent summaries that not only extract but also paraphrase content. The ATSDL framework seeks to improve upon previous models by extracting phrases and reassembling them into concise summaries. This method not only preserves semantic accuracy but also enhances syntactic correctness, distinguishing it from traditional extractive and generative approaches. The findings after experimentation on datasets such as CNN and DailyMail states that ATSDL(Song et al., 2019) outperforms existing models in generating summaries that are syntactically and semantically better.

3 Methodology

This section outlines the systematic approach employed to fine-tune pre-trained LLMs for domain-specific datasets and evaluate them for their ability to summarize texts. The overall flow of our research is as shown in Figure 1.

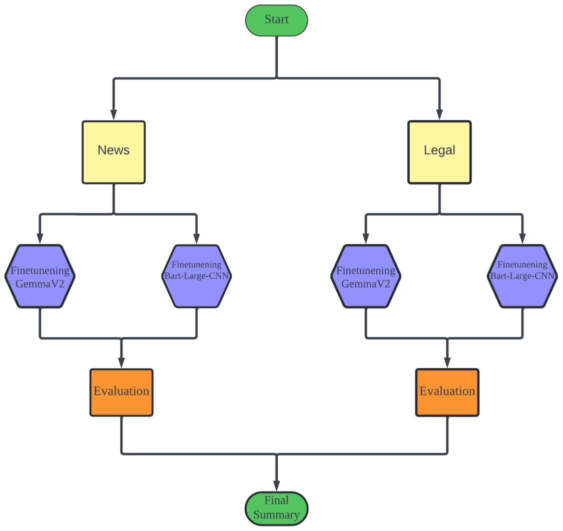

Figure 1: Workflow

3.1 Dataset

In our research, we utilize two distinct datasets to investigate the performance of our models across different domains. For the news and media domain, we employed the CNN/Daily Mail dataset (Penugonda) which we dgot it from Kaggle, which comprises over 300,000 news articles. This dataset is widely used in natural language processing for tasks such as summarization and text classification, providing a substantial corpus for evaluating model efficacy in real-world scenarios. The data is of the following format (Penugonda) :

{ "id": "id of the article",
"article": "Whole article",
"highlights": "headline of the article." }

For the legal domain, we utilized the BillSum dataset (Trivedi), which contains approximately 20,000 U.S. Congressional bills along with their summaries which we downloaded it from Kaggle. The BillSum dataset offers a unique challenge due to the specialized language and complex structure of legislative documents. It serves as an excellent resource for testing the capabilities of our models in processing and summarizing legal texts, highlighting their adaptability and precision in domain-specific applications. The format of the dataset is as follows (Kornilova and Eidelman, 2019):

{ "summary" : "some summary",
"text" : "some text.",
"title": "An act to amend Section xxx." }

We have used the the "text" variable as the source label and the "summary" variable as the target label.

3.2 Models

After conducting a literature review, we've identified a gap in existing surveys concerning the evaluation of the summarization capabilities of recent models such as GemmaV2 and Bart-Large-CNN across both general domains like news and technical domains like law. As a result, we've decided to focus our research on evaluating these two models.

- **Gemma V2:** The Gemma V2 model (Team et al., 2024) is a sophisticated transformer-based architecture that is part of a series ranging from 2 to 7 billion parameters, enabling it to deeply understand and process language. It is designed to effectively manage long-context information by incorporating modifications to the attention mechanism, such as local attention, sliding windows, and memory-compressed attention. These enhancements help the model handle lengthy inputs efficiently, without drastically increasing computational requirements.

Parameters	7B
d_model	3072
Layers	28
Feedforward hidden dims	49152
Num heads	16
Num KV heads	16
Head size	256
Vocab size	256128

Table 1: Model parameters for the GemmaV2

- **Bart-Large-CNN:** The BART-Large-CNN model, an abbreviation for Bidirectional and Auto-Regressive Transformers, serves as a denoising autoencoder used to pretrain sequence-to-sequence models. It utilizes a conventional Transformer-based architecture for neural machine translation (Varshney et al., 2023), which combines the bidirectional encoder from BERT with the left-to-right decoder from GPT. It is pretrained by deliberately adding noise to text and then learning to reconstruct the initial content, showing significant effectiveness, particularly in tasks related to generating text. (Lewis et al., 2019).

3.3 Experiments

The Aim of our experiment is to fine tune the GemmaV2 and Bart-Large-CNN LLMs on the CNN and Billsum datasets to evaluate their performance of abstractive summarization in the News and Law domains.

For that, we have fine tuned these models by using LoRa(Low-Rank Adaptation) (Hu et al., 2021). Low-Rank Adaptation (LoRA) is a training technique that adapts pre-trained language models by introducing and training a low-rank decomposition of the model's weight matrices, thus updating a small subset of parameters. This approach maintains the bulk of the pre-trained weights fixed, enabling efficient fine-tuning and specialization of the model on a target task without the computational burden of full model retraining.

Our methodology, as depicted in the figure 2, evaluates two language models, GemmaV2 and Bart-Large-CNN, which are finetuned and assessed

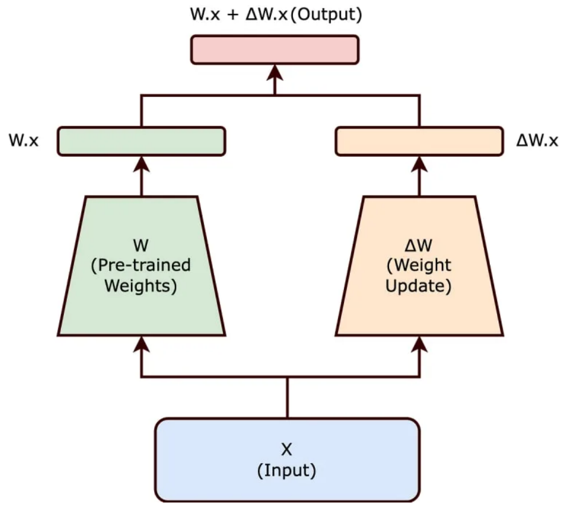

Figure 2: Working of LoRA

for their summarization capabilities in the news and legal domains. The process involves data collection,training the model by adjusting model's hyperparameters and finally evaluating their summarising capabolities using Rogue scores. Ultimately, the experiment aims to produce a conclusion that reflects the comparative effectiveness of these models in Law and News domains.

In the initial phase of our experiments, we aimed to establish baselines for our Seq2Seq models, namely Gemma V2 and Bart-Large-CNN. For Both the models, we initially assigned the learning rate as $5e^{-}5$ and dynamically modified it based on the model's results on the validation set and finalized the learning rate of Gemma to $2e^{-}4$ and Bart-Large-CNN to $3e^{-}5$.

Both models were trained for an initial number of epochs set to 30, with an early stopping mechanism in place to halt training if there was no improvement in the validation loss for two straight epochs, thus mitigating overfitting. We employed the Adam optimizer for its effectiveness in managing sparse gradients and handling varying parameter scales across the model.

To prevent the models from learning invalid patterns, the training data was shuffled at the beginning of each epoch. Additionally, gradient clipping was implemented to address the issue of exploding gradients in deep neural networks, ensuring stability during training.

Following each training epoch, both Gemma V2 and Bart-Large-CNN underwent evaluation on separate validation sets. This iterative process allowed for dynamic adjustment of training hyperparameters based on performance metrics such as validation loss and ROUGE scores, ensuring continual optimization of model performance.

Table 2: Hyperparameters for training using GemmaV2

Hyperparameter	**Value**
max_seq_length	512
per_device_train_batch_size	1
gradient_accumulation_steps	4
warmup_steps	2
max_steps	30
learning_rate	$2e^{-}4$
fp16	True
logging_steps	1

Table 3: Hyperparameters for Bart-large-CNN Model

Hyperparameter	**Value**
Batch Size	4
Number of Training Epochs	30
Learning Rate	$3e^{-}5$
Weight Decay	0.01
Max Source Length	512
Max Target Length	64

3.4 Evaluation Metrics

To evaluate the performance of our models we employed Rouge scores as our primary evaluation metric. These scores are essential for assessing the quality of text summaries (Lin, 2004) by comparing them to human-generated reference summaries. We specifically used ROUGE-N, which measures the overlap of N-grams between the machine-generated summary and the reference (Marek, 2023), aiding in evaluating the precision of content words within the summary. Additionally, ROUGE-L, which focuses on the longest common subsequence, was used to gauge the fluency and structural similarity of the generated text, reflecting the natural flow and readability of the summary.

4 Results and Analysis

4.1 Results

Our study evaluated the summarization abilities of two Seq2Seq models, GemmaV2 and Bart-CNN, across two distinct domains: news and legal texts. We utilized the Rouge Score metric to assess the performance of these models after fine-tuning them with domain-specific datasets.

News Domain (CNN/DailyNews Dataset): Table 4.

For the news domain, the GemmaV2 model achieved Rouge-1, Rouge-2, and Rouge-L (Ng and Abrecht, 2015) scores of 0.64, 0.53, and 0.64, respectively. On the other hand, the Bart-CNN model's performance showed Rouge-1, Rouge-2, and Rouge-L scores of 0.60, 0.42, and 0.60, respectively. This indicates that both models performed comparably in capturing the overall gist and critical information of the news articles, although GemmaV2 showed slightly higher precision in capturing the finer details compared to Bart-CNN.

Table 4: Rouge Scores for CNN/DailyMail dataset (News Domain)

Model	Rouge-1	Rouge-2	Rouge-L
GemmaV2	0.64	0.53	0.64
Bart-CNN	0.60	0.42	0.60

Legal Domain (BillSum Dataset): Table 5.

In the legal domain, the performance disparity between the models was more pronounced. The GemmaV2 model's scores were significantly lower than in the news domain, with Rouge-1, Rouge-2, and Rouge-L scores of 0.33, 0.10, and 0.30 respectively. Bart-CNN performed more robustly with scores of 0.63, 0.45, and 0.51 for Rouge-1, Rouge-2, and Rouge-L, respectively. This suggests that the Bart-CNN model is more adept at handling the complex sentence structures and specialized vocabulary found in legal documents.

Table 5: Rouge Scores for BillSum dataset (Legal Domain)

Model	Rouge-1	Rouge-2	Rouge-L
GemmaV2	0.33	0.10	0.30
Bart-CNN	0.63	0.45	0.51

4.2 Analysis:

When correlating the performance of the models across different domains, several trends emerged. Notably, the Bart-Large-CNN model exhibited a more consistent performance across domains, performing competitively in both news and legal text summarization tasks. In contrast, the GemmaV2 model showed domain-specific variations in performance, excelling in news summarization but encountering challenges in legal text summarization.

These findings underscore the importance of considering domain-specific characteristics when evaluating NLP models and highlight the need for targeted fine-tuning to enhance performance across diverse domains.

5 Conclusion

In conclusion, our research addresses the evaluation of text summarization models by examining Seq2Seq models with output control capabilities across various domains. Through an analysis of GemmaV2 and Bart-Large-CNN models trained on news and legal datasets, we have provided insights into their performance and effectiveness in generating coherent and informative summaries.

Our findings reveal that while both models demonstrate promising results, there are notable differences in their performance across domains. The Bart-Large-CNN model exhibits consistent performance in both news and legal text summarization tasks, underscoring its versatility and effectiveness in capturing key information from diverse sources. In contrast, the GemmaV2 model shows domain-specific variations, excelling in news summarization but facing challenges in processing legal texts.

6 Limitations and Future Work

This study highlights several limitations that require careful consideration. Primarily, the efficacy of fine-tuning large language models (LLMs) heavily relies on the availability of large, high-quality, domain-specific datasets. In fields where such data are scarce or fragmented, model performance and generalizability may be significantly hindered.

Relying solely on Rouge scores might not fully capture the qualitative aspects of summarization effectiveness across different domains.

Future scope can include exploring the effectiveness of cross-lingual summarization, enhancing techniques for summarizing low-resource languages, optimizing models for real-time performance, assessing robustness against adversarial attacks, and incorporating knowledge graphs to improve context understanding. Aditionally, future work could focus on refining the adaptation processes of sequence-to-sequence models to enhance their effectiveness across varied domains. This would involve developing more sophisticated domain-specific evaluation techniques. Additionally, it would be valuable to explore the integration of external knowledge bases or the incorporation of attention mechanisms tailored to specific types of data. Such improvements could help in overcoming the discrepancies in performance noted between different domains, as observed with GemmaV2 and Bart-Large-CNN, thereby achieving a more uniform standard of summarization quality across a broader range of text types.

7 Acknowledgements

We express our deepest gratitude to our mentors and professors who have made this research possible through their guidance and invaluable support. Special thanks to our peer mentor, Justin Ho, whose expertise and insights have greatly enhanced our understanding and execution of this project. We are also profoundly thankful to Professors Dr. Sangpil Youm and Dr. Bonnie Dorr from the Computer Science department at University of Florida, whose resources and relentless dedication have been pivotal in our research.

References

Lochan Basyal and Mihir Sanghvi. 2023. Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models. *arXiv preprint arXiv:2310.10449*.

Wafaa El-Kassas, Cherif Salama, Ahmed Rafea, and Hoda Mohamed. 2020. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Cheonsu Jeong. 2024. Fine-tuning and utilization methods of domain-specific llms.

Vivek Karjule, Jayesh Dange, Sneha Thange, Janhavi Sase, and Prof Kokate. 2023. A survey on text summarization techniques. *Journal of Natural Language Processing*.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. pages 48–56.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. pages 74–81.

Petr Marek. 2023. Dialogový management pro konverzační umělou inteligenci. *PQDT - Global*, page 202. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-12-22.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. *arXiv preprint arXiv:1508.06034*.

Gowri Shankar Penugonda. CNN-DailyMail News Text Summarization — kaggle.com. [Accessed 25-04-2024].

Shengli Song, Haitao Huang, and Tongxiao Ruan. 2019. Abstractive text summarization using lstm-cnn based deep learning. *Multimedia Tools and Applications*, 78.

Ayesha Ayub Syed, Ford Lumban Gaol, and Tokuro Matsuo. 2021. A survey of the state-of-the-art models in neural abstractive text summarization. *IEEE Access*, 9:13248–13265.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

Pawan Trivedi. BillSum Processed — kaggle.com. [Accessed 25-04-2024].

Deeksha Varshney, Aizan Zafar, Niranshu Kumar Behera, and Asif Ekbal. 2023. Knowledge graph assisted end-to-end medical dialog generation. *Artificial Intelligence in Medicine*, 139:102535.

Kun Xia, Jianguang Huang, and Hanyu Wang. 2020. Lstm-cnn architecture for human activity recognition. *IEEE Access*, 8:56855–56866.

Yuekun Yao and Alexander Koller. 2022. Structural generalization is hard for sequence-to-sequence models.

Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation*, 31(7):1235–1270.

Yong Zhang, Dan Li, Yuheng Wang, Yang Fang, and Weidong Xiao. 2019. Abstract text summarization with a convolutional seq2seq model. *Applied Sciences*, 9(8).