Text Summarization for Scientific Papers

Botong Zhang
bzhang16@sas.upenn.edu

Han Yang
hanyang3@seas.upenn.edu

Ruixi Gong
gruixi@sas.upenn.edu

Yang Dong
flankado@sas.upenn.edu

Yuxuan Gao
gaoyx@seas.upenn.edu

Abstract

Text summarization is a crucial process for condensing extensive information into concise and informative summaries. This project delves into various text summarization techniques, spanning from simple baselines to advanced models. The primary focus is on extending our initial strong baseline, a T5 (Text-to-Text Transfer Transformer) model, with four distinct approaches. The first extension employs the PyTorch Lightning pipeline for fine-tuning, accompanied by a preprocessing strategy targeting informative sections like abstracts and introductions. The second extension leverages the LongT5 model, designed for efficient handling of sequences surpassing the conventional 512-token limit. The third extension introduces Section-wise Summarization, utilizing the T5 model to independently summarize each section before concatenating for a comprehensive document summary. The fourth extension explores Hierarchical Summarization, directly extracting abstracts and conclusions while applying intermediate hierarchical layers to each section. Performance evaluation, particularly focusing on ROUGE-L scores, demonstrates the effectiveness of these extensions compared to the strong baseline. Especially for the first extension, our top-performing model, we achieved a notable ROUGE-L score of 0.7172.

1 Introduction

Abstractive Summarization involves generating a summary that surpasses mere content extraction by rephrasing and restructuring information in a concise form. In our project, we focus on this task using one of the most recent and innovative transformer models, T5 (Text-to-Text Transfer Transformer), which treats text processing as a "text-to-text" problem—taking text as input and producing new text as output (Raffel et al., 2019).

In Figure 1, the input text undergoes abstraction to produce a summarized output. The choice of text summarization as our term project's focus stems from its widespread significance in information retrieval and document understanding. With the continuous proliferation of textual data, the demand for automated summarization techniques becomes increasingly apparent. Our project aims to explore and extend existing models, specifically leveraging advanced transformer architectures like T5 and LongT5, to enhance the performance of text summarization tasks.

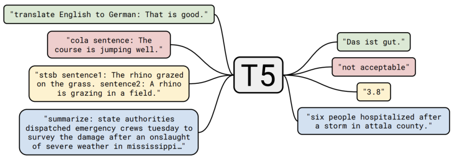
"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Figure 1: Illustration of Abstractive Summarization: A diagram of our text-to-text framework. The summarization task we are working on is cast as feeding our model text as input and training it to generate some target text. "T5" refers to the model, which is dubbed the "Text-to-Text Transfer Transformer.

Our project extends the baseline T5 model, incorporating techniques like the PyTorch Lightning pipeline, LongT5 model, Section-wise Summarization, and Hierarchical Summarization. The remainder of the paper is structured as follows: in the following section, we discuss our experimental design, describe the data, present evaluation metrics, and introduce the simple baseline model with its implementation and performance on the test set. In Section 4, we delve into the implementation of the published strong baseline (T5) and all the extensions we've experimented with, along with an error analysis of our system's output.

2 Literature Review

In the realm of information retrieval and document understanding, text summarization plays a crucial role. Our project seeks to extend the capabilities of the Text-to-Text Transfer Transformer (T5) model, a significant advancement in this field. This endeavor is grounded in a variety of key studies, each contributing to our understanding and application of sophisticated neural network architectures and natural language processing techniques.

At the core of our project lies the innovative work of Raffel et al., who not only introduced the Text-to-Text Transfer Transformer (T5) model but also profoundly explored transfer learning in natural language processing (NLP). The T5 model, pivotal in our approach, adopts a "text-to-text" framework, treating all text processing tasks as a conversion from input to output text, a concept central to our focus on abstractive summarization (Raffel et al., 2019). This approach involves more than just extracting content; it rephrases and restructures the text to create concise and meaningful summaries. Moreover, Raffel et al.'s work highlights the significance of unsupervised pre-training in NLP, using extensive, unlabeled datasets like the Colossal Clean Crawled Corpus (C4). This comprehensive view of transfer learning's potential in various NLP tasks, particularly in text summarization, is instrumental in guiding our project's methodology and objectives.

We further explore the capabilities of the LongT5 model, an extension of the T5 designed to handle extended sequences exceeding the usual token limitations. This aspect of our work is inspired by Vaswani et al. (Vaswani et al., 2023), who revolutionized sequence processing in NLP with the introduction of the Transformer architecture. The LongT5's efficiency in managing long texts is vital for summarizing extensive academic documents.

The project also integrates insights from a paper focusing on the mathematical modeling of sentence summarization (Rush et al., 2015). It defines the task as producing a condensed summary from an input sentence, represented as a sequence of M words, $\mathbf{x}_1, \ldots, \mathbf{x}_M$, from a fixed vocabulary $\mathcal{V}$. Each word is an indicator vector in $\{0,1\}^V$, with sentences as sequences of these indicators. The paper contrasts abstractive summarization, which involves finding an optimal sequence $\mathbf{y} \in \mathcal{Y}$ that maximizes a scoring function $s(\mathbf{x}, \mathbf{y})$, with extractive summarization, which focuses on selecting words directly from the input. This delineation of summarization strategies and the mathematical framework provided are integral to our project, as they offer a structured approach to understanding and developing summarization models.

Another pivotal study in our research is an exploration of abstractive text summarization using Attentional Encoder-Decoder Recurrent Neural Networks (Nallapati et al., 2016). Addressing challenges like keyword modeling and the hierarchy of sentence-to-word structures, this paper provides critical insights for advancing abstractive techniques, especially relevant as we tackle the complexity of summarizing extensive academic texts.

Collectively, these studies provide a nuanced understanding of both the theoretical and practical aspects of text summarization. They emphasize the importance of advanced neural network architectures, the necessity of model interpretability, and the critical role of mathematical frameworks in summarization. Our project aims to amalgamate the strengths of both extractive and abstractive approaches, underpinned by these foundational studies, to effectively summarize scientific and academic texts.

3 Experimental Design

3.1 Data

Our project leverages a dataset consisting of the 1,000 most cited papers from the ACL Anthology Network (AAN), focusing specifically on 1,009 papers with annotated summaries (Radev et al., 2013). We used the annotated summaries as ground truth labels for the training, testing, and further analysis of our models. The github repo link: https://github.com/Cralence/CIS5300-Project.

The dataset is divided into training, validation, and test sets, comprising 801, 99, and 109 papers respectively, used for model training, validation, and performance evaluation.

Avg. citation count	61 (Range: 21 - 928)
Avg. text length per RP	4,417 words
Avg. abstract length	110 words
Avg. sampled citing sentences per RP	15
Avg. citing sentences selected for summary	2
Inter-annotator agreement (Kappa)	0.75
Avg. length of gold summary	151 words

Table 1: ScisummNet reference paper (RP) dataset detailed statistics

3.2 Evaluation Metric

Our evaluation script uses the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric for assessing the quality of text summarization. ROUGE measures the overlap between the system-generated summaries (hypotheses) and reference (gold standard) summaries. Specifically, the script will output the average ROUGE scores, including ROUGE-1 (measures the overlap of unigrams between the generated summary and the reference summaries), ROUGE-2 (measures the overlap of bigrams between the generated summary and the reference summaries), and ROUGE-L (measures the longest common subsequence between the generated summary and the reference summaries), which capture word n-gram overlap and longest common subsequence, respectively. In specific, the ROUGE-N has the formula:

$$\text{ROUGE-N} = \frac{\sum_{s\in\{\text{Ref Sum}\}} \sum_{gram_n \in s} Count_{match}(gram_n)}{\sum_{s\in\{\text{Ref Sum}\}} \sum_{gram_n \in s} Count(gram_n)} \quad (1)$$

where $Count_{match}(gram_n)$ is the count of n-grams in the candidate summary that match the reference summary, and $Count(gram_n)$ is the count of n-grams in the reference summary. This metric was introduced by Lin (Lin, 2004). These scores provide insights into the accuracy and fluency of the generated summaries.

3.3 Simple baseline

Our simple baseline for the text summarization task processes each document by extracting the first few sentences of the abstract, assuming these sentences capture the essence or key points of the document. This method is commonly used as a naive approach in summarization tasks due to its simplicity and ease of implementation. The following graph is the experiment investigating the performance change with repect to the number of prefix sentences selected for our baseline model. From the result, we can see that the best rouge scores are reached when the first 5 sentences are chosen as the prediction, achieving ROUGE-L=0.510.

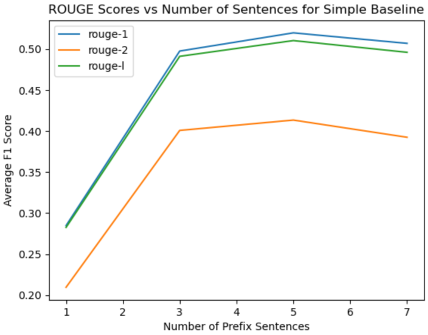

Figure 2: Rouge scores vs. number of different predix sentences for baseline model

4 Experimental Results

4.1 Published baseline

For the strong basline, our project utilized the T5 (Text-to-Text Transfer Transformer) model, as outlined in (Raffel et al., 2019). Sepcifically, we utilized the T5ForConditionalGeneration class from the Hugging Face Transformers library to implement this model. The training dataset includes training, validation, and test splits, with the AdamW optimizer and a linear learning rate scheduler facilitating effective learning. Upon evaluation on the test set, the model achieved an average ROUGE-L score of 0.5804. This performance was not quite satisfying, primarily due to our constraint on summary and text lengths, which were kept short. Future extensions will focus on optimizing these lengths to 512 and 1024 tokens, respectively, to enhance model performance.

In the paper "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer" by Raffel et al. (2019), the authors demonstrate the fine-tuning of the T5 model on various datasets and tasks through a comprehensive experimental setup (Raffel et al., 2019). For the summarization tasks, they fine-tuned on datasets like CNN/Daily Mail to compress and paraphrase long articles into concise summaries, resulting in ROUGE-L score of 0.3940.

While our fine-tuned model approaches the accuracy levels reported in the aforementioned study, a direct comparison of metrics is not feasible due to differences in datasets, training approaches, and domain specificity. Owing to the structured and consistent nature of our scientific document dataset, our model likely benefits from domain familiarity, resulting in elevated ROUGE scores.

4.2 Extensions

Model	Recall (r)	Precision (p)	F1 Score (f)
imp. T5-Base	**0.6921**	**0.7730**	**0.7172**
LongT5	0.5977	0.7772	0.6615
Section-wise 3	0.6455	0.6155	0.6043
Hierarchical 4	0.6493	0.6497	0.6343

Table 2: Comparison of ROUGE-L Scores Across Different Models

4.2.1 Extension1: Preprocessing and T5 Fine-tuning (best)

The first extension of our text summarization project involves a refined preprocessing strategy and fine-tuning the T5 model using the PyTorch Lightning framework. We specifically targeted informative sections of documents, such as abstracts and introductions, for extraction. This selective approach ensures focus on the most crucial summary points. Key training parameters included a 't5-base' model, 1024 tokens for text length, 512 tokens for summary length, a batch size of 5, and 10 epochs. This extension significantly improved performance, achieving ROUGE-1, ROUGE-2 and ROUGE-L of 0.723, 0.632 and 0.717 respectively, indicating the enhanced accuracy and coherence of the generated summaries.

4.2.2 Extension2: LongT5 Fine-tuning

In the LongT5 model implementation, we configured the text length to 4000 tokens and achieved a ROUGE-L score of 0.6615. Notably, this model's performance was slightly lower compared to the first extension. A key factor contributing to this outcome was our deliberate decision to omit the extraction of abstracts, introductions, conclusions, and the first two sentences from each paragraph. By excluding these sections, which are typically rich in summary content, we aimed to challenge the model's capability to discern and condense the essence of a document from its main body. This approach was intended to provide insights into the model's performance in scenarios where conventional summary sections are absent or less informative, thus evaluating its adaptability and effectiveness in a wider range of summarization contexts.

4.2.3 Extension3: Section-wise Summarization

In Extension 3, our approach involved Section-wise Summarization, where we used the T5-base model to independently summarize each section of the scientific document. After summarizing, these individual section summaries were concatenated to compile the final comprehensive document summary. Our intuition behind this extension is that by individually summarizing each section of the scientific paper, the context and importance of each section are maintained by the model. This approach tends to generate more coherent overall summarization with each section equally valued. This method resulted in a ROUGE-L score of 0.6043. The reason of a relatively low ROUGE scores might due to the limitation of our metric used. ROUGE is primarily based on the overlap of n-grams between the generated and reference summaries. It doesn't directly measure the semantic meaning or the quality of the information conveyed. If the model produces semantically accurate summaries that use different phrasing or synonyms not present in the reference, ROUGE scores might not reflect the actual quality of the summary.

4.2.4 Extension4: Hierarchical Summarization

In extension 4, Hierarchical Summarization was implemented by initially extracting the opening sentences of the abstract and conclusion sections. We then applied hierarchical layers to the intermediate sections of the text, ultimately leading to a prediction layer aimed at generating a summary with a word count of approximately 150, which is the average length of the ground truth summaries obtained by our previous analysis of the dataset. The method is useful when we have long and detailed text but want to obtain concise summary, which is usually the type of task we are facing when summarizing scientific documents. This strategy culminated in the creation of a comprehensive hierarchical summary. This method achieved a ROUGE-L score of 0.6343.

4.3 Error analysis

In our analysis of the text summarization model, we identified several key error types. The predictions and ground truth are saved in *sum_pred.txt* and *sum_gt.txt* respectively. The errors include repetitiveness, omission of key information, misinterpretation of focus, redundancy, and lack of coherence.

The model often repeats phrases or ideas within a single summary, as observed in about 20% of the summaries, exemplified by repetitive mentions of "The annotation scheme used a total of 26 attributes to represent flights." Important details or aspects are frequently omitted, affecting around 30% of summaries, such as crucial details about the Joyce system's tasks in the "Applied Text Generation" paper. About 15% of summaries misinterpret the main focus, like the summary of "Subjectivity in Language," which incorrectly focused on adjectives instead of gradable adjectives' impact on subjectivity.

Redundancy is another common issue, found in approximately 25% of summaries, where redundant information does not contribute to overall understanding, seen in "Statistical Significance Testing." Finally, around 10% of summaries suffer from a lack of coherence, making them difficult to follow, as in the scattered representation in "Non-Projective Dependency Parsing."

These errors indicate a need for improvements in the model's ability to understand central themes and generate concise, non-repetitive summaries.

5 Conclusions

In conclusion, our term project embarked on a journey to enhance the capabilities of text summarization using advanced neural network models, primarily focusing on the T5 (Text-to-Text Transfer Transformer) and its extensions. Throughout the project, we diligently implemented and evaluated various models. The most notable progress was observed in the first extension, where targeted preprocessing and fine-tuning within the PyTorch Lightning framework led to a significant improvement in the ROUGE-L score (from 0.5804 to 0.7172). However, while our models demonstrated significant advancements, they did not quite reach the pinnacle of state-of-the-art performance.

The primary reasons for this gap can be attributed to several factors. Firstly, the complexity of the text summarization task itself posed substantial challenges, especially when dealing with intricate academic texts. Secondly, our models, though effective, still encountered issues such as repetitiveness and omission of key details in summaries, indicating room for further optimization in understanding and processing nuanced language structures.

Moreover, resource constraints limited the extent of training and fine-tuning possible for our models, which is crucial for achieving top-tier performance in such sophisticated tasks. This was evident in the comparison of our results with current state-of-the-art models, where the lack of extensive training and optimization emerged as a key differentiator.

In essence, our project made commendable strides in text summarization, providing valuable insights into the challenges and potential solutions in this domain. It laid a solid foundation for future exploration and set a clear path for improvements that could bridge the gap to state-of-the-art performance. The learnings from this project not only contribute to academic discourse but also present practical implications for the development of more refined and efficient text summarization tools in the field of Natural Language Processing.

References

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

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *The SIGNLL Conference on Computational Natural Language Learning (CoNLL)*.

Dragomir R Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The acl anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944.

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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

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