Multi-documentation Summarization of Science Articles

Justin To, Luka Liu, Milan Dean

W 266: Natural language Processing
UC Berkeley School of Information

April 13, 2023

Abstract

In today's age of information overload, the ability to generate accurate and concise summaries of large bodies of text is more important than ever. This study focuses on the task of abstractive multi-document summarization (MDS), which involves generating a summary based on multiple related documents. Two state-of-the-art pre-trained large language models (LLMs), Centrum and LED, were selected, and different strategies, including fine-tuning and model-stacking, were explored for building an MDS model customized for Multi-XScience, a dataset previously unseen by the pre-trained models. In evaluating the summaries generated, we adopted both quantitative ROUGE-based scoring and qualitative analysis of the model outputs, in order to understand how the models adapted to unique elements specific to the task and dataset. By studying the process and strategies for adapting pre-trained models to domain-specific MDS tasks, we aim to contribute to the growing body of research on automated text summarization and address the ongoing challenge of information overload.

1 Introduction

MDS is the task of summarizing multiple texts into a concise and informative summary. Compared to single document summarization, it is challenging in the context of abstractive MDS to keep the summary coherent and comprehensive, given the documents are longer[1] and of a more complex structure[2]. Despite the challenges, MDS has potentially a wider application as many real-world tasks involve collating, digesting, and paraphrasing information from multiple sources. An important question therefore is, given the advances in the NLP domain and the availability of pre-trained LLMs, how should we build a customized abstractive MDS model if we are given a domain-specific dataset previously unseen by the models?

In this paper, we employ the Multi-XScience dataset, a relatively recent dataset that hasnot been widely used for pre-

[1] For instance, the widely used CNN/Daily Mail summarization dataset has an average token length of around 780 tokens on average. In contrast, the Multi-XScience dataset has around 30% of the samples with over 1024 tokens in the inputs. The maximum length exceeds 6,300 tokens.

[2] For instance, in MDS, the source articles are written by multiple authors with varying writing styles, and article structures that the model cannot easily leverage.

training/fine-tuning LLMs, to simulate the situation where we are given an MDS task for a broad domain, e.g. academic writing from the sciences. The defining characteristics of the dataset, namely (a) the significant portion of samples with long inputs (e.g. up to ∼6,300 tokens) and (b) how the different articles within each sample are related yet not covering the exact same event[3], led us to pick two models for experimentation. The first is the Longformer Encoder Decoder (LED) model, which was proposed by Beltagy et al. [1] and introduced the concept of local vs global attention so that attention-based transformers can still effectively handle the computation of long inputs. The second is the Centrum model, proposed by Puduppully and Steedman [6], for its capability of handling long inputs and inclusion of a centroid-based approach for dealing with multiple source documents.

Since both LED and Centrum have pretrained checkpoints available, we first tested the models off-the-shelf using the Multi-XScience dataset and then proceeded to fine-tune the models for improved performance. Additionally, we also explored the use of a two-step setup whereby the individual source articles of each sample are first shortened as a normal summarization task, and then combined again for a second-step MDS task. For evaluation, we employed the common ROUGE score metric for the off-the-shelf, fine-tuned, and two-step models, showing the improved performance from fine-tuning (ROUGE-2 from 5.2 to 6.9; ROUGE-L from 14.6 to 17.8) and close to state-of-the-art (SOTA) results. Further analyses were also conducted on the variation of performance across inputs of various lengths, and the amount of "copying" from the main article. Lastly, a qualitative analysis was done on 25 samples from the generated summaries for all models tested, and the fine-tuned and two-step models were found to have shown marked performance in extracting different information from the articles, contrasting different, and generally adapting to the writing styles of the dataset's labels.

2 Related Works

In recent years, there has been significant research in the field of multi-document summarization, with a particular focus on the use of transformer-based models for abstractive MDS. In addition, much progress has been made in providing more datasets for MDS training and evaluation.

On model development, the LED model proposed by Beltagy et al. [1] addresses the issue of processing long documents by introducing a sparse attention mechanism that allows the model to scale to documents of up to 16,384 tokens in length. This approach has outperformed previous models, including advanced models such as PEGASUS, on long-document tasks such as summarization.

Branching off to models more specialized in the MDS context, PRIMERA (Pyramid-based Masked Sentence Pre-training for Multi-document Summarization), proposed by Xiao et al. citexiao2022primera, is a pre-training method for transformer-based models that leverages hierarchical sentence representations to improve performance on multi-

[3] As we will further explain in subsequent parts, this is a unique feature of the Multi-XScience dataset. As a comparison, the commonly used Multi-news dataset has, in each sample, multiple news reports surrounding a single event. While the minute details covered in each news report could vary, the overall theme to be covered in the output summary should likely be present in every source article. On the contrary, the Multi-XScience dataset has a main article, and the task is to summarize the other articles that the author of the main article made reference to into a related work section of a paper, This requires more compare and contrast having regard to the main theme as opposed to identifying the most common ideas.

document summarization tasks. The approach involves constructing a sentence-level pyramid structure and applying a masked language modeling objective to the pyramid, leading to improved results over BART, PEGASUS, and even LED.

Advancing the research on MDS further, Puduppully and Steedman [6] proposed a centroid-based pre-training method for multi-document summarization that leverages document-level clustering to capture document-level semantics. This approach has been shown to improve performance on summarization tasks over previous models including PEGASUS, particularly for MDS datasets with a large number of documents.

On the dataset front, a commonly used dataset in the MDS domain is perhaps the Multi-news dataset, which was introduced by Fabbri et al. [2] and contains news articles from multiple sources. Other MDS datasets include Wikisum, which was generated by Liu et al. [4] from a large-scale collection of Wikipedia articles, WCEP compiled by Ghalandari at el. [3] based on news summaries from the Wikipedia Current Events Portal, and the Multi-XScience dataset by Lu et al. [5] used in this study.

Overall, these recent developments highlight the ongoing efforts to improve the performance of transformer-based models for MDS, through approaches such as sparse attention mechanisms, hierarchical sentence representations, and document-level clustering.

3 Methods

3.1 Dataset

We used the Multi-XScience MDS dataset [5] in this study, which is a collection of scientific articles from various fields, including computer science, physics, and biology. There are a total of 30369, 5066 and 5093 samples for the training, validation and test sample sets. Some key parameters of the dataset are set out as follows, while Appendix A summarizes other findings and visualizations from our EDA, as well as the data preprocessing steps adopted for the experiments.

	Training	Validation	Test
Sample size	30,369	5,066	5,093
Mean Input Token Length	899 565 82-4694	898 567 79-4183	886 547 131-6348
Samples with inputs > 1024 tokens	31.21%	30.71%	30.87%
Mean Label Token Length	142 58 24-735	141 58 25-324	142 58 26-418
Articles per sample	4.43 2.62 2-21	4.43 2.63 2-20	4.39 2.55 2-19

Table 1: Dataset description. The cells with three columns refer to mean/standard deviation/range

The dataset is chosen for two main reasons. First, we note that Multi-XScience is among the few datasets previously unseen by MDS models during pre-training[4]. Picking other datasets already seen by the models could pose difficulties in evaluation, not to mention that it would go against the overall goal to explore MDS model-building strategies for unseen data.

The second reason is that Multi-XScience is a more challenging dataset, and the task is actually more comparable to real-world data collation and summarization tasks. Many of the datasets (e.g. Multi-news, WCEP) are news-based, with the articles within a sample covering the same news event and key elements to be included in the reference summary appearing in multiple locations. The

[4] In particular, LED's pre-training data included the Multi-News dataset while Centrum's pre-training included both Multi-News and WikiSum.

Multi-XScience dataset however, expects the model to write the related works section of a journal paper based on (a) the abstract of that paper which serves to provide context; and (b) the abstracts of the other journal articles that the main paper referenced. While these additional abstracts are usually[5] closely related, they often concern different aspects or approaches to a problem, and the model is expected to be able to compare and contrast the similarities and differences. This inherent difficulty of the dataset also shows up empirically in the model runs[6] of the PRIMERA study.

As can be seen from the table above, the input length of the Multi X-Science dataset shows great variability, with some inputs so short that they can fit in more traditional summarization models, while others exceed even 4096, which is the max input length for Centrum. The same variability can also be observed in terms of the number of articles per sample (see Figure 1 below), with the two distributions showing a close-to-linear pattern. This provides an opportunity for us to use the Multi X-Science dataset to also explore possible differences in model performance between short and long samples (whether in terms of input length or number of articles).

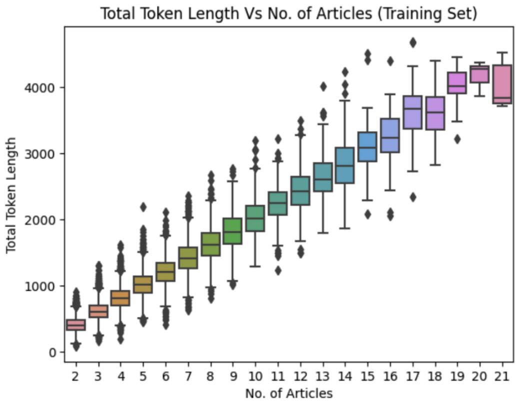

Figure 1: Total token length as a function of the number of articles in the training data set. The relationship between token length and the number of articles is linear.

3.2 Models

Due to resource constraints, we focused on two pre-trained LLMs, namely LED and Centrum for our experiments. Of the three recent and well-performing models we reviewed, we did not pick PRIMERA for two reasons - (a) Centrum is, based on results presented in its paper, likely the best-performing model of the three while LED is a common starting point for both PRIMERA and Centrum and it would be of benefit to treat it as a more advanced baseline; and (b) the authors of PRIMERA already provided performance results of PRIMERA on the Multi-XScience test set under different scenarios.

We adopted two different baselines for performance comparison. The first one, dubbed the "baseline model" is simply a lead-based model, whereby the first 3 sentences are chosen as the summary and is a common primary strategy implemented in various papers (e.g. [5], [7] and [9]). As we will discuss in the evaluation session, this also provides a useful baseline for comparing the degree of "copying" from the main article which is an undesirable trait for the Multi-XScience dataset. The second baseline is simply the publicly available checkpoint for LED, which serves as

[5] This, however, is not always the case. For example, sample 845 of the test set concerns the case of a journal article doing MDS over online product reviews, and one of the references appears to be completely unrelated to the topic and instead focuses on the physical design attributes of luggage carriers. This example is also available in the Appendix containing the qualitative summary (i.e. Appendix C)

[6] For instance, the zero- and few-short evaluation of BART, PEGASUS, LED and PRIMERA yielded a ROUGE-2 score of only 1.9 to 4.6 and a ROUGE-L score of 9.9 to 15.7 for Multi-XScience, while the score ranges for Multi-news is 3.7 to 13.6 for ROUGE-2 and 10.4 to 20.8 for ROUGE-L. A similar gap between results for the two datasets persists even after fine-tuning of the models.

the base architecture for recent MDS LLMs.

On top of the 2 baseline models, we ran experiments on the performance of the following models, with details on their fine-tuning and inference settings provided in Appendix :

1. a version of LED that was fine-tuned on the arXiv single document summarization task
2. the publicly available checkpoint of Centrum;
3. a version of the baseline LED[7] that we fine-tuned on the Multi-XScience training set;
4. a similarly fine-tuned version of Centrum; and
5. a two-step model stacking the fine-tuned LED and Centruml.

Of these experiments, models (c) and (d) stem from our wish to test out the performance of fine-tuning strategies in the current MDS context. During the fine-tuning process, we noticed that despite the larger input lengths allowed for in Centrum (4096 tokens), there are still limitations in allowing the model to access all information contained in the input as the Multi-XScience dataset can have up to ∼6 300 tokens. We, therefore, explored under model (e) a two-step model stacking process whereby the individual source articles are first condensed by our fine-tuned LED as a single document summarization task followed by an MDS process by (a) the fine-tuned LED or (b) the fine-tuned Centrum. The idea is to see if the Centrum can perform better if it is presented with more concise sources and without token truncation, see Figure 2. Refer to Appendix B for more details on the two step model explanation and observations.

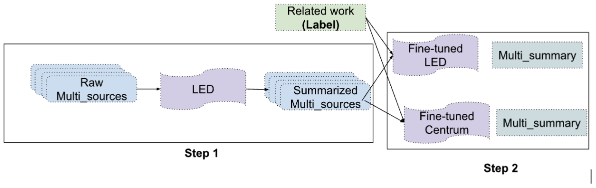

Figure 2: Two-steps model design

3.3 Evaluation metrics

As in the standard practice for summarization tasks, the primary quantitative metric we use is the ROUGE score. Among the different versions of ROUGE, we chose ROUGE-2 and ROUGE-L, , with a particular focus on the latter given the prevalence of compound nouns in academic writings[8], for gauging model performance. Compared to ROUGE-2, ROUGE-L is better suited for this task because it can capture the extent to which the generated summary includes important content and maintains the same order of important words and phrases as the reference summaries. In calculating these metrics, the precision, recall, and f-measure scores are

[7] In theory, it would be better to fine-tune the alternative LED checkpoint which (a) is a larger version of LED with ∼512M parameters as opposed to ∼200M for the basic version; and (b) has been fine-tuned on text from the scientific fields already. However, we only had a laptop RTX 3070Ti GPU with 8GB of VRAM at the time of fine-tuning and was unable to work on the larger model while maintaining the max input token length at a reasonable level, despite implementing strategies such as gradient accumulation and checkpointing as well as cutting floating point precision in half (i.e. fp16 vs fp32).

[8] This is an empirical observation gained from viewing samples in the dataset. That said, this is quite intuitive as well considering phrases in academic settings are often specialized terms glued together from multiple words (e.g. the theme of "Synchronous Optical Network Ring Assignment Problem" found in sample 4820 of the test set, and "fixed-size ordinally forgetting encoding" found in sample 3157 also from the test set). Measuring 2-grams when key phrases of the articles are often much longer may be of a more limited relevance in our context.

all considered[9], and scores were computed for the overall test dataset, as well as the subsets of short, medium, and long token length samples[10].

Unlike other MDS datasets where the key information is expected to be found in most, if not all, of the source documents, the Multi X-Science dataset has a different structure whereby the main article (i.e. first source) is expected to provide general context while the other sources will each provide a related, yet different angle of the topic. We, therefore, would like to avoid the situation of the model extracting information only from the main article, and measuring the degree of "copying" by repurposing ROUGE calculations and having it compare the words between the generated summary and the main article[11].

Finally, we supplemented our evaluation with a qualitative analysis whereby 25 random samples were selected and we reviewed the summaries written by both baselines and all 7 models, judging them based on their fluency, ability to extract information from multiple sources, and contrast information from the main article and other sources, as well as factual accuracy.

4 Results and Discussion

4.1 Overall performance

In the following table, we report the ROUGE-2 (R-2) and ROUGE-L (R-L) scores of the 2 baselines plus 5 models we tested, with precision (P), recall (R), and f-measure (F) scores presented. Scores for overall, short, medium, and long samples are also presented.

As demonstrated in the tables above, the fine-tuned models (LED and Centrum) provided a markedly improved performance over the baselines and off-the-shelf models. In fact, even though the models were only fine-tuned for up to 2 epochs[12], our figures quickly approached the SOTA figures reported in [8] (i.e. ROUGE-2 of 6.8; ROUGE-L of 18.2) demonstrating the viability of LED and Centrum to be adopted for MDS tasks. When considering the breakdown figures, the improvement for longer samples showed a much larger increment as compared to shorter samples, indicating the MDS-specific model architectures are perhaps more suitable for tasks with longer input lengths.

This result is consistent with what we observed in our qualitative analysis (Appendix C). For example, in sample 485 where the Centrum model was able to extract the key ideas of 3 relevant studies and contrast them with the main article, forming a coherent summary while the baseline and off-the-shelf models were overwhelmed by the large number of articles (9 in total) and choose only to copy one of the articles in full. This is very different from shorter samples, e.g. sample 4160 with only 3 articles, where the off-the-shelf LED and Centrum were still able to prepare a summary covering elements, albeit in a more extractive manner without efforts to change the tone or merge the extracted parts together.

[9] We also look only at the mean score but not the lower and upper ranges due to the limited scope of our study, though we recognize that looking at the variability in performance across samples is in itself a valid future investigation direction.

[10] We divided the 5 093 test samples into three groups. The "short" samples consisted of 1273 samples with less than 486 tokens (lower quartile) for the inputs; "medium" with 2546 samples with input lengths between 486 and 735 (upper quartile) tokens; and "long" samples with 1274 samples with input lengths greater than 1150 tokens.

[11] This is easily implemented using the rouge function from the HuggingFace. One can simply replace the target reference that the function takes in with the main article fed as input to the model.

[12] Training details and considerations are at Appendix B.

1	1
2	2

Table 2: Rouge-2 Scores for Models Tested (Best in green; worst in red)

1	1
2	2

Table 3: Rouge-L Scores for Models Tested (Best in green; worst in red)

For the two-step model, the ROUGE scores do not show any advantage of the approach over simple fine-tuned models. However, as we will observe from the qualitative analysis (e.g. samples 831 and 5068), there is a strong tendency for the models to write in a MDS manner by highlighting differences and similarities between the main and other articles. However, the downside is that the two-step model made quite a number of factual fallacies or even hallucinations in the process[13].

4.2 Degree of Copying

In order to assess the degree of copying from the main article, we calculated a repurposed ROUGE-L score, which is an undesirable trait for the Multi-XScience dataset. As shown in the table below, the fine-tuned and two-step models exhibited a significantly lower degree of copying from the main article, which is further supported by the qualitative analysis showing that these models were able to extract information from multiple articles. In contrast, the baseline LED model showed the highest degree of copying, often copying substantial portions of the main article to the point that the number of copied sentences exceeded even that of the lead-based baseline.

1	1
2	2

Table 4: Rouge-L Scores for Measuring "Copying" (Lower is better; best in green; worst in red)

Although the ROUGE scores did not reveal any significant copying for the off-the-shelf LED and Centrum models, it is noteworthy that these models occasionally copied from a single reference article rather than the main article (e.g., as observed in the output of the base Centrum for sample 4371). This indicates that there is still room for improvement in the "copying" metric, such as by taking the minimum of the ROUGE-L scores compared to each input article (rather than just the main article) to discourage copying from a single source.

4.3 Other observations

The qualitative analysis also allowed us to make the following observations on the models, including:

1. The fine-tuned and two-step models showed clear signs of learning common writing styles (using phrases such as "there is a large body of work...") which is desirable in real world applications;

2. There are cases where the sample contained irrelevant articles and the fine-tuned models were able to safely navigate away from them (e.g. sample 845)

3. Despite being trained on the Multi-XScience dataset, the models showed greatest difficulties when faced with highly technical papers that use common words in uncommon meanings (e.g. in mathematics in sample 4858). This

[13] One possible reason is that we need better tuning on the amount and manner of information fed into the 2nd step model, but we have not been able to carry out further explorations due to time limitations. The generation of the 1st step answers alone took around 14 hours.

shows the potential to enhance performance by adding another fine-tuning step on domain-specific texts. Alternatively, the two-step model has continued to perform well in such cases which shows the potential of such an approach.

5 Conclusion

In conclusion, we demonstrated the viability of building custom-based MDS models through fine-tuning which is able to quickly approach SOTA scores even with simple and limited tuning. Through breakdown and qualitative analyses of the results, we highlighted the importance of considering the token length of input texts when selecting a text summarization model. Our results suggest that the fine-tuned LED and Centrum models are much more adept at MDS for longer texts, while the two-step model shows potential especially for highly technical texts despite needing more tuning. Future work could focus on further optimizing the tuning techniques (e.g. fine-tuning with domain-specific texts first before the MDS data), improving on the two-step approach, as well as looking at the performance stability of the models across data samples.

Appendix

A EDA and Data Preprocessing

This appendix provides further findings from the EDA on the Multi-XScience dataset, as well as the data preprocessing procedures undertaken in the experiments.

As mentioned above, one standard summary could be summarized from more than 20 source articles. With concatenating all sources of articles, the total token length is about 4096. In addition, we've examined the token length of these standard summaries (label) and provided the details below.

To determine the input length, we look at the average length of the documents we want to summarize. With the consideration of the various numbers of source documents and our GPU capabilities, we decided to take input numbers of tokens = 512 as an experiment and numbers of tokens = 4096 (max tokens in the training dataset) as our final model.

For the output length, considering the desired level of detail in the summary and the distribution of standard summaries (label) tokens, we decided a summary length between 100-250 may be sufficient.

Multi-XScience includes a main abstract, which is from the main article; a standard summary of the related works with a separation of "@cite", which is the summary of multi-source documentations, label datasets; and reference abstracts, which are the multiple abstracts from a different source of works. We processed the dataset by defining the separator of documents as "|||||", a document separator, then concatenating reference abstracts using the document separator to replace citation references.

B Model

B.1 Baseline Model

Copying the first 3 sentences from each reference abstract was used as our baseline. It is a simple and easy-to-implement approach that provides a starting point for summarizing multiple documents. It assumes that the first few sentences of a document contain important information that should be included in the summary.

This approach can serve as a useful baseline for several reasons:

1. It is easy to implement and does not require a lot of computational resources. This makes it a good starting point for developing more advanced summarization models.

2. It is straightforward to understand and explain. This makes it a good choice for initial experiments and evaluations.

3. It can provide a quick estimate of the performance of a summarization system. By comparing the summaries generated by this approach to human-written summaries or other machine-generated summaries, it is possible to get a rough idea of how well the system is performing.

B.2 Baseline Base LED (1K and 16K) Model

The base LED model has ("allenai/led-base-16384"), with max input token length of 16384 & ∼200M parameters and consists of 12 transformer encoder layers and 2 transformer decoder layers. This model is designed to be computationally efficient and suitable for low-resource environments. The baseline LED was generated based on a pre-trained

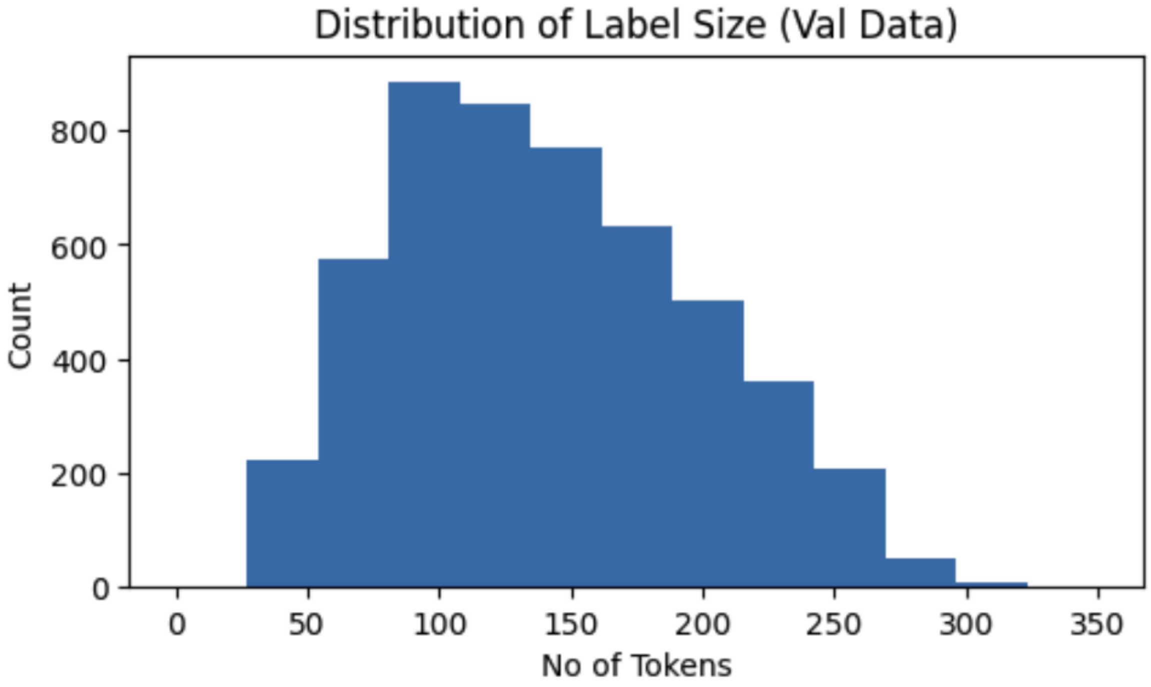

Figure 3: Distribution of the number of tokens in the label of training data. The number of tokens is right-skewed and centered around 100.

LED model by Beltagy, L (2020) [1] on X-science test datasets with input tokens varying from 1k and 6k.

B.3 Baseline Large LED (1K and 16K) Model

The "large" in the model name indicates that it has more parameters and is larger than the "base" LED model. The model has ("allenai/led-large-16384-arxiv"), with max input token length of 16384 & ∼512M parameters, and consists of 24 transformer encoder layers and 4 transformer decoder layers. This model has been pre-trained for general summarization tasks. The baseline LED was generated based on a pre-trained large LED and evaluated on X-science test datasets.

B.4 Baseline Centrum Model

The "ratishsp/Centrum" model with max input token length of 4096 & ∼192M parameters and consists of 12 transformer encoder layers and 12 transformer decoder layers. It is smaller than some of the larger transformer models, such as T5 and GPT-3, but larger than some of the smaller transformer models, such as the BART base. The publicly available Centrum checkpoint was built upon the LED model architecture using the 4096 token version.This means Centrum is not able to take in 16384 tokens.

B.5 Fine-tuned LED Model

Fine-tuning the "LED-large-16384-arxiv" model to the X-science train (size 30369) and validation (size 5066) datasets specifically for the related work summarization. We used 4096 max number of tokens in tracking and validation sets, and 256 max output tokens with a batch size of 2 (due to limited GPU resources). The model is then used as a cross-entropy loss function for 2 epochs, and evaluated on test datasets(size 5093) with ROUGH metrics.

B.6 Fine-tuned Centrum Model

Fine-tuning the "ratishsp/Centrum" model to the X-science train and validation datasets specifically for the related works summarization. We used 4096 max number of tokens in tracking and validation sets, and 256 max output tokens with a batch size of 16 (due to limited GPU resources). The model is then used as a cross-entropy loss function for 2 epochs, and evaluated on test datasets with ROUGH metrics.

B.7 Two-step LED Model

In the first step, the LED model is used to generate a summary of each related source document. By summarizing each source document, the model can condense the information and reduce redundancy, making it easier to process and analyze. For the first step, none of the individual source articles in the X-science dataset contain more than 4096 tokens, so there is no issue of the first step model (i.e. the fine-tuned LED) receiving truncated inputs.

In the second step, generate the summary based on the fine-tuned LED model to refine the summaries generated in the first step. In this way, we were hoping that the fine-tuned LED model can learn to generate even more accurate and relevant summaries.

B.8 Two-step Centrum Model

The same first step as the two-step LED Model above. In the second step, generate the summary based on the fine-tuned Centrum model to refine the summaries generated in the first step. In this way, we were hoping that the fine-tuned Centrum model can learn to generate even more accurate and relevant summaries.

B.9 Two-step Model results

However, from our evaluation and observation of this approach, the performance of the two-step model does not exceed the fine-tuned LED/Centrum model. One potential issue is that the large amount of information contained in the source documents may make it difficult for the model to generate concise and informative summaries. Ultimately, the performance is heavily dependent on the performance of step 1.

Overall, our results demonstrate the importance of utilizing advanced techniques like fine-tuning to unlock the full potential of language models in real-world applications.

C Qualitative Analysis

We also investigated how the model performed when summarizing short, medium, and long documents. The study aimed to evaluate the quality of the generated summaries and identify any trends or patterns in the model's performance based on the length of the source documents. A total of 25 randomly drawn samples (from the 5093 samples in the X-Science test dataset) are analyzed. These 25 samples include:

- 10 short samples, i.e. with total token lengths below the lower quartile
- 10 medium samples, i.e. with total token lengths between the lower and upper quartiles
- 5 long samples, i.e. with token lengths above the upper quartile

Our analysis of the review (refer to the Analysis Appendix for more detailed review) results showed that the fine-tuned LED and Centrum learned the structure for multi-documents summary for X-science dataset, with particularly good results for longer documents (ref. Sample 485) while for shorter samples the advantage over the baseline and off-the-shelf models appear to be less as those models are able to extract information from multiple sources to some degree, albeit with limited ability to merge them into a single coherent summary (ref. Sample 4160).

Additionally, we observed that LED and Centrum seem to have difficulties in digesting math topics, where the use of common language for unusual meanings might have confused the models (refer to no 4858). The two-step model somehow helps for these samples, but the downside is that the two-step model hallucinates quite a bit. In general, LED and Centrum never really hallucinate. To future improve this issue, we would suggest that further research in this area should consider:

- Add domain-specific knowledge: This can include adding specialized math dictionaries or knowledge bases to your training data.

- Use special tokens for math symbols: This can help the model recognize and differentiate between mathematical expressions and regular language. For example, use a special token such as “[math]” to indicate the start of a math expression and “[/math]” to indicate the end.

- Use more training data: Consider adding more math-specific data to the training set to help the model learn to recognize and understand math concepts.

- Fine-tune your model: fine-tuning it on a smaller dataset of math-specific documents. This can help the model learn to recognize and summarize math-related information more effectively.

Overall, our study highlights the importance of considering the length of source documents, and the training datasets topics when evaluating the performance of multi-document summarization models.

Future research in this area should focus on developing models that can effectively summarize long documents; and provide the model with more specialized knowledge to help the model recognize and differentiate between regular language and math expressions, improving its ability to summarize math-related documents effectively. Refer to the manual review documentation in more detail link.

References

[1] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

[2] Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy, July 2019. Association for Computational Linguistics.

[3] Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. A large-scale multi-document summarization dataset from the wikipedia current events portal, 2020.

[4] Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences, 2018.

[5] Yao Lu, Yue Dong, and Laurent Charlin. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074, Online, November 2020. Association for Computational Linguistics.

[6] Ratish Puduppully and Mark Steedman. Multi-document summarization with centroid-based pretraining, 2022.

[7] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks, 2017.

[8] Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. Primera: Pyramid-based masked sentence pre-training for multi-document summarization, 2022.

[9] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2020.