Abstractive Summarization Using a Pretrained T5 Model

Viktoria Tsiokou
University of the Basque Country
San Sebastián
Tutor: Eneko Agirre
vtsiokou001@ikasle.ehu.eus

Abstract

In today's world, large language models are increasingly being deployed for a variety of tasks, including summarizing longer texts. This report describes an experiment using a pretrained T5 model for abstractive summarization. The experiment also involved a linguistic analysis to identify the similarities and differences between the reference and generated summaries.

1 Introduction

In the age of information overload, the ability to efficiently extract meaning from vast amounts of text is crucial. Automatic text summarization has emerged as a powerful tool to address this challenge, condensing information into concise and informative summaries. The rise of large language models (LLMs) like ChatGPT and the T5 model have revolutionized this field, enabling the generation of abstractive summaries that capture the essence of the source text while going beyond mere extraction of key sentences. This project explores the potential of the T5 model for abstractive summarization, aiming to leverage its pre-trained capabilities to generate high-quality summaries that meet the information needs of users in today's information-rich environment.

2 Related Works

In the domain of text summarization with deep learning, a pivotal contribution comes from the exhaustive review conducted by Guanghua Wang and Weili Wu(Wang and Wu, 2023), which meticulously explores the multifaceted landscape of text summarization techniques. This comprehensive survey not only elucidates the historical evolution of text summarization but also delves into the intricate methodologies powered by deep learning algorithms, providing a holistic view of the field's advancements and challenges.

Mingye Wang et al.'s(Wang et al., 2023) research introduces a novel T5-based model tailored specifically for abstractive summarization tasks. Their approach stands out for its innovative integration of semi-supervised learning techniques augmented with consistency loss functions. By leveraging the T5 model's capabilities within a semi-supervised framework, their work aims to enhance the quality and coherence of abstractive summaries generated, thereby pushing the boundaries of summarization performance.

Abdul Ghafoor Etemad et al.'s (Etemad et al., 2021) study focuses on fine-tuning the T5 model to optimize its performance for abstractive summarization tasks. By tailoring fine-tuning strategies specifically for abstractive summarization requirements, their work aims to improve the model's ability to generate concise and contextually rich summaries. This research underscores the importance of customizing pre-trained models like T5 to suit the nuances of abstractive summarization, thereby enhancing the overall efficacy and relevance of generated summaries.

Upon a thorough examination of Wang and Wu's comprehensive review, it became apparent that amidst the diverse array of text summarization approaches, abstractive summarization emerged as a particularly captivating avenue for exploration. The allure of creatively rephrasing and synthesizing textual content to distill its essence resonated strongly, prompting a focused exploration into the intricacies and advancements within the realm of abstractive summarization.These seminal works collectively contribute to enriching the discourse on abstractive summarization methodologies, showcasing how cutting-edge deep learning models like T5 can be harnessed to elevate the quality and efficiency of text summarization processes.

3 Dataset

The foundation of this work was the CNN/DailyMail dataset[1], a widely used benchmark for training and evaluating abstractive summarization models. This dataset, accessible

[1] https://huggingface.co/datasets/ccdv/cnn_dailymail

through the `load_dataset` function from the Hugging Face library (version 3.0.0 in this case), provides a rich collection of text pairs suitable for abstractive summarization tasks. Each data point consists of two elements:

1. **Source Article:** This is the main body of text, a news article from either CNN or Daily Mail websites. The dataset includes the full content of the article, allowing the model to capture the context and key points for summarization.

2. **Abstractive Summary:** This is a human-written concise summary of the corresponding source article. These summaries are abstractive in nature, meaning they go beyond simply extracting key sentences and instead present a new, condensed version of the article that captures the essential information while potentially using different phrasings and vocabulary.

3.1 Preprocessing

The raw text data underwent a meticulous preprocessing pipeline to prepare it for training the T5 model. This pipeline involved several key steps.

First, the source and target texts were segmented into individual words, a process known as tokenization. This transformation, achieved using the `word_tokenize` function from the NLTK library[2], breaks down sentences into a sequence of tokens that the model can understand and process.

Next, part-of-speech (POS) tags were assigned to each token using NLTK's `pos_tag` function. These tags, like noun, verb, adjective, and adverb, provide valuable grammatical context for language understanding. By understanding the grammatical role of each word, the model can learn more nuanced relationships between words.

Following POS tagging, lemmatization was performed using the NLTK WordNetLemmatizer. This process simplifies words to their base forms (lemmas), reducing variations caused by inflections. Leveraging the POS tags during lemmatization ensures accuracy, as word meanings can change depending on their grammatical role.

Once the text was tokenized, lemmatized, and enriched with part-of-speech information, it was fed into the T5Tokenizer, a pre-trained model based on the T5-small architecture. This tokenizer converted the textual data into numerical token IDs, a format suitable for the T5 model[3] Truncation was applied during this step to limit the maximum sequence length and improve computational efficiency.

Right after, the token IDs were transformed into PyTorch tensors, the data format required for training the T5 model. Additionally, attention masks were created to identify padding tokens within the sequences. These masks ensure the model focuses only on relevant parts of the input during training, ignoring any padding elements used for sequence length consistency. The token IDs for the target text were further converted into "labels," which serve as the ground truth information for supervised learning, allowing the model to compare its generated summaries with the human-written ones and learn from the differences.

By following these preprocessing steps, the text data was transformed into a structured and informative representation that the T5 model could effectively utilize for learning abstractive summarization.

Finally, I had to downsize the initial dataset from 287113 articles to a more manageable subset of 5000 randomly selected articles for training, in addition to selecting 1000 articles for validation and 200 articles for testing. This downsampling approach allowed for a focused and efficient training process, while ensuring diversity and comprehensiveness within the dataset. The preprocessed data was then prepared for model training by converting each selected article and its highlights into sequences as we see in Table 1, further optimizing the dataset for effective deep learning model fine-tuning.

Dataset	Size
Training	77829
Validation	13223
Test	2570

Table 1: Dataset Sizes (in sequences)

4 Training the Model

The pre-trained model employed in this work leverages the Transformer-based architecture introduced by Vaswani et al. (Vaswani et al., 2017), specifically utilizing the T5-small variant (Raffel et al.,

[2] https://www.nltk.org/

[3] https://github.com/google-research/text-to-text-transfer-transformer?tab=readme-ov-file#released-model-checkpoints

2020), considering computational limitations. This architecture has demonstrated exceptional capabilities in various text-to-text tasks, including summarization. The following sections will delve deeper into the inner workings of the T5 architecture and how it is specifically adapted for the task of abstractive summarization.

4.1 T5 and fine-tuning

My project made use of a publicly available implementation for fine-tuning a pre-trained T5 model on the abstractive summarization task. The code, originally developed by Pratiksahu631 and available on GitHub [4], provides a well-structured framework for training the T5 model on custom datasets.

At its core, the T5 model employs an encoder-decoder architecture(Zhang et al., 2020a). The encoder processes the source text (article) and generates a contextual representation, capturing the key information and relationships within the document. This encoded representation is then fed into the decoder, which utilizes an attention mechanism to selectively focus on relevant parts of the encoded information while generating the abstractive summary. The attention mechanism allows the decoder to dynamically attend to specific sections of the source text as it builds the summary, ensuring a coherent and informative output.

To adapt the pre-trained T5 model for abstractive summarization, the fine-tuning process utilizes the provided code. This code first loads the pre-trained T5 model (T5ForConditionalGeneration) and its corresponding tokenizer (T5Tokenizer) from the Hugging Face Transformers library. The training, validation, and test datasets are then loaded using a custom TextDataset class, which tokenizes the text data and handles batching during training.

The code defines a DataCollatorForLanguageModeling specifically tailored for this summarization task. By setting the mlm argument to False, this data collator avoids masked language modeling, which would be counterproductive for summarization. Instead, it focuses on preparing batches for the model to learn the task of directly generating summaries from the provided source text.

Finally, the code defines training arguments specifying hyperparameters like epochs, batch sizes, and evaluation strategies. I ran the code for 3 epochs, considering the GPU limitation on my Google Colab Account. A Trainer object from the Transformers library is then instantiated to manage the training process. This object utilizes the loaded model, tokenizer, training arguments, data collator, and datasets to perform the fine-tuning process. The fine-tuned model and tokenizer are ultimately saved for further evaluation and deployment.

Looking at the training and validation loss scores, we see that the model demonstrates effective learning, as shown by the reduction in training loss across epochs in Table 2, though a slight uptick in validation loss by the third epoch calls for vigilant oversight to avert potential overfitting.

Epoch	Training Loss	Validation Loss
1	0.107700	0.003870
2	0.008600	0.002835
3	0.006000	0.006038

Table 2: Training and Validation Losses Over Epochs

5 Evaluation

This work employs the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric for model evaluation. ROUGE has established itself as a standard metric within the field of Natural Language Processing (NLP) for tasks involving text summarization (Lin, 2004). Its widespread adoption facilitates objective comparisons between various summarization models.

More specifically, I focused on ROUGE-1, ROUGE-2 and ROUGE-L metrics. ROUGE-1 evaluates word-level overlap, assessing how well the model captures the core vocabulary of the source text. ROUGE-2 delves deeper, examining bigram co-occurrence to gauge the model's ability to represent phrasal relationships. Finally, ROUGE-L goes beyond n-grams, identifying the longest matching sequence of words (not necessarily consecutive) to capture overall semantic similarity and gist. This combination provides a comprehensive evaluation of the model's performance in capturing essential information and mimicking human-written summaries.

Metric	Recall	Precision	F1-Score
ROUGE-1	0.32567	0.37368	0.33848
ROUGE-2	0.14041	0.15940	0.14424
ROUGE-L	0.30703	0.35280	0.31945

Table 3: Summary Evaluation Metrics

[4] https://github.com/Pratiksahu631/Abstractive_Text_Summarization/blob/main/Final_year_Project.ipynb

While ROUGE is a common metric for summarization tasks, it has limitations. Firstly, ROUGE scores can be skewed by the quality of reference summaries. Secondly, its focus on n-gram overlap may reward summaries mimicking surface-level features without capturing deeper semantics. Finally, it doesn't explicitly evaluate informativeness. Future work could explore incorporating complementary metrics like METEOR (Lavie and Agarwal, 2007) (which considers word order and synonyms) and BERTSCORE (Zhang et al., 2020b) (which leverages pre-trained models for semantic similarity assessment) to provide a more comprehensive evaluation of the model's performance.

6 Linguistic Analysis: Going Beyond ROUGE scores

While ROUGE scores provide a quantitative measure of similarity between generated summaries and reference summaries, they don't offer insights into the nuanced linguistic aspects of these summaries. To gain a deeper understanding of the strengths and weaknesses of the T5 model in abstractive summarization, this section delves into a qualitative linguistic analysis. This analysis compares and contrasts three example reference summaries with the summaries generated by the T5 model (found in the Appendix), focusing on four key aspects:

1. **Content Coverage**: The analysis of content coverage demonstrates that, while generated summaries are adept at capturing the overarching themes of source materials, they frequently miss finer details and contextual nuances essential for comprehensive understanding. This tendency is evident across several case studies. For instance, the omission of Leonard H. Thomas's diaries in Example 1 removes a layer of historical depth, while Example summary 2 fails to convey the significance of George H.W. Bush's presence at a tennis event, underscoring the importance of narrative details. Additionally, Example 3 neglects the groundbreaking legal implications of serving divorce papers via Facebook, highlighting a gap in capturing novel developments. These specific instances underscore a broader trend where summaries, despite their effectiveness in distilling primary themes, often fall short in integrating critical details and contextual background. This balance between overarching content capture and the integration of nuanced details is pivotal in enriching narratives and fully leveraging the descriptive power of summaries.

2. **Coherence**: This facet examines the logical flow and internal consistency of the information presented in the summary. It evaluates whether the generated summary presents a clear and well-organized narrative, or if it suffers from disjointedness or lack of clear connections between ideas. In reviewing both generated and reference summaries across various examples, it's evident that while both types generally uphold narrative flow, reference summaries often achieve a more holistic and interconnected presentation. This distinction is notably seen through their ability to seamlessly weave together diverse story elements, a quality vividly demonstrated in the first two case studies. This pattern suggests a broader observation: summaries that adeptly link different parts of the narrative not only enhance coherence but also enrich the reader's comprehension, illustrating the importance of narrative flow in the effectiveness of summary generation.

3. **Conciseness**: This analysis evaluates the ability of the summary to convey essential information in a succinct and focused manner. While brevity is desirable, it shouldn't come at the expense of omitting crucial details or context that enriches the understanding of the source text. The examples demonstrate how generated summaries achieve conciseness but may sacrifice valuable details that enhance the story.

4. **Tone**: The analysis of tone reveals that both generated and reference summaries generally adhere to a neutral and objective style, aligning with the expectation for summarization tasks. However, the reference summaries occasionally imbue a slight emotive quality or stylistic flair that enhances the reader's engagement, as seen in Example 3's portrayal of social media's evolving role in legal service. Such nuances in tone, while subtle, can significantly affect the summary's impact, demonstrating an area where automated summaries could further adapt to mirror the source text's emotional undertones more closely.

By employing this multifaceted linguistic analysis alongside ROUGE scores, we gain a more comprehensive picture of the T5 model's performance. This allows for the identification of areas where the model excels, such as concise content delivery, and areas where improvement is needed, such as capturing nuanced information and maintaining a balanced level of detail.

6.1 ChatGPT and T5 Model: Comparing Results

As part of my linguistic analysis, I wanted to compare the generated summaries to those of ChatGPT4, and mark down their differences. After giving ChatGPT one of the example articles I used for my analysis, I prompted it to summarize it in no more than 250 tokens. Interestingly, the ChatGPT4 summary exhibited a more holistic view of the original article compared to the T5 model's output. This suggests a potential difference in their summarization strategies. While the T5 model's summary may have prioritized conciseness, ChatGPT4's approach appeared to favor a more comprehensive representation, incorporating various details through the use of complex sentence structures. In terms of informative value, the ChatGPT4 summary presented a more well-rounded picture of the source article. The result summaries can be seen in Table 4.

7 Conclusion

This report has briefly described the steps taken to work with a pretrained T5 model for the task of abstractive summarization using the CNN/Dailynews dataset. The experiment also explored preprocessing techniques, evaluated the model's performance using the ROUGE metric, and analyzed the linguistic differences between the reference and generated summaries. Finally, this research suggests two potential areas for improvement: exploring alternative evaluation methods and using a larger T5 model for enhanced linguistic performance.

References

Abdul Ghafoor Etemad, Ali Imam Abidi, and Megha Chhabra. 2021. Fine-tuned t5 for abstractive summarization. *International Journal of Performability Engineering*, 17(10):900.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second ...*

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, and Yanqi Zhou. 2020. mt5: A massively multilingual pretrained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008.

Guanghua Wang and Weili Wu. 2023. Surveying the landscape of text summarization with deep learning: A comprehensive review.

Mingye Wang, Pan Xie, Yao Du, and Xiaohui Hu. 2023. T5-based model for abstractive summarization: A semi-supervised learning approach with consistency loss functions. *Applied Sciences*, 13(12).

Lei Zhang, Furu Liu, Dongyan Zhao, Yang Wang, and Qun Liu. 2020a. An encoder-decoder architecture with graph convolutional networks for abstractive summarization. *arXiv preprint arXiv:2010.11934*.

Tianyi Zhang, Ryan Skerry-Ryan, and Geoffrey Hinton. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

Example 1

Generated Summary: evidence has emerged of a wartime blunder that resulted in millions of pounds of Russian bullion being lost in the river Clyde. the bullion was to pay the Americans for arms shipped to Russia during one of the most crucial stages of the Second World War.

Target Summary: Edinburgh engineer detailed how bullion was lost during secret mission . Leonard H. Thomas served on HMS Ulster Queen during Arctic Convoys . His secret diaries reveal a crate of bullion was lost while being transferred . Mr Thomas wrote it fell into river Clyde while being moved to another ship . The mission was so secretive it is not known if bullion was ever recovered .

Example 2

Generated Summary: former president, 90, fell ill in December and was hospitalized after complaining of shortness of breath. he and his wife

Table 4: Comparison of Summaries

Summary Type	Text
Reference Summary	Former US president George H.W. Bush looked in good health on Friday as he accompanied Australian tennis player Roy Emerson to a tennis game. Bush, 90, lives in the Houston Area with his wife Barbara and has attended many ATP fundraisers and charity events throughout the years. Back in December, Bush, who suffers from Parkinson's, was hospitalized for shortness of breath for about a week.
Generated (T5) Summary	former president, 90, fell ill in December and was hospitalized after complaining of shortness of breath. he and his wife have attended many ATP fundraisers and charity events throughout the years.
ChatGPT4 Summary	Former US President George H.W. Bush, looking healthier, attended the U.S. Men's Clay Court Championship in Houston with tennis legend Roy Emerson. After being hospitalized in December for health issues, Bush appeared in good spirits at the event. An avid fan of ATP World Tour 250 events, he, alongside his wife Barbara, has supported many ATP fundraisers. Despite battling Parkinson's and past health scares, Bush's outing reflects his resilience and ongoing interest in tennis.

have attended many ATP fundraisers and charity events throughout the years.

Target Summary: Former US president George H.W. Bush looked in good health on Friday as he accompanied Australian tennis player Roy Emerson to a tennis game . Bush, 90, lives in the Houston Area with his wife Barbara and has attended many ATP fundraisers and charity events throughout the years . Back in December, Bush, who suffers from Parkinson's, was hospitalized for shortness of breath for about a week .

Example 3

Generated Summary: service of process is an industry and its own body of law premised on one guiding principle. if you are going to sue someone, you should at least let them know about it. if you are going to sue someone, you may avoid bad news in life.

Target Summary: A court allowed a wife to serve divorce papers via Facebook . Danny Cevallos: Why not let people be found via social media?