Improving Multi-Stage Long Document Summarization with Enhanced Coarse Summarizer

Jinhyeong Lim and Hyun-Je Song
Department of Computer Science and Artificial Intelligence
Jeonbuk National University
Jeonju, 54896, Korea
{dlawlsgud, hyunje.song}@jbnu.ac.kr

Abstract

Multi-stage long document summarization, which splits a long document as multiple segments and each of which is used to generate a coarse summary in multiple stage, and then the final summary is produced using the last coarse summary, is a flexible approach to capture salient information from the long document. Even if the coarse summary affects the final summary, however, the coarse summarizer in the existing multi-stage summarization is coarsely trained using data segments that are not useful to generate the final summary. In this paper, we propose a novel method for multi-stage long document summarization. The proposed method first generates new segment pairs, ensuring that all of them are relevant to generating the final summary. We then incorporate contrastive learning into the training of the coarse summarizer, which tries to maximize the similarities between source segments and the target summary during training. Through extensive experiments on six long document summarization datasets, we demonstrate that our proposed method not only enhances the existing multi-stage long document summarization approach, but also achieves performance comparable to state-of-the-art methods, including those utilizing large language models for long document summarization.

1 Introduction

Long document summarization aims to compress a long document, such as meeting minutes, reports, and scientific articles, into a concise text that captures salient information. Since the number of tokens in a long document usually exceeds the limit of the summarization models, various summarization approaches (Mao et al., 2022; Beltagy et al., 2020; Tay et al., 2020; Rohde et al., 2021; Pu et al., 2023a; Xie et al., 2022) to deal with long document has proposed. One promising approach among the long document summarization approaches is the multi-stage split-then-summarization approach (Zhang et al., 2022). It first splits the long document into source segments and each of which is used to generate a coarse summary. After splitting into source segments and generating a coarse summary in multiple stages, it produces the final summary using the last coarse summaries. This approach offers flexibility in processing documents of arbitrary length by adjusting the number of stages and has achieved state-of-the-art performance on several long document summarization benchmark datasets. However, there is still room for improvement in terms of the quality of the coarse summary.

It should be noted that the quality of the final summary depends on the quality of the coarse summaries. In the previous multi-stage split-then-summarization approach, Zhang et al. (2022) construct the training data using pairs of long documents and target summaries for the coarse summarizer because there is no official data for the coarse summarizer. That is, it first splits the long document and the target summary into segments, respectively. Each document segment is then aligned with a subset of the target segments, maximizing the ROUGE-1 score between the document segment and the subset. The resulting aligned pairs are used to train the coarse summarizer. Even though some document segments may be irrelevant to the target summary, aligned pairs derived from these segments are employed to train the coarse summarizer. It results in the generation of coarse summaries that may contain some noise. Figure 1 shows an example of the construction of the training data for the coarse summarizer, where the dotted line indicates the alignment of one document segment with a subset of the target segments. As all document segments are aligned and used to train the coarse summarizer, the coarse summarizer may generate low quality and excessively verbose summaries.

In this paper, we propose a novel approach suitable for multi-stage summarization. The proposed

Proceedings of The 4th New Frontiers in Summarization Workshop, pages 135–144

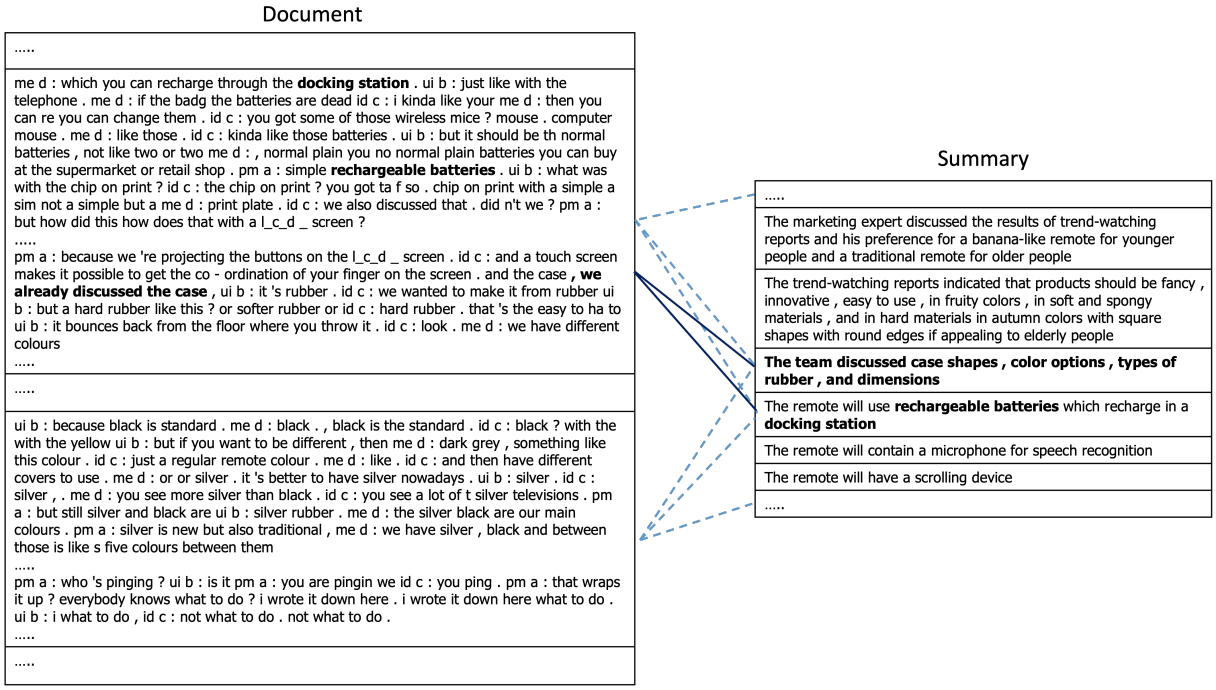

Figure 1: An example from AMI dataset to show an alignment between document segments and target segments. Dotted lines indicate the source-focused aligned pairs, while bold lines represent the target-focused aligned pairs.

method first generates aligned pairs that are relevant to the generation of the final summary. Unlike the previous alignment discussed, the proposed method executes an alignment in a reverse direction. Specifically, the proposed method aligns each target segment with a document segment. By aligning in this reverse direction, we ensure that only relevant document segments are included in the training data for the coarse summarizer. Furthermore, the proposed method incorporates a sub-summary generation contrastive objective (Liu et al., 2021) during training of the coarse summarizer to explicitly model the similarity between the target segment and the document segment. This addition of contrastive objective encourages the coarse summarizer to focus on relevant document segments and target segments and contributes to further improving the summarization quality of the coarse summarizer.

We conduct extensive experiments on six long document summarization datasets to show the superiority of the proposed method. We also compare the proposed method with large large language models-based long document summarization. Experimental results imply that the proposed method contributes to enhancing the efficiency and effectiveness of the multi-stage long document summarization approach compared to the existing method[1].

2 Multi-Stage Long Document Summarization

Let $\mathcal{D} = \{(S_i, T_i)\}_{i=1}^{N}$ be a set of document-target summary pairs, where S_i is the i-th long document and T_i is its corresponding summary. The multi-stage long document summarization approach segments the long document and then summarizes the segmented text in multiple stages because the number of tokens in S_i exceeds the limit of the summarizer. It consists of two stages: C coarse stages and one fine-grained stage. In each coarse stage, an input document is divided into document segments, and then a coarse summarizer generates coarse summaries from the document segments. In the fine-grained stage, a fine-grained summarizer generates the final summary from the last generated coarse summary.

Let K be the number of the maximum input tokens of the summarizer. S_i is divided into multiple segments, each with a length of fewer than K tokens. That is, $S_i = \{s_{i1}, \ldots, s_{in_i}\}$, where n_i is the number of segments of S_i. Similarly, the target summary $T_i = \{t_{i1}, \ldots, t_{im_i}\}$ is also divided into multiple segments, usually split into separate sen-

[1]The proposed method is publicly available at https://github.com/Jinhyeong-Lim/Summ-N-ECS

tences. That is, t_{ij} and m_i are the j-th sentence and the number of sentences in T_i, respectively. To generate training data for the coarse summarizer, Zhang et al. (2022) adopt the ROUGE-based greedy target alignment function, aligning each document segment s_{ij} with a subset of T_i such that the ROUGE-1 score between s_{ij} and the subset is maximized. The training data for the coarse summarizer in the k-th stage is constructed as follows:

$$\mathcal{D}^k_{coarse-s} = \cup_{i=1}^{N} \cup_{j=1}^{n_i} \text{align}_{source}(s_{ij}, T_i).$$

Here, $\text{align}_{source}(\cdot,\cdot)$ is a function to align each document segment with a subset of target segments. Since the alignment is executed for each source segment, all document segments are contained in $\mathcal{D}^k_{coarse-s}$. The k-th coarse summarizer is trained with $\mathcal{D}^k_{coarse-s}$ to minimize the negative log-likelihood (NLL) $\mathcal{L}_{nll}$ between the word distributions predicted by the summarizer and the target segments.

After training the coarse summarizer, a coarse summary is obtained using the trained summarizer from each document segment s_{ij}. All n_i coarse summaries are then concatenated to form a new input for the next stage. The target summary for the next stage is copied from the original target summary. It is worth noting that the number of coarse stages is estimated based on the length of the long document and the characteristics of the summarizer. Further details can be found in Zhang et al. (2022) and Section 4.3.

In the fine-grained stage, the coarse summaries from the C-th coarse stage are concatenated and used as input for the fine-grained stage. Since the number of tokens in the input is shorter than K, a fine-grained summarizer can be modeled similarly to a well-known vanilla abstractive summarizer (Lewis et al., 2020; Zhang et al., 2020). This means that the fine-grained summarizer is trained on the dataset from the last coarse stage and produces the final summary using the last coarse summaries.

3 Improving Multi-Stage Summarization with Enhanced Coarse Summarizer

This paper presents a new approach to multi-stage summarization. The proposed method generates new aligned pairs that include only relevant document segments because the greedy target alignment function (Zhang et al., 2022) generates some pairs that contain irrelevant document segments. In addition, the proposed method incorporates a contrastive learning into the training of the coarse summarizer that a document segment and the corresponding target summary should convey the same meaning, which is not modeled explicitly by the NLL loss (Xu et al., 2022).

3.1 Target-focused Aligned Pairs

To generate aligned pairs that includes the relevant document segments, the proposed method designs a new alignment function. The proposed alignment function focuses on the target segment that each target segment t_{ij} is aligned with a document segment s_{il}, maximizing the ROUGE scores between between t_{ij} and s_{il} (Bold lines in Figure 1. The training data for the coarse summarizer in the k-th stage can be constructed as follows:

$$\mathcal{D}^k_{coarse-t} = \cup_{i=1}^{N} \cup_{j=1}^{m_i} \text{align}_{target}(t_{ij}, S_i), \quad (1)$$

where $\text{align}_{target}(\cdot,\cdot)$ is a function to align each target segment with a document segment. This alignment ensures that irrelevant document segments are not included in the training data.

The training data constructed by the new alignment function, however, has one problem that there is one-to-many mappings in pairs of document segment-target segment, which is also known as multi-modality problem (Gu et al., 2017; Wei et al., 2019). That is, one document segment is mapped to multiple target segments. This can result in the generation of low quality of coarse summaries. To alleviate this problem, the proposed method merges the multi-modal data by concatenating target segments. For example, if the proposed alignment function generates pairs $\{(s_{i1}, t_{i1}), (s_{i1}, t_{i4}), (s_{i1}, t_{i6})\}$, these three pairs are merged into one pair by concatenating three target segments such that $\{(s_{i1}, t_{i1} \oplus t_{i4} \oplus t_{i6})\}$, where $\oplus$ is a string concatenate operator.

3.2 Contrastive Learning with Sub-summary Generation Objective

The summary of a long document comprises multiple sentences, each of which can be seen as a sub-summary. Given that a single long document may encompass multiple subjects, we can consider the coarse summarizer as mapping each subject to its corresponding sub-summary, and these subjects are inherently present within the document's segments. To achieve improved mapping, the proposed method employs contrastive learning with a sub-summary generation objective (Liu et al., 2021).

Table 1: The statistics of data set used in experiments. The document length and summary length are the averaged numbers across the dataset.

Dataset	Size ($\|\mathcal{D}\|$)	Document length	Summary length	Type	Domain
AMI	137	6007.7	296.6	Dialogue	Meeting
ICSI	59	13317.2	488.5	Dialogue	Meeting
QMSum	1808	9069.8	69.6	Dialogue	Meeting
SummScreen-FD	4348	7605.4	113.7	Dialogue	TV Series
SummScreen-TMS	22503	6420.7	380.6	Dialogue	TV Series
GovReport	19466	9409.4	553.4	Document	Reports

To conduct contrastive learning, the proposed method initially constructs contrastive sub-summary generation pairs, comprising both positive and corresponding negative examples. Positive examples are obtained from data pairs as defined by Equation (1) while their corresponding negative examples are derived by substituting document segments with alternative ones. Here, the method selects the document segment that exhibits the lowest ROUGE score in comparison to the target segment.

Let $\{(s^k_{pos}, t^k), (s^k_{neg}, t^k)\}_{k=1}^{N*m_i}$ be a constructed the contrastive sub-summary generation pairs, where s^k_* and t^k are segments of document and target, respectively. With the contrastive pairs, the proposed method calculates the NLL values as follows:

$$L^{t^k}_{pos} = -\log \Pi_{j=1}^{|t^k|} p(t^k_j \mid t^k_{i:j-1}, s^k_{pos}),$$
$$L^{t^k}_{neg} = -\log \Pi_{j=1}^{|t^k|} p(t^k_j \mid t^k_{i:j-1}, s^k_{neg}),$$

where $L^{t_i}_{pos}$ and $L^{t_i}_{pos}$ are the negative log likelihood values of the positive example and negative example, respectively. t^k_j is the j-th token in t^k. Then, the normalized score is obtained by applying the softmax function to the two NLL values:

$$su(s^k_{pos}), su(s^k_{neg}) = softmax([L^{t^k}_{pos}, L^{t^k}_{neg}]),$$

where $su(s^k_{pos})$ and $su(s^k_{neg})$ represent the normalized scores of the positive example and negative example, respectively, indicating their relative relevance scores.

Then, the sub-summary generation contrastive objective, denoted as $\mathcal{L}_{ctr}$, is defined as follows:

$$\mathcal{L}_{ctr} = \frac{1}{N*m_i} \sum_{k=1}^{N*m_i} \max(0, \delta - (su(s^k_{neg}) - su(s^k_{pos}))),$$

where δ is a margin that the relevance score between a positive document segment and a target segment to be at least larger than the relevance score of the negative example. The δ is set as 1.

The final loss for the coarse summarizer is defined as

$$\mathcal{L}_{coarse} = \lambda * \mathcal{L}_{ctr} + \mathcal{L}_{NLL}, \quad (2)$$

where λ is a hyperparameter to balance the contrastive loss.

4 Experiments

4.1 Experimental Settings

This paper conducts the experiments on six long document summarization datasets: AMI (Carletta et al., 2006), ICSI (Janin et al., 2003), QMSum (Zhong et al., 2021), SummScreen-FD, SummScreen-TMS (Chen et al., 2022), and GovReport (Huang et al., 2021). Table 1 shows data statistics of the datasets. For the query-focused document summarization task using the QMSum dataset, we follow the settings of Zhang et al. (2022) such that the query is concatenated into the beginning of the document at both the training and the inference time.

The proposed method follows the experimental settings of Zhang et al. (2022) except the backbone summarization models. The proposed method adopts the BART-large model, pretrained on the CNN/DM dataset, as the backbone model for the fine-grained summarizer. The backbone model for coarse summarizer in the GovReport dataset is set to BART-large while for other datasets, it is set to DialogLM (Zhong et al., 2022).

All experiments are conducted on one NVIDIA RTX A6000 GPU with a 48 GiB memory. The multi-stage long document summarization with the proposed coarse summarizer is trained for 30 epochs with a batch size of 4. We set the optimizer to AdamW and the learning rate is set to 2e-5 in the coarse stage and 3e-5 in the fine-grained stage. The best checkpoint is chosen by early stopping based

Table 2: Performances of long document summarization on six long document summarization data sets. * denotes the ROUGE-L scores without the sentence split. Best performance is in bold, and second best is underlined.

	AMI			ICSI			QMSum-All		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
HMNET (Zhu et al., 2020)	53.02	18.57	24.85*	46.28	10.60	19.12*	-	-	-
DDAMS (Feng et al., 2021)	53.15	**22.32**	25.67*	40.41	11.02	19.18*	-	-	-
UniLM-CP(Dong et al., 2019)	52.67	19.33	50.55	48.43	12.39	46.24	29.19	6.73	25.52
BART$_{Large}$-SLED (Ivgi et al., 2023)	-	-	-	-	-	-	34.20	**11.00**	22.00*
DYLE (Mao et al., 2022)	-	-	-	-	-	-	34.42	9.71	30.10
DialogLM (Zhong et al., 2022)	54.49	20.03	50.92	49.56	12.53	47.08	33.69	9.32	30.01
DialogLED (Zhong et al., 2022)	54.80	20.37	52.26	50.11	13.23	47.25	34.50	9.92	30.27
SUMM^N (Zhang et al., 2022)	53.44	20.30	51.39	45.57	11.49	43.32	34.03	9.28	29.48
Proposed model	**54.85**	21.18	**52.28**	**50.27**	**13.38**	**47.30**	**35.31**	10.13	**30.58**

	SummScreen-FD			SummScreen-TMS			GovReport		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
UniLM-CP (Dong et al., 2019)	33.29	6.74	28.21	44.07	9.96	41.73	-	-	-
TopDownFormer (Pang et al., 2023)	**36.84**	**9.19**	31.12	**51.02**	**14.66**	**49.01**	-	-	-
BART$_{Large}$-SLED (Ivgi et al., 2023)	-	-	-	-	-	-	57.50	26.30	27.40*
PageSum (Liu et al., 2022)	-	-	-	-	-	-	60.04	27.17	57.21
DYLE (Mao et al., 2022)	-	-	-	-	-	-	**61.01**	**28.83**	**57.82**
DialogLM (Zhong et al., 2022)	35.75	8.27	30.76	45.58	10.75	43.31	-	-	-
DialogLED (Zhong et al., 2022)	36.70	8.68	**31.38**	45.22	11.69	42.86	-	-	-
SUMM^N (Zhang et al., 2022)	32.48	5.85	27.55	44.64	11.87	42.53	56.77	23.25	53.90
Proposed model	36.81	9.07	31.21	45.81	11.97	43.35	58.01	25.66	55.30

on the highest average of ROUGE-1/2/L scores on the validation set. The λ in Equation (2) is set to 1.

The proposed method is compared with the previous state-of-the-art methods on the datasets. The performance is measured with ROUGE (Lin, 2004).

4.2 Experimental Results

Table 2 shows the ROUGE scores of the proposed model compared to the baselines. The proposed model outperforms SUMM^N on all datasets, which indicates that the enhanced coarse summarizer in the proposed method improves the multi-stage summarization approach. Furthermore, the proposed model achieves similarly or even better performance compared to other state-of-the-art models. These results demonstrate that utilizing the proposed coarse summarizer in the multi-stage summarization approach is a reasonable solution for long document summarization.

We investigate the effectiveness of target-focused alignment and contrastive learning in the proposed model. Table 3 shows the results of an ablation study. If the coarse summarizer is trained only using $\mathcal{D}^k_{coarse-t}$, the performance is dropped by 1.28 on the AMI dataset and 3.2 on the ICSI dataset compared to the proposed method.

Table 3: Ablation study on test set of AMI and ICSI datasets. Performance is the ROUGE-1 score.

	AMI	ICSI
$\mathcal{D}^k_{coarse-t}$ + contrastive learning	54.85	50.27
$\mathcal{D}^k_{coarse-t}$	53.57	47.07
$\mathcal{D}^k_{coarse-s}$ + contrastive learning	53.79	46.12
$\mathcal{D}^k_{coarse-s}$ (= SUMM^N)	53.44	45.57

However, its performance is better than the one of SUMM^N, which is trained using $\mathcal{D}^k_{coarse-s}$. This implies that the target-focused alignment and the contrastive learning helps improve multi-stage summarization.

4.3 Coarse Summary Analysis

To verify the quality of coarse summaries generated by the proposed method, we compare the coarse summaries at different stages with the target summary. Figure 2 depicts the ROUGE-1 scores of all datasets for each stage. Stage 1 represents the model with only one coarse stage and no fine-grained stage, while Stage i ($i > 1$) represents i - 1 coarse stages and one fine-grained stage. When comparing the ROUGE-1 scores of the proposed

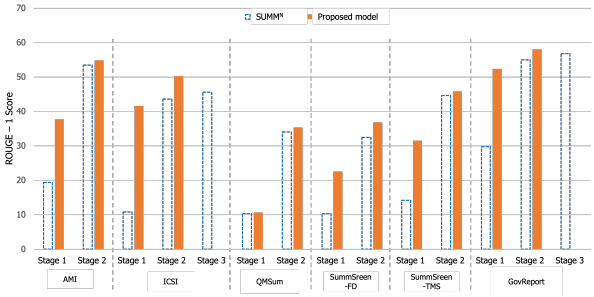

Figure 2: ROUGE-1 scores at different stages.

Table 4: Values for the coarse stage estimation

	ICSI	QMSum	GovReport
d_1	13317.2	9069.8	9409.4
K	1024	1024	1024
c_1 (Ours)	24.9	37.2	38.4
c_1 (SUMM^N)	276.8	68.4	243.2
$\hat{N}$ (Ours)	1 (0.7)	1 (0.7)	1 (0.6)
$\hat{N}$ (SUMM^N)	2 (1.9)	1 (0.9)	2 (1.5)

method with SUMM^N, the proposed method consistently outperforms SUMM^N at all stages. Notably, in the ICSI and GovReport datasets, the proposed method at Stage 2 achieves higher performance compared to SUMM^N at Stage 3. This indicates that our proposed method is more effective than SUMM^N.

We also investigate the effectiveness of the proposed method by estimating the number of coarse stages. In the multi-stage summarization, the fewer the coarse stages performed, the greater the effectiveness of the method. SUMM^N proposes a method to estimate the number of coarse stages. It is based on the length of the long document and the characteristics of the summarizer. The number of coarse stages is computed as follows:

$$\hat{N} = \left\lceil \frac{\log K - \log d_1}{\log c_1 - \log K} \right\rceil,$$

where d_1 and c_1 are the average length of document and coarse segments in coarse stage 1. K represents the maximum input tokens of the backbone model.

Table 4 shows the values used to estimate the coarse stage on three data sets. The estimated number of coarse stages for the proposed model on ICSI and GovReport is smaller than those of SUMM^N. This difference arises because the coarse summarizer in SUMM^N is trained using pairs of a document segment and a set of target segments, whereas the coarse summarizer in the proposed method is trained using pairs of a document segment and a target segment[2]. As a result, the coarse summarizer in the proposed method exhibits a tendency to generate more succinct coarse summaries, which in turn facilitates the generation of superior final summaries.

4.4 Human Evaluation

This paper conducts human evaluation with AMI and ICSI dataset to validate the quality of the generated summaries with respect to fluency and coverage. The experimental settings for the human evaluation follow those of Zhang et al. (2022). Specifically, the quality is assessed through three metrics: Readability, Conciseness, and Coverage. Readability takes into account word and grammatical error rate to assess the fluency of the summary. Conciseness measures how well the summary discards the redundant information, while Coverage gauges how well the summary covers each part of the dialogue.

We compare the results of the proposed method and SUMM^N because both methods are grounded in the multi-stage summarization. For source documents within AMI and ICSI datasets, three human annotators evaluate the quality of the summaries generated by each model. In this process, each annotator reviews the source document, the gold summary, and the generated summary, subsequently rating each summary from 1 to 5 (with higher scores indicating superior quality) across the aforementioned metrics.

Table 5 summarizes the performance of the proposed method and the baseline, SUMM^N. The proposed model outperforms SUMM^N in both the AMI and ICSI datasets. Specifically, the Coverage score of the proposed method is notably higher than that of SUMM^N. It seems that the coarse summarizer in the proposed method effectively generates concise summaries for each document segment, while the fine-grained summarizer captures the essence of each segment more comprehensively. The results of human evaluation experiments demonstrate that the proposed method generates better summaries than SUMM^N.

[2]For the sake of simplicity, there is no one-to-many mappings.

Table 5: Human evaluation scores.

	AMI			ICSI		
	Readability	Conciseness	Coverage	Readability	Conciseness	Coverage
SUMM^N	3.90	3.45	3.48	3.56	3.28	3.33
Proposed model	**4.15**	**3.58**	**3.78**	**3.77**	**3.55**	**4.17**

Table 6: Performances of LLM-based summarization on three long document summarization data sets.

	AMI			ICSI			QMSum-All		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
LongChat-7B-16K (Zero-shot)	26.15	6.77	24.50	18.06	2.07	17.12	23.40	3.86	20.78
Llama 2-13B (Zero-shot)	29.20	6.23	27.77	9.99	0.79	9.46	15.10	1.94	13.40
Llama 2-13B (Few-shot, Source-focused)	28.59	7.09	27.18	23.35	2.62	22.42	21.32	3.16	18.82
Llama 2-13B (Few-shot, Target-focused)	30.49	7.28	29.18	24.66	2.99	23.47	21.48	3.16	19.16
Proposed model	**54.85**	**21.18**	**52.28**	**50.27**	**13.38**	**47.30**	**35.31**	**10.13**	**30.58**

4.5 Comparison with LLM-based Long Document Summarization

Recent studies have shown that while summaries generated by large language models (LLMs) consistently outperform those of fine-tuned summarization methods in short document summarization (Zhang et al., 2023; Pu et al., 2023b), they typically fall short in the context of long document summarization (Yang et al., 2023; Nijkamp et al., 2023; Touvron et al., 2023). For example, ChatGPT achieved a score of 28.34 R-1 on the QMSum dataset (Yang et al., 2023), whose score is below the 35.31 R-1 obtained in our study. According to the study (Nijkamp et al., 2023), the performance of XGen-7B for GovReports was only at 21.28 R-1, which is substantially inferior compared to existing long document summarization methods. The recently introduced Llama 2 (Touvron et al., 2023) also encountered challenges in long document summarization, achieving only a 15.08 R-1 score on the QMSum dataset. Apart from LLMs specifically engineered to handle extended contexts, the majority of LLMs have input context length about 4K (e.g., Llama2-13B). This implies that even when using LLMs, a multi-stage summarization approach remains essential to process long documents.

To probe the efficacy of LLM-based summarization in our experiments, we conducted experiments using the LongChat (Li et al., 2023) and Llama 2 (Touvron et al., 2023) models and evaluated their performance in comparison with the proposed method. We employed two types of LLM-based summarization approaches: zero-shot summarization and few-shot multi-stage summarization. In zero-shot summarization, the model is provided with both a task prompt and the document as input, and then it generates a summary of the given document. In few-shot multi-stage summarization, we replace both the coarse and fine-grained summarizers in the multi-stage summarization with a few-shot LLM summarization model. Here, we also examined two types of few-shot LLMs for the coarse summarizer: a source-focused model, where the few-shot examples are sampled from source-focused pairs, and a target-focused model, where the examples are sampled from target-focused pairs.

Table 6 shows the performance of the LLM-based summarization across three long-document datasets. The efficacy of the summary model using LLMs is inferior to that of existing multi-stage summarization approaches. In other words, LLMs have not yet demonstrated strong performance in long document summarization, indicating an area for future research. Another interesting observation is that when using LLMs for multi-stage summarization, LLMs with few-shot examples from target-focused pairs achieve better performance than those with few-shot examples from source-focused pairs. These results suggest that target-focused pairs play an important role in improving performance in multi-stage summarization.

4.6 Case study

Table 7 shows sample summaries generated by the proposed method and SUMM^N. Both summaries

Table 7: Sample output summary of the proposed method on the AMI dataset.

SUMM^N	The project manager opens the meeting by **recapping the events of the previous meeting**. The marketing expert presents the results of market research , which shows that users want a fancy-looking remote control that is easy to use and has a **fancy look** and feel. **The user interface designer** presents the user interface concept for the remote , which is based on the idea that a **remote** should be simple and user-friendly. The industrial designer presents about the internal components of a remote control. The group discusses using kinetic **energy** to power the device , using a simple battery for the **LCD screen** , and using an advanced chip for the advanced **chip**. The project manager closes the meeting , telling the team members what their tasks will be for the next meeting. $\cdots$ The Marketing Expert will research how to produce a remote that is technologically innovative. The User Interface Designer will look at how to make a remote out of wood or plastic with either a wooden or plastic cover. The Group will not work with teletext. There was a lack of information on the cost of components and materials.
Proposed method	The project manager **recapped the decisions made in the previous meeting**. The marketing expert presented the results of a market study which shows that users of remotes want remotes to have a **fancy look** and feel, are interested in speech recognition, and are willing to pay extra for an **LCD screen**. The interface specialist presents, suggesting that the remote should be user-friendly and attractive, using large buttons, **speech recognition**, and using a little display. The industrial designer presents, presenting the components design. The group discusses the features they would like to include in the remote, including an LCD screen and a kinetic **energy** source. The user interface designer and industrial designer will work together on the look-and-feel design the group will use default materials. The remote will be single curved, single curved or double curved. The case will be made of plastic or rubber. The company will use **wood**. What type of **chip** to use. Whether to have a double curved or single curved case $\cdots$ the device will have a **docking station** for the remote to put the remote in when not in use. what sort of chip the device should have. What kind of display to include. What shape the remote is to be. Whether speech recognition is a good idea or not. Whether the remote has to be a changeable case. **Choosing between an LCD screen or speech recognition.**
Gold	The project manager opened the meeting and **recapped the decisions made in the previous meeting**. The marketing expert discussed his personal preferences for the design of the remote and presented the results of trend-watching reports , which indicated that there is a need for products which are **fancy** , innovative , easy to use , in dark colors , in recognizable shapes , and in a familiar material like wood. The user interface designer discussed the option to include speech recognition and which functions to include on the remote. The industrial designer discussed which options he preferred for the remote in terms of energy sources , casing , case supplements , buttons , and chips. The team then discussed and made decisions regarding energy sources , speech recognition , **LCD screens** , **chips** , case materials and colors , case shape and orientation , and button orientation. $\cdots$ The case covers will be available in wood or plastic. The case will be single curved. Whether to use kinetic energy or a conventional battery with a **docking station** which recharges the remote. Whether to implement an LCD screen on the remote. **Choosing between an LCD screen or speech recognition**. Using wood for the case

capture the content of the source long document and exhibit similarities to the gold summary. However, the summary from the proposed method contains more relevant phrases compared to the one generated by SUMM^N such as "recapped the decision made in the previous meeting" and "choosing between an LCD screen or speech recognition". We posit that the proposed summarizer has the capacity to generate concise coarse summaries, which subsequently facilitate the production of final summaries enriched with key phrases.

5 Conclusion

In this paper, we have demonstrated the effectiveness of enhancing the coarse summarizer in the multi-stage split-then-summarization approach. To enhance the coarser summarizer, the proposed method focuses on constructing target-focused aligned pairs, ensuring that only relevant source segments are included. Then, the proposed method incorporates contrastive learning into the training of the coarse summarizer to reinforce the conveyance of the similar meaning between source and target segments. Experimental results on six long document summarization datasets show that the proposed method helps improve the performance of the multi-stage long document summarization approach.

Acknowledgments

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Korean Government (MSIT) under Grant 2021R1F1A1048181, in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) funded by the Korean Government (MSIT) through Artificial Intelligence Innovation Hub under Grant 2021-0-02068.

Limitations

While the multi-stage split-then-summarization approach offers flexibility in processing long documents by adjusting the number of stages, it requires lots GPUs with large memory sizes. In our experiments, we used A6000 GPUs with a 48 GiB memory, and the training process took up to three days per dataset. It is important to note that the proposed model is based on the existing multi-stage long document summarization approach, which means that the memory usage during training is similar to the existing approach, and the footprint remains the same during inference.

Ethics Statement

We have conducted the proposed model training and testing using publicly accessible datasets. To the best of our knowledge, this work does not involve any ethical issues. We believe that this work complies with the ethical code of ACL.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The ami meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. *Unified Language Model Pre-Training for Natural Language Understanding and Generation*. Curran Associates Inc., Red Hook, NY, USA.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2021. Dialogue discourse-aware graph model and data augmentation for meeting summarization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3808–3814. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2017. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Maor Ivgi, Uri Shaham, and Jonathan Berant. 2023. Efficient Long-Text Understanding with Short-Text Models. *Transactions of the Association for Computational Linguistics*, 11:284–299.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph E. Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. 2023. How long can open-source llms truly promise on context length?

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Junpeng Liu, Yanyan Zou, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Caixia Yuan, and Xiaojie Wang. 2021. Topic-aware contrastive learning for abstractive dialogue summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1229–1243, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yixin Liu, Ansong Ni, Linyong Nan, Budhaditya Deb, Chenguang Zhu, Ahmed Hassan Awadallah, and Dragomir Radev. 2022. Leveraging locality in abstractive text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6081–6093, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ziming Mao, Chen Henry Wu, Ansong Ni, Yusen Zhang, Rui Zhang, Tao Yu, Budhaditya Deb, Chenguang Zhu, Ahmed Awadallah, and Dragomir Radev. 2022.

DYLE: Dynamic latent extraction for abstractive long-input summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1687–1698, Dublin, Ireland. Association for Computational Linguistics.

Erik Nijkamp, Tian Xie, Hiroaki Hayashi, Bo Pang, Congying Xia, Chen Xing, Jesse Vig, Semih Yavuz, Philippe Laban, Ben Krause, Senthil Purushwalkam, Tong Niu, Wojciech Kryscinski, Lidiya Murakhovs'ka, Prafulla Kumar Choubey, Alex Fabbri, Ye Liu, Rui Meng, Lifu Tu, Meghana Bhat, Chien-Sheng Wu, Silvio Savarese, Yingbo Zhou, Shafiq Rayhan Joty, and Caiming Xiong. 2023. Long sequence modeling with xgen: A 7b llm trained on 8k input sequence length. Salesforce AI Research Blog.

Bo Pang, Erik Nijkamp, Wojciech Kryscinski, Silvio Savarese, Yingbo Zhou, and Caiming Xiong. 2023. Long document summarization with top-down and bottom-up inference. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1267–1284, Dubrovnik, Croatia. Association for Computational Linguistics.

Dongqi Pu, Yifan Wang, and Vera Demberg. 2023a. Incorporating distributions of discourse structure for long document abstractive summarization.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023b. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Tobias Rohde, Xiaoxia Wu, and Yinhan Liu. 2021. Hierarchical learning for generation with long source sequences. *ArXiv*, abs/2104.07545.

Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. 2020. Sparse Sinkhorn attention. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9438–9447. PMLR.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Bolin Wei, Shuai Lu, Lili Mou, Hao Zhou, Pascal Poupart, Ge Li, and Zhi Jin. 2019. Why do neural dialog systems generate short and meaningless replies? a comparison between dialog and translation. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7290–7294. IEEE.

Qianqian Xie, Jimin Huang, Tulika Saha, and Sophia Ananiadou. 2022. GRETEL: Graph contrastive topic enhanced language model for long document extractive summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6259–6269, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shusheng Xu, Xingxing Zhang, Yi Wu, and Furu Wei. 2022. Sequence level contrastive learning for text summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11556–11565.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summ^n: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11765–11773.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online. Association for Computational Linguistics.