Long Document Summarization with Prioritized Input

*David Ristau**   *Ratan Deep Singh**
Berkeley School of Information
dfristau@ischool.berkeley.edu, *ratan_singh@ischool.berkeley.edu*

Abstract

Document summarization is an important sub-field of natural language processing, with the potential to help people save time and make decisions on what content to read in depth. Much of the success in summarization has been applied to short documents, however the summarization of longer documents has the potential to be far more beneficial. In this paper we explore multiple techniques of prioritizing the content of long documents before using a transformer model to generate summaries. We report positive results for both rouge-1 and rouge 2 scores using prioritized text to fine-tune a transformer model and generate summaries compared to using standard input text.

1. Introduction:

The field of long document summarization has made significant progress in recent years, but remains an open problem. The crux of long document summarization lies in the limitations of the maximum size of document that can be fed into modern day transformer models [8]. Some of the most successful models in long document summarization leverage a sparse attention mechanism to enable longer contexts to be evaluated [1,8].

Despite the increase in context length capabilities of sparse attention models, there remain lots of documents that exceed the capacity of these larger models. The pubmed dataset, used in this paper, is commonly used in research for long document summarization. 36% of documents in this dataset exceed the maximum token length of 4096 for a sparse attention model. There are two primary strategies for effectively summarizing documents that are longer than the input token length. The first is truncating the document to fit within the capacity of the input sequence length. The second is to use a form of sliding window approach [5]. In this paper we focus on the prior.

The primary contribution of this paper is to evaluate the impact of text prioritization on summary generation. We approach this problem by using text prioritization with models that are fine tuned on both standard and prioritized text. We find that fine tuning a model on prioritized text has a positive impact on the rouge-1 and rouge-2 scores of the resultant summaries.

2. Background:

Within the domain of automatic text summarization, there are two primary methods: extractive and abstractive. Extractive methods rely on a model to extract the most important sentences within a document, and use those sentences as the summary. Abstractive methods generate a summary text sequence using the body of the document as an input sequence. Extractive methods have historically been the primary focus of much of the research in the

*equal contribution
Code available at: https://github.com/dfristau/ristau_singh_w266_project

field, but with the popularity of large sequence to sequence models, abstractive summarization is becoming more popular. Sub domains within the field of automatic text summarization include short document summarization, and long-document summarization [5]. Of these sub-domains, long document summarization is largely seen as the more challenging, with the least satisfactory solutions thus far [5].

A simple approach to automatic text summarization, published in 2004, is called lexrank [6]. This approach uses a graph based methodology to identify sentences with a high centrality, and extract them as a summary. This approach can be applied to long documents and short documents as algorithms can accommodate large texts with relative ease. This method is largely out of date, but represents common early approaches to automatic text summarization.

More recent approaches leverage transformer models. Some of these models use a full attention mechanism, Like BERT [4]. This can work with short documents, or documents that have most of the important information in the first 512 tokens, but for tasks that require a longer context these models struggle to effectively summarize the content [8].

The most recent and successful models in this domain modify the attention mechanism in transformer models. These approaches substitute the standard full attention mechanism with a sparse attention mechanism. This results in a model that scales linearly in memory usage as opposed to quadratically with input sequence length [1, 8]. Two common models that share this approach are Longformer, which implements global and local-window attention [1], and BigBird, which implements a combination of global, local-window, and random attention [8]. Both of these models are capable of taking in an input sequence up to 4096 tokens.

This substantial input sequence length increase shows promise in the domain of summarizing long documents. However, If the documents to be summarized exceed this length then a form of truncation or rolling window approach must still be used.

3. Data:

This paper uses the PubMed dataset, which was originally introduced by *Cohaen et al. 2018* [3] for use in the domain of long document summarization. This dataset consists of 119,924 training documents, 6,633 validation documents, and 6,658 test documents. Each document represents a scientific paper acquired from the National Center for Biotechnology Information (NCBI), and contains the body of the paper and the abstract, which is used as the gold standard summary.

We generate two modified versions of this dataset, prioritizing the sentences of the body of the scientific papers, but leaving the abstracts the unmodified. The details of how we generate these modified datasets are discussed in the methods section.

4. Methods:

In this paper we pursue two paths of research, both leveraging the concept of prioritizing the text input to a transformer model. The details of each method are outlined in this section. We use rouge metrics [2] for all evaluation criteria, in accordance with common practices in the field automatic text summarization.

Data Modification:

A primary task of this research is to generate an effective method for prioritizing the input text before generating a summary via a pre-trained transformer model. This approach can ensure that the most pertinent information is included in the transformer input sequence, even with a limit of 4096 tokens.

For the task of prioritization we use the lexrank algorithm [6], but we propose alternative methods of prioritization could be used in future research. Lexrank is a graph based approach to determining the relative importance of sentences in a document. Sentences are ranked highly if they have a high cosine similarity to many other sentences that are also ranked highly, much like the page rank algorithm [7]. A visual representation of a graph derived using the lexrank algorithm is shown in figure 4.1. In this figure the individual nodes represent a sentence and the edges represent the weight of the connections between other sentences. A node with an overall high degree of edge weights is hypothesized to represent a central topic to the document.

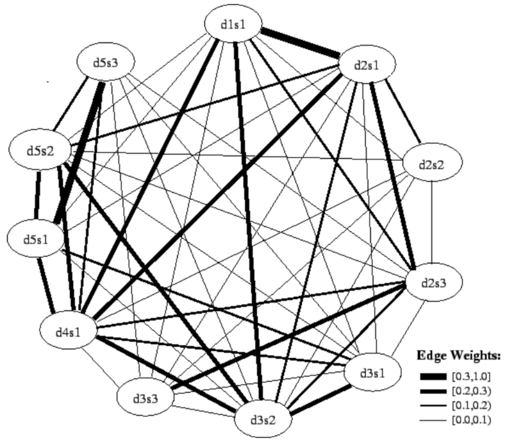

Figure 4.1. Visualization of a lexrank network from the paper: LexRank: Graph-based lexical centrality as salience in text summarization [6].

We use the lexrank algorithm to generate two prioritized datasets from the original pubmed data.

The first, and simpler approach is achieved by simply running each document through the lexrank algorithm and outputting the sentences of the document in order of importance. This ensures that if the document must be truncated (in the case that it exceeded 4096 tokens) then it will contain the most important sentences to that document. This procedure was executed for the train, validation, and test splits of the pubmed dataset.

The second approach that we implement is to first use the lexrank algorithm to rank each sentence in the document. We then evaluate the tokenized length of the document and iteratively remove the least important sentences until the document length is below the threshold of 4096 tokens. After this truncation procedure is complete, we then reorder the sentences temporally. This has the effect of removing the least important sentences within a document, but maintaining the original order of each sentence and section within the document.

Each of the above procedures were executed on the train, validation, and test splits of the pubmed dataset. The output of these procedures is a file with the prioritized documents and the corresponding unmodified abstract for use in evaluating the predicted summary.

Data Prioritization Without Fine Tuning:

In this task, we use the original prioritized datasets described in the above section with the ‘bigbird-pegasus-large-pubmed’ model from hugging face. The model is pre-trained, then fine tuned on the original pubmed dataset, but not fine tuned with the prioritized data. The baseline for this task is achieved by evaluating summaries generated from the original pubmed dataset using this model. We then generate summaries using the BigBird-Pegasus model using our prioritized data as input, but still without any pre-traing. We repeat this process for both the ranked, and ranked and temporally corrected datasets, and compare performance to the baseline. The results of these experiments are presented in the discussion section of this paper.

Fine Tuning on Prioritized Data:
Our second task is to pre-train a transformer model on the prioritized datasets to determine if further performance improvements can be achieved.

Due to both runtime and financial limitations, several compromises were made at this stage of the project. We select the 'longformer2roberta-cnn_dailymail-fp16' from hugging face, which can be fine-tuned with fewer resources than the BigBird-Pegasus model mentioned in the previous section. Additionally, only 40,000 documents of the available 119,924 were used in training. These compromises enable us to generate results to determine if pre-training on our prioritized data has a positive effect. We acknowledge that these results do not compete with current state of the art methods. As a result our claims are limited to the directional shift in performance that prioritized data fine tuning has, and not the absolute magnitude of the increase or decrease.

A baseline for this task is achieved by fine-tuning the pre-trained Longformer-roBERTa model on the unmodified pubmed dataset. We train with 40,000 training documents and evaluate loss on 3,000 validation documents for 1 epoc. The same training procedure was then repeated for the ranked and ranked and temporally corrected datasets. The results of these experiments are presented in the discussion section of this paper.

5. Results and Discussion:

Prioritized Data Without Fine Tuning:
The results of our first task are shown in table 5.1. In each rouge metric, the best performance is achieved with our ranked, then temporally aligned dataset.

To evaluate if this performance increase is statistically significant, we perform a welch t-test. The results of this test give insignificant p-values for each metric. While we cannot claim a significant difference in performance using prioritized data, this test does demonstrate that prioritizing the input to a summarization model does not negatively affect performance.

Model \Metric	**R1**	**R2**	**RLsum**
Standard Input	42.16	17.48	26.14
Ranked Input	42.60	17.24	25.88
Ranked and Temporarily Aligned Input	**42.51**	**18.02**	**26.48**

Table 5.1 - Results generated using a BigBird-Pegasus model trained on the pubmed dataset with standard and prioritized input.

We show the models performance on specifically documents longer than 4096 tokens, in table 5.2. These results show uniform improvement using either method of prioritization as input. These results still do not show statistical significance, however the lack of negative impact is notable given the model was not fine tuned on prioritized data. These results motivate our second task of fine tuning with prioritized data.

Model \Metric	R1	R2	RLsum
Standard Input	40.60	14.62	24.17
Ranked Input	**41.27**	15.14	**24.76**
Ranked and Temporarily Aligned Input	41.26	**15.35**	24.53

Table 5.2 - Results generated using a BigBird-Pegasus model trained on the pubmed dataset with standard and prioritized input of documents of token length > 4096.

Prioritized Data With Fine Tuning:

Table 5.3 presents results generated by fine tuning the Longformer-roBERTa model. These results were generated by training on the standard pubmed dataset and our prioritized datasets, then evaluating the performance of each model on a test set with the same prioritization treatment as the training data.

The baseline model, fine tuned and evaluated using the standard pubmed dataset, shows the lowest performance in both rouge-1 and rouge-2 score but the best performance in rouge-lsum score by a large margin. Comparing the distributions shown in figure 5.1 demonstrates highly significant results for each rouge metric. We therefore reject the null hypothesis that these rouge scores are drawn from the same distribution. This result supports the alternative hypothesis that our prioritization tasks have a positive impact on rouge-1 and rouge-2. These results additionally demonstrate that this treatment has a negative impact on rouge-lsum.

The best performance in rouge-1 and rouge-2 score is observed when the model is fine tuned on text that is in ranked order by sentence importance, and evaluated using a test set that is prioritized in the same manner. This differs from our expectations, based on intuition. Intuitively, reading a document that has its sentences re-ordered based on the importance will be confusing. We anticipated this intuition would

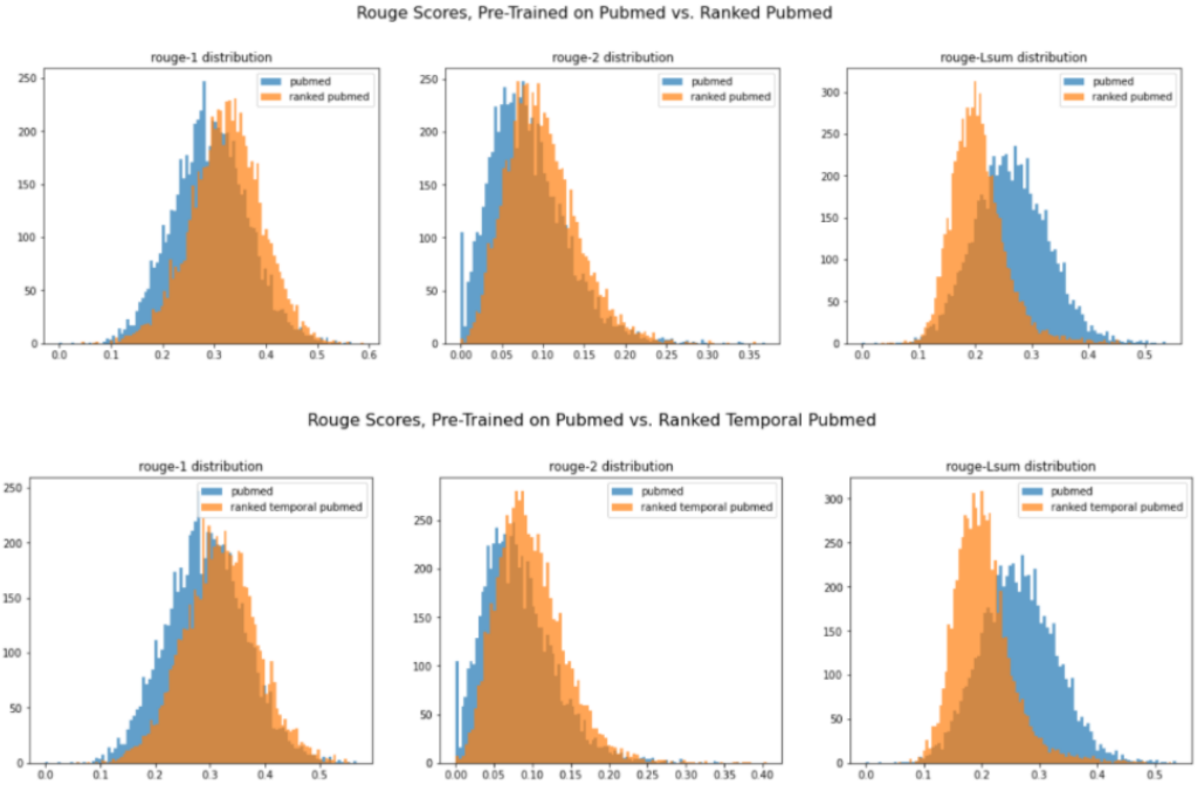

Figure 5.1 - Rouge score distributions of baseline longsum-roberta model and prioritized text techniques.

be reflected in these results and are surprised to find the opposite to be true.

Model \Metric	R1	R2	RLsum
Standard fine tuning and Input	29.24	8.20	**26.40**
Ranked fine tuning and Input	**32.39***	**9.86***	20.56*
Ranked and Temporarily Aligned fine tuning and Input	31.93*	9.81*	20.28*

**indicates statistically significant results.*

Table 5.3 - Results generated using fine tuning a Longformer-roBERTa model trained and evaluated on standard and prioritized pubmed documents

We evaluate these models again, considering only documents longer than 4096 tokens to determine if these prioritization methods outperform the baseline when truncation is applied. In this context, the ranked dataset will consist of the first 4096 tokens from the most important sentences in the document. The dataset with a temporal correction, will contain the most important sentences up to the 4096 token limit, and then put each sentence back in original order, effectively removing non-informative sentences within the body of the document.

The results reported in table 5.4 demonstrate that both prioritization methods yield better results for rouge-1 and rouge-2 score, and show a similar reduction in performance in rouge-lsum score on longer documents.. These results also show highly significant p-values. Notably, our method of truncating by removing the least important sentences followed by a temporal correction appears to outperform both the pure ranking approach and baseline. This indicates this method of prioritization is better able to cope with documents longer than the input sequence of the Longformer model.

Model \Metric	R1	R2	RLsum
Standard fine tuning and Input	27.77	7.21	**25.12**
Ranked fine tuning and Input	30.65*	8.83*	19.70*
Ranked and Temporarily Aligned fine tuning and Input	**31.13***	**9.16***	19.81*

**indicates statistically significant results.*

Table 5.4 - Results generated using fine tuning a Longformer-roBERTa model trained and evaluated on standard and prioritized pubmed documents of token length > 4096.

To evaluate if these prioritization methods have a negative impact on documents less than 4096 tokens in length, we compare the performance of short documents only. We hypothesized that since these documents are able to fit entirely in the input sequence of a Longformer encoder, the prioritization methods would have a negative impact on performance.

The results shown in table 5.5 show that models fine tuned and evaluated on prioritized text outperformed the model with non-prioritized text on shorter documents. Each of these results show highly significant p-values. These results are unexpected. The most surprising results come in the case of the ranked and temporally corrected results. This is due to the fact that documents shorter than 4096 tokens will be the same in the original dataset and the ranked and temporally corrected dataset because no truncation is necessary. These results suggest that fine tuning on prioritized text has a benefit even in the case of shorter documents where truncation is not an issue.

Model \Metric	R1	R2	RLsum
Standard fine tuning and Input	30.04	8.73	**27.09**
Ranked fine tuning and Input	**33.34***	**10.42***	21.02*
Ranked and Temporarily Aligned fine tuning and Input	32.37*	10.17*	20.54*

**indicates statistically significant results.*

Table 5.5 - Results generated using fine tuning a Longformer-roBERTa model trained and evaluated on standard and prioritized pubmed documents of token length < 4096.

The results generated by pre-training and evaluating on prioritized text are promising and indicate that this strategy can be applied to improve rouge-1 and rouge-2 summarization scores.

6. Limitations

The primary limitations to this study are twofold. While the results of our experiments demonstrate improvement in rouge-1 and rouge-2 scores, the corresponding reduction in rouge-lsum score indicates these techniques do not uniformly improve summarization. We believe this result warrants further research to determine if further training can correct this reduction. The second limitation to this study is our inability to train our models long enough due to time and financial constraints. Therefore the results generated from our pre-training task do not reflect a model trained to its fullest performance potential. We acknowledge that further training may shift the distributions of performance, and thus change the results of this study. We believe this research to be a promising starting point for prioritized truncation techniques nonetheless and believe the strategy should be further researched.

7. Conclusion

Prioritizing text prior to applying a transformer based summarization model can have significant impacts on the effectiveness of summaries generated for long documents. We have shown that fine tuning on text that is prioritized both by pure sentence rank, and by sentence rank with a temporal correction has significant benefits to rouge-1 and rouge-2 sores. Most notably the technique of pre-trainng on prioritized text outperforms pre-training on standard text in documents that are both longer and shorter than the input sequence of the transformer model. These results indicate an inherent benefit of fine-tuning on prioritized input for the task of summarization.

8. References

[1] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. *ArXiv*. Retrieved November 22, 2021, from https://arxiv.org/abs/2004.05150.

[2] Chin-Yew, L. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Association for Computational Linguistics*.

[3] Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., & Goharian, N. (2018). A discourse-aware attention model for abstractive summarization of long documents. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. https://doi.org/10.18653/v1/n18-2097

[4] Devlin, J., Chang, M.-W., Lee, K., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*. Retrieved November 22, 2021, from https://arxiv.org/pdf/1810.04805v2.pdf.

[5] El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2020). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 113679.

[6] Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479. https://doi.org/10.1613/jair.1523

[7] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. *Stanford InfoLab*.

[8] Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2021). Big Bird: Transformers for Longer Sequences. *ArXiv*. Retrieved November 22, 2021, from https://arxiv.org/abs/2007.14062.