SUMM^N: A Multi-Stage Summarization Framework for Long Input Dialogues and Documents

Anonymous ACL submission

Abstract

Text summarization helps readers capture salient information from documents, news, interviews, and meetings. However, most state-of-the-art pretrained language models (LM) are unable to efficiently process long text for many summarization tasks. In this paper, we propose SUMM^N, a simple, flexible, and effective multi-stage framework for input texts that are longer than the maximum context length of typical pretrained LMs. SUMM^N first splits the data samples and generates a coarse summary in multiple stages and then produces the final fine-grained summary based on it. Our framework can process input text of arbitrary length by adjusting the number of stages, while keeping the LM input size fixed. Moreover, it can deal with both single-source documents and dialogues, and it can be used on top of different backbone abstractive summarization models. To the best of our knowledge, SUMM^N is the first multi-stage split-then-summarize framework for long input summarization. Our experiments demonstrate that SUMM^N outperforms previous state-of-the-art methods by improving ROUGE scores on three long meeting summarization datasets AMI, ICSI, and QMSum, two long TV series datasets from SummScreen, and a long document summarization dataset GovReport. Our data and code are available at https://github.com/ANONYMOUS/Summ-N.

1 Introduction

Abstractive summarization helps readers capture salient information from various sources such as documents, news, interviews, and meetings. Previous work has primarily focused on short texts of news (Gehrmann et al., 2018; Zhang et al., 2019) and short conversations (Gliwa et al., 2019; Chen and Yang, 2021). Recently proposed longer dialogue and document summarization tasks (Zhong et al., 2021b; Huang et al., 2021; Chen et al., 2021) pose challenges for current large pretrained language models due to the time and memory complexity of training, as well as limited input lengths these models can consume.

A common method to handle long text reduces the input to a shorter one. This can be accomplished by truncating inputs (Lewis et al., 2020) or employing retrieve-then-summarize pipelines (Zhong et al., 2021b). However, these methods break the dependency of the context and decrease the number of tokens that the model can read, i.e., the receptive field of the model. The cutting-off model depends on the lead bias of the source text, while the retrieve-then-summarize models heavily rely on the independence of retrieved units (turns or sentences) which are usually scattered throughout the source text.

Another approach optimizes the attention mechanism in Transformers to accommodate longer inputs by reducing the impact of quadratic complexity of the attention process using Locality-sensitive hashing (LSH) attention (Kitaev et al., 2020) and Sinkhorn attention (Tay et al., 2020). Additionally, HMNet (Zhu et al., 2020) and HAT-BART (Rohde et al., 2021) use hierarchical self-attention to extend the input limitation of typical self-attention models. However, the simplified attention mechanism weakens the power of pretrained Transformer model, e.g. HMNet does not pretrained on external large-scaled unsupervised dataset as BART did.

In this paper, we propose SUMM^N, a multi-stage framework for long dialogue and document summarization. Figure 1 shows the structure of SUMM^N. First, it divides each source text into segments so that each can be completely fed into the backbone abstractive summarization model. Then, it matches each of them with the subset of target text using a ROUGE-based greedy algorithm. Next, each stage generates a coarse summary for each segment and concatenates them together as the input

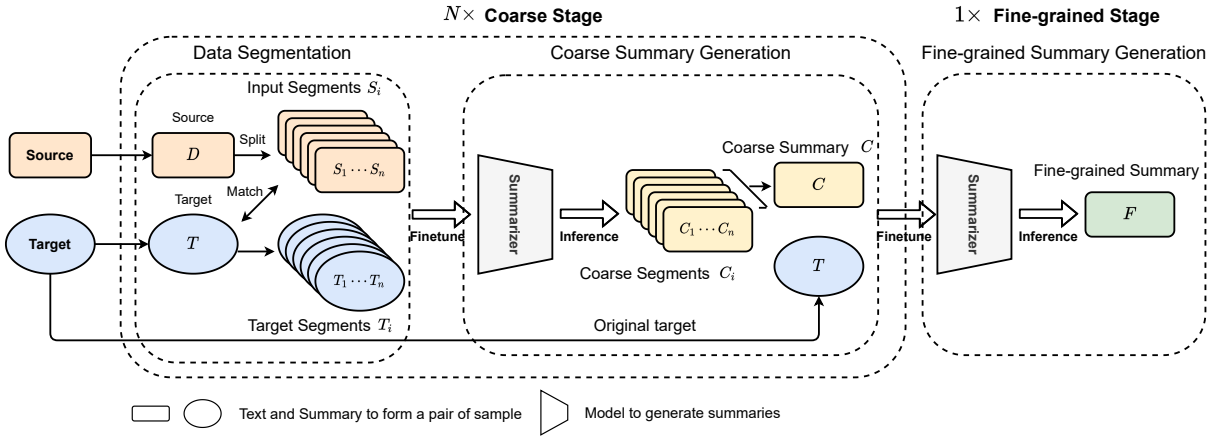

Figure 1: Workflow of the proposed SUMM^N framework. It contains N coarse stages and 1 fine-grained stage. At each coarse stage, source and target text is segmented and paired using a ROUGE-based greedy algorithm, and then a backbone summarization model is used to generate the summary for each segment. After multiple coarse stages, the last fine-graded stage produces the final summary output.

to the next stage. After multiple stages of compression and summarization, the final stage produces a fine-grained summary. The process expands the model context to the full reception field, meaning that the proposed model can read the full input no matter how long the input is. Unlike the retrieve-then-summarize pipelines (Zhang et al., 2019) which extracts sentences usually without their context, SUMM^N only cuts the source text at the end of each segment, so that the context of most sentences remains. In other words, it relies much less on the independence of the context than retrieve-then-summarize pipelines. It does not assume lead bias because each part of the source is fully used. In addition, in each stage, it leverages a backbone abstractive summarization model to recursively generate the summaries. Therefore, it enjoys the full power of the pretrained language models because the framework preserves the intact structure of Transformers.

SUMM^N is flexible to inputs with different lengths by adjusting the number of stages. SUMM^N can change the number of coarse stages according to the compression ratio between source and target, the input limit of the backbone model, and the input source length. We give the empirical formula to decide the number of needed stages for every tested dataset. Experiments show that ROUGE increases on all datasets when increasing the number of stages from one to the appropriate number. Additionally, SUMM^N is flexible because it can be applied to different backbone summarization models. We found that the ROUGE scores increase sharply on AMI dataset when replacing backbone model with T5 (Raffel et al., 2020), and PEGASUS (Zhang et al., 2019).

We conduct extensive experiments on long-input summarization datasets in multiple domains. The results demonstrate that the proposed model significantly outperforms previous state-of-the-art methods according to automatic and human evaluations on three long meeting summarization datasets (AMI, ICSI, QMSum) and one long TV series summarization dataset (SummScreen). It also achieves state-of-the-art performance on a long document summarization dataset (GovReport). These datasets include document summarization as well as both query-based and query-independent long dialogue summarization tasks.

Our contributions are: (1) We propose SUMM^N, a simple, flexible, and effective framework for long dialogue and document summarization. To the best of our knowledge, SUMM^N is the first multi-stage split-then-summarize framework to solve long text summarization tasks. (2) We evaluate SUMM^N on both dialogue and document domains and improve the baseline model by a large margin. (3) We analyze and compare the proposed framework with baselines and discuss its merits in details.

2 Related Work

Long Document Summarization Long document summarization has been studied in multiple domains, such as news (Nallapati et al., 2016), patterns (Trappey et al., 2009), books (Kryściński et al., 2021; Wu et al., 2021), scientific publications (Qazvinian and Radev, 2008), and med-

ical records (Cohan et al., 2018). Gidiotis and Tsoumakas (2020) proposed a divide-and-conquer method by splitting the input into multiple segments, summarizing them separately, and combining the summary pieces. Grail et al. (2021) proposed a hierarchical neural model to process segmented input blocks. Compared with SUMM^N, these models only split the input once, implying the lack of flexibility when handling longer input.

The GovReport dataset was recently introduced containing documents with more than 9000 words, thus greatly challenging the capabilities of current models such as PEGASUS (Zhang et al., 2019), TLM (Subramanian et al., 2019), and BIGBIRD (Zaheer et al., 2020). To handle this dataset, Huang et al. (2021) proposed head-wise positional strides to reduce the cost of the encoder-decoder attention. Similarly, models such as Longformer (Beltagy et al., 2020) and Reformer (Kitaev et al., 2020) adjust attention mechanisms in Transformers to consume longer inputs. However, these models sparsify the attention structure of the pretrained model to fit the longer source text. By contrast, SUMM^N is able to maintain the full structure of various pretrained models.

Long Dialogue Summarization Various models have also been proposed to handle long dialogue summarization. HMNet (Zhu et al., 2020) and HAT-BART (Rohde et al., 2021) leverage a two-level transformer-based model to obtain word level and sentence level representations. DialLM (Zhong et al., 2021a), Longformer-BART-arg (Fabbri et al., 2021) use finetuning or data augmentation to incorporate the external knowledge to maintain the accuracy of lengthy input. Different from these models, SUMM^N is a framework without modifying the structure of the backbone attention model.

Multi-Stage Text Generation Multiple multi-stage coarse-to-fine frameworks have been studied in many other text generation tasks, such as dialogue state tracking (Chen et al., 2020), neural story generation (Fan et al., 2018), and extractive summarization (Xu and Lapata, 2020). In a summarization task, a two-stage extract-and-summarize pipeline is commonly used (Zhang et al., 2019; Subramanian et al., 2019; Zhao et al., 2020). However, unlike that work, our framework aims at long input summarization with fully abstractive intermediate summaries, meaning that SUMM^N can be viewed as a summarize-then-summarize pipeline.

3 Method

Figure 1 shows the workflow of SUMM^N. The workflow includes two types of stages, N coarse stages, and one fine-grained stage. Coarse stages include the data segmentation and coarse summary generation, while the fine-grained stage directly generates the summary as the final result. Besides, we have separate models for each stage and each was separately trained. SUMM^N can adjust and compute the number of coarse stages N according to the stats of dataset and model.

To formulate our task, we denote one sample of the source text as $D = \{D_1, D_2, \cdots, D_m\}$, where D_i indicates one sentence in a document or a dialogue. For query-based summarization, there is also a query Q. The goal is to produce a well-formed summary T, given D and the optional Q.

3.1 Data Segmentation

In long text summarization, the number of tokens in the source data usually exceeds the limit of the backbone summarization models, thus reducing the quality of the summary. To make sure that the model can capture information about all source tokens, we apply a segmentation algorithm for long input summarization datasets. First, we segment the source text so that the data input to the backbone model does not exceed the length limit. Then, we apply a greedy algorithm to find the best target summary that matches the source segments.

Source Segmentation Assume that the number of the maximum input tokens of the backbone model is K. To completely input the source information, we cut the input D (between sentences) into multiple segments, each of them containing fewer than K tokens. Given the input D, we will have n segments $S = \{S_1, S_2, \cdots, S_n\}$ where $S_i \in D$ is a segment in D. For query-based summarization tasks, we simply concatenate the query to the beginning of the S, i.e. $S_i \leftarrow Q \bigoplus S_i$. In both cases, the number of tokens in each segment is less than the hyper-parameter K.

Target Segmentation Segmenting the source text results in n source pieces S_i. We assign each S_i a target $T_i \in T$ to form the new pair (S_i, T_i) for the next step. We use a greedy matching algorithm for target segmentation. We first split T into separate sentences $T_s = \{T_{s_1}, T_{s_2}, \cdots, T_{s_k}\}$. Then, each segment S_i is matched with a subset of T_s such that the ROUGE-1 score between T_s

Algorithm 1 Greedy Target Segmentation

Input: $S_i, T_s = \{T_{s_1}, T_{s_2}, \cdots, T_{s_k}\}$
Output: (S_i, T_i)
  $T_i \leftarrow \Phi$
  loop
    $T_i' \leftarrow T_i$
    for $T_s' \in T_s - T_i$ **do**
      $\tau' \leftarrow \text{ROUGE}_1(S_i, T_i')$
      $\tau \leftarrow \text{ROUGE}_1(S_i, T_i \bigoplus T_s')$
      if $\tau' < \tau$ **then**
        $T_i' \leftarrow T_i \bigoplus T_s'$
      end if
    end for
    if $T_i' = T_i$ **then**
      Break the loop.
    else
      $T_i \leftarrow T_i'$
    end if
  end loop
  return (S_i, T_i)

and S_i is maximized. However, it is not feasible to find the optimal set due to the huge time cost. We apply a simple greedy approximation to find such a subset. From a null set T_i, we iteratively add to the subset the sentence with the highest ROUGE-1 gain between T_s and S_i. Algorithm 1 shows how we obtain the new training pair (S_i, T_i). $\bigoplus$ indicates the concatenation of sentences while keeping them in the same order as in the original text. We use ROUGE-1 as the matching criterion because the higher ROUGE-1 score usually implies higher scores on the other metrics such as ROUGE-2 or ROUGE-L, while ROUGE-1 enjoys lower time complexity compared with other ROUGE metrics.

3.2 Coarse Summary Generation

In coarse summary generation, we train a summarization model, that takes the segmented data as input. Data segmentation helps the summarizer to better learn the task of the current stage. We first collect the training samples (S_i, T_i) generated by data segmentation to form a new dataset. This augments the source data to L_s/K times compared with the cut-off methods, where L_s indicates the length of source text of original dataset. Additionally, because we incorporate the full input using segmentation, it does not rely on the leading bias in the cut-off method that only considers the first segment S_1. Afterward, we use these data to train a neural summarizer. This way, our model treats each part of the source text as equally important.

Given a source segment S_i and an optional query Q, we obtain the coarse summary segments using a backbone summarization model:

$$\hat{T}_i^l = \text{SUMM}_l(Q, S_i)$$

Where l is the index of the current stage. Then, the n coarse summaries corresponding to the original source $S = \{S_1, S_2, \cdots, S_n\}$ are concatenated: $\hat{T}^l = \hat{T}_1^l \bigoplus \hat{T}_2^l \bigoplus \cdots \bigoplus \hat{T}_n^l$. We use $\hat{T}^l$ as the new source text of next stage, which compresses the input source data D^l. i.e. $D^{l+1} = \hat{T}^l$. To pair with the D^{l+1}, the target to the next stage is copied from the original dataset, i.e. $T^{l+1} = T$.

The proposed framework is applicable to different backbone models $\text{SUMM}_l(*)$, such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020). We pick BART as the backbone model because it can best illustrate the benefits of our framework (Section 4.1).

3.3 Number of Coarse Stages

The number of stages can be computed by data stats and model characteristics. In SUMM^N, each coarse stage compresses the input to a shorter length. After N turns of coarse stages, the averaged length of source text is below K, the dataset is then fed into the fine-grained stage. Hence, the number of coarse stage can be computed by the following equation (details can be found in appendix):

$$\frac{L_s}{K^N} \times |T_i|^N \leq K$$

$$N = \lceil \frac{\log K - \log L_s}{\log |T_i| - \log K} \rceil$$

Where $*^N$ indicates the N-th power of $*$, and $|T_i|$ is the averaged length of the segmented targets. Table 1 shows the N for each dataset.

The greedy algorithm in SUMM^N for target segmentation is critical to the performance. Consider a duplication algorithm where each segment S_i is simply paired with the target T, i.e. $T_i = T$. Since the target text is longer than segmented text, the generated summary of each coarse stage will be longer as well, leading to a lower compression speed and larger N. Besides, the duplication of the target will confuse the model, because some source segments will probably be paired with the same target, causing the model to generate duplicated content. Experiments (Table 7, "- stage 2" versus "- stage 2 tar. seg.") show that ROUGE scores declines a lot when greedy target segmentation is replaced by the duplication algorithm .

Dataset	Type	Domain	Size	Source length	Target length	Query	$N+1$
AMI	Dialogue	Meetings	137	6007.7	296.6	✗	2
ICSI	Dialogue	Meetings	59	13317.3	488.5	✗	3
QMSum	Dialogue	Meetings	1808	9069.8	69.6	✓	2
SummScreen	Dialogue	TV shows	26851	6612.5	337.4	✗	2
GovReport	Document	Reports	19466	9409.4	553.4	✗	3

Table 1: The summarization datasets for evaluation. The source length and target length is the averaged number across the dataset. N indicates the number of coarse stages.

3.4 Fine-Grained Summary Generation

When the input source of D^l is shorter than K, we can precede to the fine-grained stage. In this stage, D^l is used to train a summarization model from scratch to obtain the final summary. The fine-grained stage works the same way as the vanilla backbone model. In fact, SUMM^N with $N = 0$ is the backbone summarizer. In the fine-grained stage, the model is directly trained on dataset (D^{L_c}, T) from the last coarse stage, and obtain the summary as the final output of SUMM^N:

$$\hat{T}^{L_c+1} = \text{SUMM}_{L_c+1}(Q, D^{L_c})$$

It is worth noting that, although source text may be shorter than 2 segments, i.e. $L_s \leq K$, we still add them in all stages, so that each summarization model can be trained on the full dataset.

4 Experiment Setup

We first list the datasets and metrics to evaluate the model. Then, we introduce the backbone model and baselines for comparisons. Finally, we present some implementation details.

4.1 Datasets and Metrics

Table 1 shows data statistics for the datasets.

AMI & ICSI (McCowan et al., 2005; Janin et al., 2003) are meeting scripts generated by Automatic Speech Recognition (ASR) systems. AMI is collected from product design meetings in a company while ICSI is collected from academic group meetings. Because the transcript is produced by ASR, there is a word error rate of 36% for AMI and 37% for ICSI.

QMSum (Zhong et al., 2021b) is a query-based meeting summarization dataset. It consists of meetings from three domains, including AMI and ICSI, and the committee meetings of the Welsh Parliament and the Parliament of Canada. Each query and sample are written by experts.

SummScreen (Chen et al., 2021) consists of community-contributed transcripts of television show episodes from The TVMegaSite, Inc. (TMS) and ForeverDream (FD). The summary of each transcript is the recap from TMS, or a recap of the FD shows from Wikipedia and TVMaze.

GovReport (Huang et al., 2021) is a large-scale long document summarization dataset with 19,466 long reports published by the U.S. Government Accountability Office on national policy issues.

We use ROUGE (Lin, 2004) as the automatic evaluation metric for all experiments. We use the `pyrouge` library [1] as the implementation. We split summary outputs into sentences to calculate the ROUGE-L score.

4.2 Backbone Model

We pick BART (Lewis et al., 2020) as our backbone summarization model because it performs well on short text summarization but not as good on longer texts, illustrating the benefits of our framework. Compared with other pretrained parameters, the BART-large model pretrained on the CNN/DM dataset yields the best performance (Zhang et al., 2021). So we use BART-large-cnn parameter as a better starting point.

It is worth noting that we use separate backbone models for each stage and each was separately trained. We experimented with reusing the model parameters in multiple stages but obtained a lower score, e.g. the ROUGE-1 score of stage 2 on the QMSum dataset decreases around two points if we use the best parameters of stage 1 summarizer as the starting point of training stage 2 summarizer. This is because the tasks of the different stages differ significantly. For instance, the input to the first stage of dialogue summarization is dialogue turns while the input to the latter stages is documents.

[1] https://github.com/bheinzerling/pyrouge

	AMI			ICSI			QMSum-All			QMSum-Gold		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
PGNet	42.60	14.01	22.62*	35.89	6.92	15.67*	28.74	5.98	25.13	31.52	8.69	27.63
TopicSeg	51.53	12.23	25.47*	-	-	-	-	-	-	-	-	-
HMNET	52.36	18.63	24.00*	45.97	10.14	18.54*	32.29	8.67	28.17	36.06	11.36	31.27
TextRank	35.19	6.13	16.70*	30.72	4.69	12.97*	16.27	2.69	15.41	-	-	-
HAT-BART	52.27	20.15	50.57	43.98	10.83	41.36	-	-	-	-	-	-
DDAMS	53.15	**22.32**	25.67*	40.41	11.02	19.18*	-	-	-	-	-	-
SUMM^N	**53.44**	20.30	**51.39**	**48.87**	**12.17**	**46.38**	**34.03**	**9.28**	**29.48**	**40.20**	**15.32**	**35.62**

Table 2: ROUGE scores on three meeting summarizing tasks, AMI, ICSI, and QMSum. QMSum-ALL use inputs with all turns while MSum-Gold use inputs with only the gold turns. * denote the ROUGE-L scores without sentence split.

4.3 Baselines

We compare the proposed framework with various baselines. **PGNet** (See et al., 2017) uses a pointer mechanism to copy the token from the training sample. **TopicSeg** (Li et al., 2019) is a multi-modal model jointly learning the segmentation and summarization. **HMNet** (Zhu et al., 2020) uses a hierarchical attention structure and cross-domain pre-training for meeting summarization. **TextRank** (Mihalcea and Tarau, 2004) is a graph-based ranking model for text processing. **HAT-BART** (Rohde et al., 2021) is a new hierarchical attention transformer-based architecture that outperforms standard Transformers. **DDAMS** (Feng et al., 2021) uses a relational graph to model the interaction between utterances by modeling different discourse relations.

For the SummScreen dataset, we use the neural and hybrid model scores reported by Chen et al. (2021). We rename these two baselines as **Longformer+ATT** and **NN+BM25+Neural** to clarify the difference between other baselines.

The baseline scores we report on GovReport are from the original paper (Huang et al., 2021). **BART Variant** indicates self-attention variants with full attention. **BART HEPOS** indicates encoder variants with head-wise positional strides (HEPOS) encoder-decoder attention.

4.4 Implementation Details

We fit all models into a single RTX A6000 GPU with a 48 GiB memory. We adopt the fairseq[2] implementation for BART. The learning rate is set to 2e-5 and the beam width is set to 2 for coarse stages and 10 for fine-grained stages. The maximum number of tokens in each batch is set to 2048. The maximum number of tokens in each source text is set to 1024 because we tried to extend the positional embeddings to 2048 or longer but obtained worse performance. For the output of each intermediate stage, we use <s> and </s> to separate each generated target segments $\hat{T}_i^l$.

[2]https://github.com/pytorch/fairseq

5 Results and Analysis

We discuss the evaluation results and effects of each component of SUMM^N in this section.

5.1 Overall Results

Meeting Summarization Table 2 shows the ROUGE scores on AMI, ICSI, and QMSum. Compared with the baseline models, SUMM^N achieves state-of-the-art results on almost all metrics. Specifically, SUMM^N improves SOTA on ICSI by **2.9**, and **0.83** ROUGE-1/2 scores, improves SOTA on QMSum-Gold by **4.14**, **3.96**, and **4.35** ROUGE-1/2/L scores. These results demonstrate the effectiveness of SUMM^N on long dialogue summarization tasks.

TV Series Summarization Table 3 shows ROUGE scores on SummScreen. SUMM^N outperforms on almost all metrics on two SummScreen datasets. Specifically, we improve **6.58**, **1.92**, and **3.34** ROUGE-1/2/L scores on the SummScreen-FD dataset. This result demonstrates the generalizability of SUMM^N over various domains including meetings and TV series.

Document Summarization Table 4 shows ROUGE scores on GoveReport. SUMM^N achieves state-of-the-art performance on ROUGE-2 and ROUGE-L, and compatible results on ROUGE-1. The results show that SUMM^N is applicable to

	SummScreen-FD			SummScreen-TMS		
	R1	R2	RL	R1	R2	RL
Longformer+ATT	25.90	4.20	23.80	42.90	**11.90**	41.60
NN+BM25+Neural	25.30	3.90	23.10	38.80	10.20	36.90
SUMM^N	**32.48**	**6.12**	**27.14**	**44.64**	11.87	**42.53**

Table 3: ROUGE scores on the SummScreen datasets including ForeverDreaming (SummScreen-FD) and TV MegaSite, Inc. (SummScreen-TMS).

	R-1	R-2	R-L
BART Variants			
Full (1024)	52.83	20.50	50.14
Stride (4096)	54.29	20.80	51.35
LIN. (3072)	44.84	13.87	41.94
LSH (4096)	54.75	21.36	51.27
Sinkhorn (5120)	55.45	21.45	52.48
BART HEPOS			
LSH (7168)	55.00	21.13	51.67
Sinkhorn (10240)	**56.86**	22.62	53.82
SUMM^N	56.77	**23.25**	**53.90**

Table 4: ROUGE scores on GovReport. For each baseline model, the number in parentheses is the maximum input length.

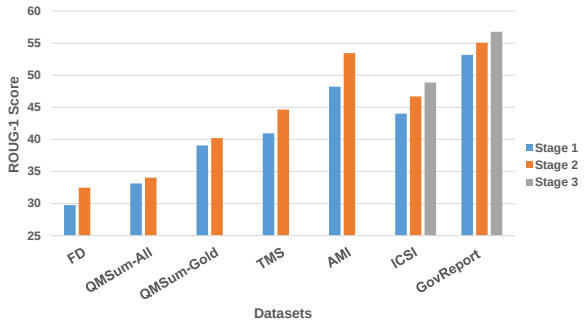

Figure 2: ROUGE-1 scores of various datasets at different stages. ICSI and GovReport have 3 stages, while the others have 2 stages. In all datasets, ROUGE-1 score increases with the increasing number of stages.

both long dialogue and document summarization tasks.

5.2 Effects of Number of Stages

We also notice that the performance increases consistently when the number of stages goes up until the predefined number of stages. Figure 2 shows the ROUGE-1 scores of different tasks across stages. **Stage 1** indicates the model with only one coarse stage and no fine-grained stage. In this model, We directly use the first segment of the coarse summary as the output, i.e. $\hat{T}_1^1$ of each sample. **Stage i** ($i > 1$) model contains $i - 1$ coarse stages and one fine-grained stage, the generated summary is from fine-grained summarization models, i.e. $\hat{T}^i$.

Model	Method	R1	R-2	R-L
AMI	Backbone	46.57	16.41	44.61
	SUMM^N	**53.44**	**20.30**	**51.39**
ICSI	Backbone	39.91	9.98	38.17
	SUMM^N	**48.87**	**12.17**	**46.38**
QMSum-All	Backbone	29.20	6.37	25.49
	SUMM^N	**34.03**	**9.28**	**29.48**
QMSum-Gold	Backbone	32.18	8.48	28.56
	SUMM^N	**40.20**	**15.32**	**35.62**

Table 5: Improvements of SUMM^N over backbone BART models on AMI, ICSI, and QMSum datasets.

Although stage 2 of SUMM^N on the ICSI dataset has already outperformed the baselines, the scores can be further improved by adding one more coarse stage. In fact, on all datasets, increasing the number of stages leads to a performance gain. This gain can be explained as the following: if the output of the current stage is longer than K tokens, adding one more coarse stage will help since the model will receive more information from the source text compared with simply truncating them. On the contrary, if the input is smaller than K, there is no need to add more stages, because there is only one segment.

5.3 Improvements over Backbone Models

SUMM^N also boosts the performance of a backbone model by a large margin. As shown in Table 5, it improves the BART-large model by **6.87**, **3.89**, **6.78** ROUGE-1/2/L on AMI. This indicates the capability of SUMM^N to boost the performance of a weak learner on long summarization tasks. In particular, when the backbone model is well pretrained on short input texts and performs well on short summarization tasks, SUMM^N could greatly increase the capability of the backbone model to process and read long source texts. Also, the backbone of SUMM^N can be easily replaced by some other models, and models do not necessarily have to be identical at every stage. For example, one can try different learners such as T5 as the backbone model and replace the model in stage 1 with a dialogue-to-document model.

5.4 Generalizability over Backbone Models

To demonstrate our framework can generalize to different backbone summarization models, we re-

Model	Method	R-1	R-2	R-L	Input
BART-base	Backbone	41.54	13.80	38.75	1024
	SUMM^N	46.60	18.80	45.23	1024
T5-large	Backbone	47.81	16.06	45.77	512
	SUMM^N	51.85	19.40	49.94	512
PEGASUS-cnn_dailymail	Backbone	46.37	16.21	44.75	1024
	SUMM^N	50.15	19.07	48.28	1024

Table 6: ROUGE scores of different backbone models on AMI. For all backbone models with various maximum input lengths, ROUGE scores increase with the help of proposed framework. Input indicates the maximum number of tokens the model can take.

	R-1	R-2	R-L
SUMM^N	**53.44**	**20.30**	**51.39**
- stage 2	48.21	18.59	46.46
- data seg.	46.83	15.91	45.00
- stage 2 & tar. seg.	46.24	16.03	44.45
only BART	46.57	16.41	44.61

Table 7: Ablations on the test set of AMI. "- data seg." indicates removing data segmentation (the same as cut-off at limitation), "- tar. seg." indicates source segmentation paired with duplicated targets.

place the BART-large-cnn model in previous experiments with other neural summarization models including T5 (Raffel et al., 2020) and PEGASUS (Zhang et al., 2019).[3] Table 6 shows the ROUGE scores of three different models that are trained and evaluated on AMI. In all models, SUMM^N improves the performance of backbone models by a large margin. For instance, although BART-base is a weaker summarizer compared with BART-large model, the framework is still able to improve the ROUGE-1 score by **5.06**.

5.5 Ablations

Table 7 shows the ablation study results of SUMM^N on the AMI test set. Removing stage 2 (using the first segment of the coarse summary $\hat{T}_1^1$ as the generated summary) leads to a 5.23 ROUGE-1 score drop. Without data segmentation, the ROUGE-1 score decreases by 6.61 using the same fine-grained stage. Removing both stage 2 and target segmentation (use duplication algorithm instead) further decreases the performance. It even hurts the performance of the original BART model because the duplication of targets will introduce some biases towards the common part of the targets.

[3]We use huggingface to implement the T5 and PEGASUS models https://huggingface.co/

	AMI			ICSI		
	Read.	Conc.	Cove.	Read.	Conc.	Cove.
HMNet	3.93	4.05	4.15	3.21	3.33	3.84
SUMM^N	**4.45**	**4.13**	**4.23**	**4.12**	**3.55**	**4.06**

Table 8: Human evaluation scores. Read. indicates *Readability*, Conc. indicates *Conciseness*, and Cove. indicates *Coverage*.

5.6 Human Evaluation

We conduct a human evaluation to assess the following: *Readability* takes into account word and grammatical error rate to evaluate how fluent the summary language is; *Conciseness* measures how well the summary discards the redundant information; *Coverage* measures how well the summary covers each part of the dialogue.

We compare the results of SUMM^N and HMNet because HMNet is a baseline model with the good capability to read whole input. For each meeting in AMI and ICSI dataset, we ask 3 different annotators with English expertise to label the summaries. Each annotator was asked to read the meeting transcript, gold summaries, and generated summaries using the SummVis (Vig et al., 2021) toolkit. They were asked to rate each summary from 1 to 5 (higher is better) for each metric. We also shuffle the summaries of two models to reduce the bias.

Table 8 shows that SUMM^N achieves higher scores in *Readability*, *Conciseness*, and *Coverage* than HMNet in both AMI and ICSI dataset. Specifically, the *Readability* of SUMM^N greatly surpasses the baseline by around 0.5/1 point on AMI/ICSI dataset. This is because BART is well-pretrained and is able to generate more readable text and SUMM^N successfully maintains this capability of BART.

6 Conclusion

In this paper, we propose SUMM^N, a simple, flexible, and effective framework for long dialogue and document summarization. It consists of multiple coarse stages and one fine-grained stage to iteratively compress the long source input. It enjoys the full power of backbone models while ensuring the full receptive field of the summarization model. We evaluate the model on various datasets and improve the baselines by a large margin.

References

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

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2021. Summscreen: A dataset for abstractive screenplay summarization. *arXiv preprint arXiv:2104.07091*.

Zhi Chen, Lu Chen, Zihan Xu, Yanbin Zhao, Su Zhu, and Kai Yu. 2020. Credit: Coarse-to-fine sequence generation for dialogue state tracking. *arXiv preprint arXiv:2009.10435*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6866–6880, Online. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

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 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3808–3814. ijcai.org.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Quentin Grail, Julien Perez, and Eric Gaussier. 2021. Globalizing BERT-based transformer architectures for long document summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1792–1810, Online. Association for Computational Linguistics.

Luyang Huang, Shuyang Cao, Nikolaus Nova 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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1419–1436. Association for Computational Linguistics.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 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. IEEE.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*.

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.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

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.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al. 2005. The ami meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, page 100. Citeseer.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 689–696, Manchester, UK. Coling 2008 Organizing Committee.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Tobias Rohde, Xiaoxia Wu, and Yinhan Liu. 2021. Hierarchical learning for generation with long source sequences. *arXiv preprint arXiv:2104.07545*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Sandeep Subramanian, Raymond Li, Jonathan Pilault, and Christopher Pal. 2019. On extractive and abstractive neural document summarization with transformer language models. *arXiv preprint arXiv:1909.03186*.

Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. 2020. Sparse sinkhorn attention. In *International Conference on Machine Learning*, pages 9438–9447. PMLR.

Amy JC Trappey, Charles V Trappey, and Chun-Yi Wu. 2009. Automatic patent document summarization for collaborative knowledge systems and services. *Journal of Systems Science and Systems Engineering*, 18(1):71–94.

Jesse Vig, Wojciech Kryściński, Karan Goel, and Nazneen Fatema Rajani. 2021. Summvis: Interactive visual analysis of models, data, and evaluation for text summarization.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nissan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Yumo Xu and Mirella Lapata. 2020. Coarse-to-fine query focused multi-document summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3632–3645, Online. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. In *NeurIPS*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, and Dragomir Radev. 2021. An exploratory study on long dialogue summarization: What works and what's next. *arXiv preprint arXiv:2109.04609*.

Yao Zhao, Mohammad Saleh, and Peter J Liu. 2020. Seal: Segment-wise extractive-abstractive long-form text summarization. *arXiv preprint arXiv:2006.10213*.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021a. Dialoglm: Pre-trained model for long dialogue understanding and summarization. *arXiv preprint arXiv:2109.02492*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021b. 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.

A Case Study

Table 9 shows a concrete sample summary generated by SUMM^N. It captures the topics of the source text and smoothly follows the outline of the gold summary. Also, SUMM^N is able to evenly generate the information of the whole summary, including the last part of source text which is truncated in the standard BART-large models.

B Computing the Number of Stages

With regard to text length, the source text of each stage needs to be compressed gradually to ensure that the summary with proper length can be generated in the final stage. Also, the compression level determines the required number of stages, which is a significant indicator of time cost. Suppose the source of stage i contains L_s^i words, while the target contains L_t^i words, and the maximum input length of the model is K, $|T_i|$ indicates the averaged number of tokens in the segmented target. L_t^i can be expressed by the number of segment $\frac{L_s^i}{K}$ times $|T_i|$.

In each stage, we have:

$$L_t^i = \frac{L_s^i}{K} \times |T_i|$$

$$L_s^i = L_t^{i-1}$$

By iterating this equation for N time, the number of needed coarse stages N for a dataset can be decided in this way:

$$\frac{L_s^0}{K^N} \times |T_i|^N \leq K$$

$$N = \lceil \frac{\log K - \log L_s^0}{\log |T_i| - \log K} \rceil$$

Where $*^N$ indicates the N-th power of $*$, while L_s^0 indicates the source text of original dataset. For target segmentation, the compression level $|T_i|/|T|$ of duplication segmentation is 1 and greedy segmentation is 0.5 to 0.9. So that target segmentation algorithm helps reduce stages.

	ICSI
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. · · · 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.
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.· · · 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.

Table 9: Sample output summary SUMM^N on ICSI dataset. Tokens marked in grey indicate the out-of-boundary contents of truncation models. Brown tokens are the keywords emerged in the gold summary. Tokens marked in red indicate the concepts of out-of-boundary text.