arXiv:2109.02492v1 [cs.CL] 6 Sep 2021

DIALOGLM: Pre-trained Model for Long Dialogue Understanding and Summarization

Ming Zhong[*1], **Yang Liu**[2], **Yichong Xu**[2], **Chenguang Zhu**[2], **Michael Zeng**[2]
[1]University of Illinois at Urbana-Champaign
[2]Microsoft Cognitive Services Research Group
mingz5@illinois.edu, {yaliu10, yichong.xu, chezhu, nzeng}@microsoft.com

Abstract

Dialogue is an essential part of human communication and cooperation. Existing research mainly focuses on short dialogue scenarios in a one-on-one fashion. However, multi-person interactions in the real world, such as meetings or interviews, are frequently over a few thousand words. There is still a lack of corresponding research and powerful tools to understand and process such long dialogues. Therefore, in this work, we present a pre-training framework for long dialogue understanding and summarization. Considering the nature of long conversations, we propose a window-based denoising approach for generative pre-training. For a dialogue, it corrupts a window of text with dialogue-inspired noise, and guides the model to reconstruct this window based on the content of the remaining conversation. Furthermore, to process longer input, we augment the model with sparse attention which is combined with conventional attention in a hybrid manner. We conduct extensive experiments on five datasets of long dialogues, covering tasks of dialogue summarization, abstractive question answering and topic segmentation. Experimentally, we show that our pre-trained model DIALOGLM significantly surpasses the state-of-the-art models across datasets and tasks.

Introduction

Dialogue plays a vital role in interpersonal interaction in daily life, workplace or online forums, and it has drawn extensive attention from both academia and industry (Zhang et al. 2020b). With the development of speech recognition systems and the growing need of remote work, an increasing number of long conversations are recorded and transcribed, such as meeting minutes, interviews and debates. These long dialogues are dense medium of information, bringing challenges for users to quickly understand the gist and extract related information. To address these challenges, many NLP tasks are proposed, including dialogue summarization, dialogue-based question answering and dialogue segmentation (Feng, Feng, and Qin 2021; Feng et al. 2021; Zhong et al. 2021; Zou et al. 2021a,b; Chen et al. 2021b; Koay et al. 2021; Hsueh, Moore, and Renals 2006). However, different from monologic texts like news, long conversations have dialogic structures and lengthy input which is difficult for current NLP systems to process them. Thus, exploring a model that can better understand and summarize an entire long dialogue is practically needed.

Recently, pre-trained neural language models achieve remarkable success on a spectrum of natural language tasks (Devlin et al. 2018; Liu et al. 2019). However, these general-purpose models are pre-trained on free-form text data with universal objectives. Although this can obtain powerful contextualized language representations, it also limits their ability in specific domains. Motivated by this, several dialogue-related pre-trained models have been proposed to tackle different tasks like conversational response generation (Zhang et al. 2020b), dialogue response ranking (Gao et al. 2020) and multi-party conversation understanding (Gu et al. 2021). Nonetheless, these models are limited to short conversations (e.g, usually fewer than 200 words), and hence are not capable of handling long dialogues (usually longer than 5,000 words) with more speakers and utterances. On the other hand, when it comes to long sequences, subsequent research focuses on improving self-attention method (Kitaev, Kaiser, and Levskaya 2020; Wang et al. 2020a) and facilitating the interaction of local and global information (Beltagy, Peters, and Cohan 2020; Zaheer et al. 2020). However, these systems are not designed for dialogues and thus they learn limited knowledge of the dialogic structures. In general, existing models all have their own dilemmas when dealing with long conversations.

In this paper, we present DIALOGLM, a pre-trained neural encoder-decoder model for long dialogue understanding and summarization. DIALOGLM is established on the sequence-to-sequence model architecture and can be applied to a wide range of natural language processing tasks. As shown in Figure 1, we propose a window-based denoising pre-training task on a large dialogue corpus: (1) select a window containing multiple consecutive turns from a conversation; (2) inject arbitrary dialogue-related noise into the window, and (3) train the model to restore this window based on the rest of the conversation. Intuitively, a pre-trained model should be able to reconstruct the noisy window, as speaking style of the interlocutors and topic content exist distributedly in a long conversation. Compared with sentence-level masking like PEGASUS (Zhang et al. 2020a), a window composed of multiple turns contains more coherent and informa-

[*]Ming completed this work during his internship at Microsoft. Correspondence to: Yang Liu (yaliu10@microsoft.com).
Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

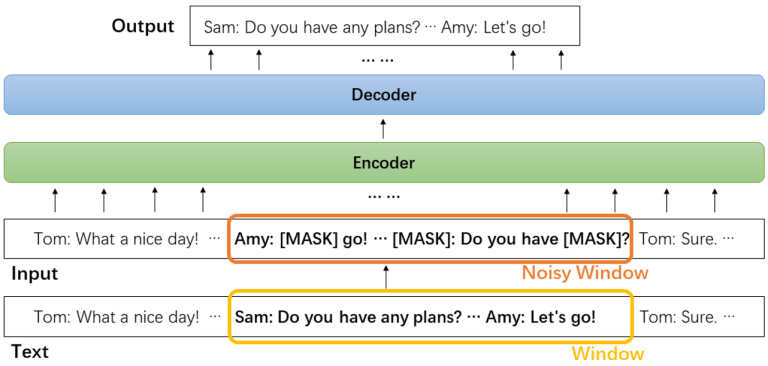

Figure 1: Pre-train task for DIALOGLM: window-based denoising. We firstly select a window containing multiple turns, and inject different dialogue-inspired noises into it. Finally, we train the model to restore this window based on the noisy window and the rest of the dialogue.

tive text, which is important for recognizing the format of dialogue. Compared to full-text denoising like BART (Lewis et al. 2020), window-based methods not only require less computational resource, which has significant advantages when dealing with long sequences, but also are better suited to downstream tasks like dialogue summarization.

Furthermore, we design five types of pre-train noises to generate a noisy window based on the characteristics of dialogues: *Speaker Mask*, *Turn Splitting*, *Turn Merging*, *Text Infilling* and *Turn Permutation*. These challenging transformations disrupt both content and order of speakers and utterances. Therefore, to reconstruct the window, DIALOGLM has to fully understand the special format and text style of speaker-utterance pairs, and grasp the general content of the complete dialogue. Moreover, to process longer sequences and reduce training time, we introduce a hybrid attention approach into our model. For most neural layers, we utilize a sparse attention method (Tay et al. 2020) to capture local information; for other layers, global self-attention is used for perceiving the full dialogue semantics. This hybrid attention approach allows DIALOGLM to accept more than 8,000 input words while achieving excellent model performance.

Experimentally, DIALOGLM surpasses previous models by a large margin in long dialogue understanding and summarization tasks. Specifically, for dialogue summarization and abstractive question answering, our model outperforms the pre-trained model BART and Longformer (Beltagy, Peters, and Cohan 2020) on five datasets including meeting and screenplay domains, achieving new state-of-the-art results across multiple datasets. DIALOGLM also shows its superiority over strong baseline models for dialogue segmentation task. Ablation studies verified the effectiveness of each component in our pre-training framework. The results demonstrate that each dialogue-inspired noise and the proposed hybrid attention methods can bring further improvements to the model. In addition to automatic evaluation, for generation tasks, we also perform human evaluation on the generated sequences from three dimensions: fluency, informativeness, and faithfulness to the original dialogue. Compared with previous powerful models, DialogLM provides considerable benefits in different perspectives.

Related Work

Pre-trained Neural Models for Dialogues

Most dialogue-related pre-trained models focus on specific tasks, such as dialogue response generation (Zhang et al. 2020b; Bao et al. 2020b; Cao et al. 2020; Wang et al. 2020b), dialogue response selection (Wu et al. 2020; Gao et al. 2020) and multi-party conversation understanding (Gu et al. 2021). Generally speaking, they either further pre-train general-purpose pre-train models on open-domain conversational data like Reddit for dialogue response generation and selection (Henderson et al. 2020; Zhang et al. 2020b; Bao et al. 2020b), or conduct task-specific training for downstream applications (Li et al. 2020; Wu et al. 2020; Gu et al. 2021). Unlike these previous studies, our pre-train task is not limited to concrete tasks. We hope that through window-based denoising, the model can learn the format and characteristics of dialogues in a general way, thereby performing better in various dialogue-oriented tasks. On the other hand, these work only focus on short dialogue scenes, and usually limit the length of input dialogue. As a result, we still lack powerful NLP tools for long conversations with more speakers and more utterances.

Pre-trained Neural Models for Long Sequences

Processing long sequences is a natural need in many NLP tasks. For the Transformer (Vaswani et al. 2017) architecture, the core difficulty lies in the computational complexity of the self-attention module, which grows quadratically with the sequence length. Recently, many methods are proposed to tackle the long sequence problem by improving the self-attention mechanism. Specifically, Linformer (Wang et al. 2020a) uses linear mapping to compress the input sequences under the assumption that the attention mechanism matrix is low-rank. Block/bucket-based local attention (Kitaev, Kaiser, and Levskaya 2020; Wang et al. 2021; Roy et al. 2021) utilize the random-projection hashing function or clustering approach to allocate highly similar tokens into a same bucket. Sliding window-based attention (Beltagy, Peters, and Cohan 2020; Zaheer et al. 2020; Zhang et al. 2021a) introduce sliding window attention to capture local informa-

Noise Type	Original Dialogue	Noisy Dialogue
Speaker Mask	Tom: The weather is good today!	[MASK]: The weather is good today!
Turn Splitting	Tom: The weather is good today! Do you have any plans? How about we go to play basketball?	Tom: The weather is good today! [MASK]:Do you have any plans? [MASK]:How about we go to play basketball?
Turn Merging	Tom: The weather is good today! Do you have any plans? Bob: I still have homework to do today. I'm afraid I can't go out to play.	Tom: The weather is good today! Do you have any plans? I still have homework to do today. I'm afraid I can't go out to play.
Text Infilling	Tom: The weather is good today! Do you have any plans? How about we go to play basketball?	Tom: The weather is [MASK] Do you have [MASK] any plans? [MASK] we go to play basketball?
Turn Permutation	Tom: Do you have any plans? Bob: How about we go to play basketball? Sam: I still have homework to do today. I'm afraid I can't go out to play.	Sam: I still have homework to do today. I'm afraid I can't go out to play. Tom: Do you have any plans? Bob: How about we go to play basketball?

Table 1: Dialogue-related noise for generating a noisy window. The blue [MASK] token means that a speaker name is masked, and the red [MASK] token indicates that a text span in the utterance is masked.

tion and retain part of full attention for global information We adopt the last two approaches in this paper to reduce computational cost by mixing Sinkhorn attention (Tay et al. 2020) and global attention in the Transformer structure.

Method

In this section, we first introduce the pre-training task for DIALOGLM: window-based denoising and five types of dialogue-inspired noises. Then we describe the overall architecture of our pre-trained model.

Window-based Denoising

A long conversation usually involves a core theme and multiple main speakers. For instance, the meetings in the AMI corpus (Carletta et al. 2005) are about product design in the industrial setting, including discussions among product managers, industrial designers, marketing experts, and user interface designers. Long dialogue with thousands of words can portray the speaking style of different people, e.g., product managers speak actively and energize the audience to help them brainstorm, while marketing experts usually use statistics to state their opinions. Also, a conversation is coherent and its content in different parts are closely related. Therefore, it is possible to infer the speakers and general content of part of the conversation based on the rest context.

Inspired by this, we propose a novel pre-training task for DIALOGLM: window-based denoising. Formally, given a long dialogue $D = (x_1, x_2, \cdots, x_n)$ consisting of n turns, where each turn x_i represents a speaker-utterance pair $x_i = (s_i, u_i)$, we firstly select a random window containing multiple consecutive turns $W = (x_j, x_{j+1}, \cdots, x_{j+m})$. Next, we inject several dialogue-related noise into it to generate a noisy window $W' = (x'_j, x'_{j+1}, \cdots, x'_{j+m'})$. During the pre-training phase, we concatenate all the turns into a long sequence and replace the window with the noisy version as input to the model, i.e., $X = (x_1, \cdots, x'_j, \cdots, x'_{j+m'}, \cdots, x_n)$. The objective is to restore this selected window W by modeling the conditional distribution $p(x_j, x_{j+1}, \cdots, x_{j+m}|X)$. As illustrated in Figure 1, several turns are chosen as a window, and we generate a noisy window by disrupting their order and masking part of the content and speaker information. The decoder is trained to reconstruct the original window based on the noisy window and the rest of the conversation.

The most relevant work to our proposed pre-train task is full-text denoising by BART (Lewis et al. 2020) and sentence-level masking by PEGASUS (Zhang et al. 2020a). However, for sequences with more than 5,000 words, full-text denoising requires unaffordable computational resources. Correspondingly, our window-based approach is a flexible alternative and allows us to add more completely transformed noise without worrying about the model being unable to recover it. On the other hand, unlike documents, a large number of individual turns in a conversation are not informative, such as merely greeting others or chatting about daily routines that have nothing to do with the theme. Therefore, sentence/turn-level masking does not necessarily enable the model to understand the core content of the whole dialogue, but a window with multiple successive turns is more likely to contain meaningful and coherent information. So we argue that, compared with previous frameworks, window-based denoising can be more suitable for pre-training a model for processing long conversations.

Dialogue-Inspired Noise

The next question is, how do we generate a noisy window? In order to make the model aware of the characteristics of dialogues and their special speaker-utterance format, we design the following five types of noise (see Table 1):

Speaker Mask For speaker names of each turn in the window, 50% of them are randomly sampled and replaced with a special [MASK_SPEAKER] token.

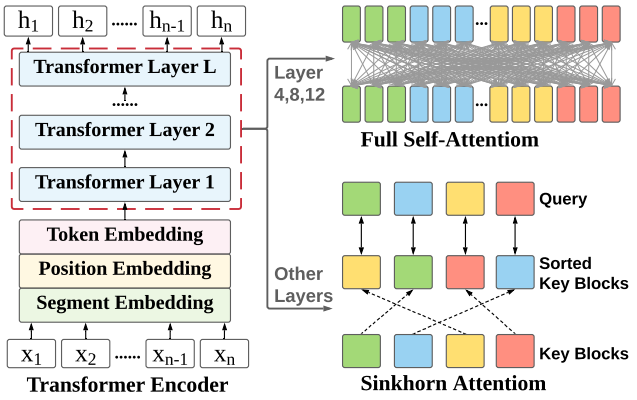

Figure 2: Model architecture for DIALOGLM. We introduce a hybrid attention approach in Transformer architecture: most layers are equipped with a sparse attention method (Sinkhorn attention) and the rest retain global self-attention.

Turn Splitting A single turn in a conversation can be composed of multiple sentences. We select the turn with the largest number of sentences in the window and split it into multiple turns. We keep the speaker of the first split turn unchanged and take [MASK_SPEAKER] as the speaker of all subsequent newly split turns.

Turn Merging We randomly sample multiple consecutive turns and merge them into one turn. Keep the speaker of the first turn unchanged, and delete all the speakers of subsequent turns. The number of merged turns is drawn from the Poisson distribution ($\lambda = 3$), and is set to be at least 2.

Text Infilling In a window, we randomly sample several text spans and replace each span with a [MASK] token. The length of text span is also drawn from the Poisson distribution ($\lambda = 3$). 0-length span corresponds to the insertion of a [MASK] token as in Lewis et al. (2020).

Turn Permutation We shuffle all turns in the window in random order. This noise is added after *Turn Merging* and *Turn Splitting*. It can further disrupt the speaker and turn information, so that the model can only restore the window when it fully comprehends the context.

Model Architecture

We choose Transformer as our backbone neural architecture because it shows promising performance and flexibility on various NLP tasks. For long dialogue processing, pre-trained models based on Transformer like BART and UNILM (Dong et al. 2019) have two limitations: 1) they have no pre-training data in dialogue format and no pre-training tasks designed for modeling dialogues; 2) the text length used during pre-training is short (1,024 for BART and 512 for UNILM). Regarding the first issue, we use window-based denoising approach to pre-train our model DIALOGLM to introduce more dialogue-related knowledge. For the second issue, we leverage hybrid attention method in the Transformer architecture.

Figure 2 depicts the hybrid attention approach for our model. When dealing with long sequences, encoder self-attention accounts for the largest computational overhead, so we improve it with the recently proposed sparse Sinkhorn attention (Tay et al. 2020; Huang et al. 2021). Local attention method such as block-based attention divide the input into several blocks and restricts the words to only attend to the words in their own block. This greatly reduces the computational burden but also loses global information. Sinkhorn attention extends this by additionally introducing a differentiable sorting network. It sorts the original blocks in a new order, and allows each block to not only attend to itself, but also to attend to the corresponding block in the new order. As shown in Figure 2, the green block can attend to the yellow block because yellow block has the same position with green block after permutation. With Sinkhorn attention, different layers learn different permutations, so each block can access information in multiple locations on different layers.

However, full dialogue semantics is still indispensable for many applications such as text summarization. Therefore, we keep self-attention of several encoder layers unchanged. In other words, we still use full self-attention in these layers. This hybrid manner enables the interaction of local and global information. Compared with the model that does not introduce sparse attention, it can achieve similar or better performance under the premise of inputting a longer sequence and reducing training time. It is worth noting that the pre-training task and model modifications we proposed are orthogonal to all Transformer-based pre-trained models. In this paper, we initialize our model with the base version of UNILMV2 (Bao et al. 2020a). And the 4th, 8th and 12th encoder layers of UNILMV2 are kept with full self-attention.

Experiments

Implementation Details

To pre-train DIALOGLM, we further train UNILM with the window-based denoising framework for total 200,000 steps on dialogue data, of which 20,000 are warmup steps. We set batch size to 64 and the maximum learning rate to 2e-5. Pre-training data is the combination of MediaSum dataset (Zhu et al. 2021) and OpenSubtitles Corpus (Lison and Tiedemann 2016) (see Table 2). MediaSum is a media interview dataset consisting of 463.6K transcripts. OpenSubtitles[1] is compiled from a large database of movie and TV subtitles across 60 languages. We use the English part as the pre-training corpus. These two large-scale pre-training datasets contain a wealth of long dialogues with multiple participants and have clear dialogic text structures. During pre-training, the window size is set to 10% of the input length, and the maximum window size is limited to 512 tokens. When generating a noisy window, we first mask 50% of the speakers, then randomly inject *Turn Splitting* or *Turn Merging*, and utilize *Text Infilling* to mask 15% tokens in the utterances. Finally, *Turn Permutation* is performed. 8 A100 GPUs with 40GB memory are used to complete the experiments in this paper. All the results listed in this paper are the average of 3 runs. We pre-train two versions of DIALOGLM:

[1]This corpus is crawled from http://www.opensubtitles.org

Datasets	# Dialogues	# Turns	# Speakers	# Len. of Dialo.	Task	Domain
AMI	137	535.6	4.0	5,570.4	Summarization & TS	Meeting
ICSI	59	819.0	6.3	8,567.7	Summarization & TS	Meeting
QMSum	232	556.8	9.2	12,026.3	Summarization & QA & TS	Meeting
ForeverDreaming	4,348	330.7	-	7,653.5	Summarization	Screenplay
TVMegaSite	22,503	327.0	-	6,466.6	Summarization	Screenplay
MediaSum	463,596	30.0	6.5	1,927.2	Pre-training Data	Interview
OpenSubtitles	138,086	760.0	-	5,615.3	Pre-training Data	TV show

Table 2: Statistics of our pre-training data and downstream datasets. QA represents the abstractive question answering task and TS indicates the topic segmentation task.

DIALOGLM is obtained by further pre-training UNILM-base with the window-based denoising method. Its maximum input length is 5,120 and the tokens exceeding this length is truncated in the experiments.

DIALOGLM-sparse additionally introduces the hybrid attention approach in the pre-training process of DIALOGLM, so its maximum length is increased to 8,192 tokens.

Downstream Tasks and Datasets

After pre-training, we apply DIALOGLM to three different long dialogue tasks over meeting and screenplay domains, which includes five datasets in total.

Tasks

Long Dialogue Summarization: Given a long conversation ($>$ 5,000 words), output a concise summary ($<$ 512 words) containing its core content.

Abstractive Question Answering: Given a long dialogue and a specific question, generate several sentences as the answer based on relevant content in the dialogue.

Topic Segmentation: Given a long conversation, segment it into multiple parts based on their main topics. Each segment is consist of multiple consecutive utterances.

Datasets

We employ five popular benchmarks for the above tasks: `AMI` (Carletta et al. 2005), `ICSI` (Janin et al. 2003), `QMSum` (Zhong et al. 2021), `ForeverDreaming` and `TVMegaSite` (Chen et al. 2021a). Detailed statistics are given in Table 2. As shown, these datasets can be divided into two domains: meeting and screenplay.

Meeting Domain: `AMI` and `ICSI` are meeting transcripts collected from product design meetings in company and academic group meetings in school, respectively. For each meeting, it contains a meeting summary and human-annotated topic boundaries. `QMSum` is a benchmark for query-based multi-domain meeting summarization task. Query type in this dataset can be divided into general query and specific query. The former can be used as the summarization task and the latter can be regarded as abstractive QA task. It also contains human-annotated topic boundaries.

Screenplay Domain: `ForeverDreaming` and `TVMegaSite` are composed of pairs of TV series transcripts and human-written recaps. They have different dialogue styles from different sources, thus can serve as a challenging testbed for abstractive dialogue summarization.

Baselines

We compare DIALOGLM with strong baselines as follows:

UNILM-CP refers to the UNILM which is further trained using its original pre-training objective on MediaSum and OpenSubtitles corpora. The comparison with it can show whether the pre-training framework we proposed is more effective in long conversation scene.

Longformer is a powerful pre-trained model for long sequence processing. We report the results of the Longformer-based model in the previous work on different datasets[2]. The variant model Longformer-BART-arg (Fabbri et al. 2021) is initialized with the BART-large-CNN[3] parameters and uses argument-mining-based source as input.

BART is the state-of-the-art denoising sequence-to-sequence pre-trained model for various generation tasks. We use BART-large in all the experiments.

HAT-BART (Rohde, Wu, and Liu 2021) is a new hierarchical attention Transformer-based architecture that outperforms standard Transformers on several seq2seq tasks.

HMNET (Zhu et al. 2020) is the state-of-the-art meeting summarization model. It has hierarchical structure and utilizes cross-domain pre-training to recognize the special format of dialogues.

DDAMS (Feng et al. 2020) is a dialogue discourse-aware summarization model, which utilizes a relational graph encoder to explicitly model the interaction between utterances in a meeting by modeling different discourse relations.

Hybrid Model (Chen et al. 2021a) first extracts salient information (up to 1,024 words) from the dialogue and produces summary using BART.

Experimental Results

Results on Meeting Domain In this domain, we experiment on two popular summarization datasets `AMI` and `ICSI`, and the query-based summarization benchmark `QMSum`. Since most of the queries are specific on meeting content, `QMSum` can also be considered as an abstractive question answering task. We use ROUGE (Lin 2004) as the evaluation metric.

[2]The results of the Longformer-based model in `AMI` and `ICSI` are from (Fabbri et al. 2021), and the results of it in `QMSum` come from (Zhang et al. 2021b). In screenplay domain, the results of Longformer are from Chen et al. (2021a).

[3]BART-large-CNN refers to further fine-tuning BART-large on the news summarization dataset CNN/DailyMail.

Models	Meeting Summarization						Abstractive QA		
	AMI			ICSI			QMSum		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
PGNET ($l = 2,048$)	42.60	14.01	22.62*	35.89	6.92	15.67*	28.74	5.98	25.13
HMNET ($l = 8,192$)	53.02	18.57	-	46.28	10.60	-	32.29	8.67	28.17
DDAMS ($l = 15,000$)	53.15	**22.32**	25.67*	40.41	11.02	19.18*	-	-	-
BART-large ($l = 3,072$)	51.77	18.83	49.67	46.23	10.17	44.83	32.16	8.01	27.72
HAT-BART ($l = 3,072$)	52.27	20.15	50.57	43.98	10.83	41.36			
Longfomer ($l = 8,192$)	54.20	20.72	51.36	43.03	12.14	40.26	31.60	7.80	20.50
Longformer-BART-arg ($l = 8,192$)	54.47	20.83	51.74	44.17	11.67	41.33	-	-	-
UNILM-base ($l = 5,120$)	51.92	18.42	49.89	46.75	11.39	45.13	29.14	6.25	25.46
UNILM-CP ($l = 5,120$)	52.67	19.33	50.55	48.43	12.93	46.24	29.19	6.73	25.52
DIALOGLM ($l = 5,120$)	**54.49**	20.03	**51.92**	49.25	12.31	46.80	**34.02**	9.19	29.77
DIALOGLM-sparse ($l = 8,192$)	53.72	19.61	51.83	**49.56**	**12.53**	**47.08**	33.69	**9.32**	**30.01**

Table 3: Experimental results on meeting-style datasets. l is the maximum number of input tokens for the corresponding model. Results with * indicate that ROUGE-L is calculated without sentence splitting.

As illustrated in Table 3, DIALOGLM achieves state-of-the-art results on most metrics across datasets. Compared to the backbone model UNILM-base, our proposed pre-train framework brings clear improvements on all the three datasets. Specifically, on AMI and ICSI, DIALOGLM increases the ROUGE-1 score by more than 2.5 (51.92→54.49 and 46.75→49.25), and this improvement reaches to about 5.0 on QMSum (29.14→34.02. This demonstrates that window-based denoising approach can assist models to further understand the content of long dialogues, and hence to produce better summaries or answers. However, when returning to the original objective of UNILM (i.e., UNILM-CP), further pre-training does not yield substantial gains. It indicates that general-purpose pre-training objective limits model's ability to comprehend long dialogues. In comparison with previous state-of-the-art models, the benefits of our model are more pronounced in datasets with longer input text (ICSI and QMSum). For example, DIALOGLM outperforms HMNET by 2.97 ROUGE-1 points on ICSI and 1.73 R-1 points on QMSum, which indicates that DIALOGLM, with further pre-training, can be a powerful tool for dealing with long conversation scenarios.

In addition, DIALOGLM-sparse is able to process longer input with the same memory consumption after introducing hybrid attention mechanism. It not only reduces the training time, but also performs better on longer meetings. In AMI dataset, since the average length of meetings is about 5,000 (see Table 2), DIALOGLM is already competent for dialogues of this length. As a result, the introduction of hybrid attention has a slightly negative impact on the performance. However, on ICSI and QMSum datasets, as the meeting length exceeds 5,000 words, which is also the common length of one-to-two-hour meetings in real applications, DIALOGLM-sparse progressively reveals its benefits. It obtains the state-of-the-art ROUGE-2 and ROUGE-L scores on these two datasets, and surpasses DIALOGLM by about 0.2 points in these metrics. Therefore, the proposed hybrid attention is genuinely required and can benefit realistic long dialogue scenes.

The results of topic segmentation task on AMI and QMSum are listed in Table 4. A long conversation comprises several segments based on their topics, and topic segmentation is the task to discover these boundaries. In our experiment, we treat it as a turn-level binary classification task, where a positive label indicates that this turn is the end of a main topic. We follow Liu and Lapata (2019) to insert a special token [CLS] to the start of each each turn, and utilize the hidden state of [CLS] as turn-level representation for classification.

Models	AMI		QMSum	
	Pk	Wd	Pk	Wd
RANDOM	0.47	0.65	0.52	0.70
EVEN	0.52	0.72	0.56	0.59
UNILM-base ($l = 5,120$)	0.44	0.57	0.49	0.56
UNILM-CP ($l = 5,120$)	0.43	0.50	0.47	0.53
DIALOGLM ($l = 5,120$)	0.38	0.39	0.44	0.48
DIALOGLM-sparse ($l = 8,192$)	**0.36**	**0.34**	**0.38**	**0.40**

Table 4: Experimental results on dialogue segmentation task.

We use standard metrics *Pk* (Beeferman, Berger, and Lafferty 1999) and *WinDiff* (Pevzner and Hearst 2002) to evaluate segmentation models. Lower scores of these two metrics indicate that the predicted segmentation is closer to the ground truth. RANDOM is the baseline that randomly selects boundary points throughout the conversation, while EVEN segments the whole dialogue evenly. The results on two datasets show the same trend: dialogue-related pre-training can boost performance, and the hybrid attention augmentation could further improve upon this.

Results on Screenplay Domain In this domain, plot details emerge indirectly in character dialogues and are scattered throughout the full transcript. To succinct plot summary, the model needs to locate and incorporate these aspects in the long dialogue.

As shown in Table 5, similar to the meeting domain, our two models achieve superior performance across the majority of criteria. On both ForeverDreaming and TVMegaSite, DIALOGLM-sparse outperforms BART-Large by 1.93 and 2.04 ROUGE-1 points (33.82→35.75 and 43.54→45.58), respectively. This improvement becomes more prominent when compared to UNILM-base which

Models	ForeverDreaming			TVMegaSite		
	R-1	R-2	R-L	R-1	R-2	R-L
Longformer	25.90	4.20	23.80	42.90	**11.90**	41.60
Hybrid Model	25.30	3.90	23.10	38.80	10.20	36.90
BART-large	33.82	7.48	29.07	43.54	10.31	41.35
UNILM-base	32.16	5.93	27.27	43.42	9.62	41.19
UNILM-CP	33.29	6.74	28.21	44.07	9.96	41.73
DIALOGLM	35.42	8.23	30.61	45.04	10.45	42.71
DIALOGLM-sparse	**35.75**	**8.27**	**30.76**	**45.58**	10.75	**43.31**

Table 5: Experimental results on screenplay-style datasets: ForeverDreaming and TVMegaSite.

has no dialogue-oriented pre-training and sparse attention. With merely dialogue pre-training data and no dialogue-specific pre-training framework, UNILM-CP can be enhanced but still inferior to DIALOGLM. On the other hand, DIALOGLM-sparse consistently outperforms DIALOGLM in these datasets. We believe this is due to the relatively long length of screenplay (over 6,000 words, see Table 2), and the tail part of its transcript is still critical. Models that can deal with longer sequences, such as DIALOGLM-sparse, are able to capture more dialogue content and storylines, leading to further improvement in a variety of circumstances.

Ablation Study To better understand the contribution of each component in our pre-trained model, we conduct comprehensive ablation studies on `QMSum` and `TVMegaSite`, which can be viewed as representatives of the meeting and screenplay domain respectively. Overall, our proposed pre-training framework, i.e. window-based denoising, greatly strengthens the capacity of the general-purpose pre-trained model to process long conversations. It is reflected in Table 6: removing "Pre-train" results in substantial performance degradation on both datasets. Furthermore, all five dialog-inspired noises contribute to the pre-training process. The most important of these are *Turn Split* and *Turn Merging*. We think this is because the model can not denoise them without being aware of the dialogic structure and the main content. *Speaker Mask* brings the least benefit, due to the restoration of other noises also implicitly requires the model to figure out who is the interlocutor of each turn. Additionally, whether to introduce the hybrid attention mechanism allows our model to be flexibly applicable to more scenarios, so we release two versions of DIALOGLM to accommodate the circumstance of varying durations of dialogues.

Human Evaluation As factual inconsistency between the source document and the generated sequence is critical in the dialogue domain (Zhong et al. 2021), it is indispensable to evaluate the abstractive model manually. Specifically, we ask 10 graduate students to rank four different summaries (generated by UNILM-base, BART, DIALOGLM-sparse and reference summary) according to three metrics: fluency, informativeness and faithfulness to the original dialogue. Ranking first means the best performance on this metric. We randomly select 30 samples from each test set of `QMSum` and `TVMegaSite` for human evaluation. The results are provided in Table 7. From the perspective of fluency, after

Model	QMSum	TVMegaSite
DIALOGLM-sparse	33.69	**45.58**
- Sparse Attention	**34.02**	45.04
- Pre-train	29.14	43.42
- Speaker Mask	33.52	45.31
- Turn Splitting / Merging	32.76	44.23
- Text Infilling	33.27	44.79
- Turn Permutation	33.22	44.64

Table 6: Ablation study of DIALOGLM (ROUGE-1 score). '-' means we remove the module from the original model.

Models	QMSum			TVMegaSite		
	Ful.	Info.	Faith.	Ful.	Info.	Faith.
UNILM	3.20	3.07	3.13	2.83	3.23	2.63
BART	2.67	3.27	3.43	**2.43**	2.73	2.50
DIALOGLM	**2.43**	**2.37**	**2.20**	2.53	**2.37**	**2.13**
Ref. Sum.	1.70	1.30	1.23	2.20	1.67	2.73

Table 7: Results of human evaluation by ranking. Ful., Info., and Faith. represent fluency, infomativeness and faithfulness to the original dialogue, respectively. Ref. Sum. refers to the human-annotated reference summary.

further pre-training, DIALOGLM can output more coherent sentences than UNILM, and it is comparable to BART. However, the performance of all neural models is still far from the human-annotated answers or summaries. In terms of the other two metrics, DIALOGLM is substantially more informative and reliable when comparing to the prior state-of-the-art model BART. The capacity of DIALOGLM to better grasp the structure and content of conversations is largely responsible for this improvement. Regarding the reference summary, it obtains high scores on `QMSum`, but performs poorly on the faithfulness of `TVMegaSite`. This is because for screenplay domain, some useful information is displayed in the video rather than in the dialogue transcript. For example, if two people are eating in a restaurant, we can easily see this from the video of the TV series, but this may not be present explicitly in the transcript. It leads to the fact that the reference summaries written by human can be informative but contain details that are not visible in the conversation, which are regarded as unfaithful to the original dialogue.

Conclusion

In this paper, we propose a novel pre-training framework for long dialogue understanding and summarization. In particular, given a long conversation, we substitute a portion of it with a noisy window comprising five dialogue-inspired noises, and let the model generate the original dialogue window. As a consequence, the pre-trained model can efficiently realize the dialogic structure and capture the essential information, allowing it to reconstruct any section of the conversation. Moreover, we present a hybrid attention approach to adapt to longer dialogue scenarios. Experiments show that our pre-trained model DIALOGLM outperforms the previous state-of-the-art models on five benchmarks with three long dialogue understanding and summarization tasks.

References

Bao, H.; Dong, L.; Wei, F.; Wang, W.; Yang, N.; Liu, X.; Wang, Y.; Gao, J.; Piao, S.; Zhou, M.; et al. 2020a. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International Conference on Machine Learning*, 642–652. PMLR.

Bao, S.; He, H.; Wang, F.; Wu, H.; and Wang, H. 2020b. PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 85–96.

Beeferman, D.; Berger, A.; and Lafferty, J. 1999. Statistical Models for Text Segmentation. *Machine Learning*, 1(34): 177–210.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Cao, Y.; Bi, W.; Fang, M.; and Tao, D. 2020. Pretrained Language Models for Dialogue Generation with Multiple Input Sources. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 909–917.

Carletta, J.; Ashby, S.; Bourban, S.; Flynn, M.; Guillemot, M.; Hain, T.; Kadlec, J.; Karaiskos, V.; Kraaij, W.; Kronenthal, M.; et al. 2005. The AMI meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, 28–39. Springer.

Chen, M.; Chu, Z.; Wiseman, S.; and Gimpel, K. 2021a. SummScreen: A Dataset for Abstractive Screenplay Summarization. *arXiv preprint arXiv:2104.07091*.

Chen, Y.; Liu, Y.; Chen, L.; and Zhang, Y. 2021b. DialSumm: A Real-Life Scenario Dialogue Summarization Dataset. *arXiv preprint arXiv:2105.06762*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; and Hon, H.-W. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. *Advances in Neural Information Processing Systems*, 32: 13063–13075.

Fabbri, A. R.; Rahman, F.; Rizvi, I.; Wang, B.; Li, H.; Mehdad, Y.; and Radev, D. R. 2021. ConvoSumm: Conversation Summarization Benchmark and Improved Abstractive Summarization with Argument Mining. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 6866–6880. Association for Computational Linguistics.

Feng, X.; Feng, X.; and Qin, B. 2021. A Survey on Dialogue Summarization: Recent Advances and New Frontiers. *arXiv preprint arXiv:2107.03175*.

Feng, X.; Feng, X.; Qin, B.; Geng, X.; and Liu, T. 2020. Dialogue discourse-aware graph convolutional networks for abstractive meeting summarization. *arXiv preprint arXiv:2012.03502*.

Feng, X.; Feng, X.; Qin, L.; Qin, B.; and Liu, T. 2021. Language Model as an Annotator: Exploring DialoGPT for Dialogue Summarization. *arXiv preprint arXiv:2105.12544*.

Gao, X.; Zhang, Y.; Galley, M.; Brockett, C.; and Dolan, W. B. 2020. Dialogue Response Ranking Training with Large-Scale Human Feedback Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 386–395.

Gu, J.-C.; Tao, C.; Ling, Z.-H.; Xu, C.; Geng, X.; and Jiang, D. 2021. MPC-BERT: A Pre-Trained Language Model for Multi-Party Conversation Understanding. *arXiv preprint arXiv:2106.01541*.

Henderson, M.; Casanueva, I.; Mrkšić, N.; Su, P.-H.; Wen, T.-H.; and Vulić, I. 2020. ConveRT: Efficient and Accurate Conversational Representations from Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2161–2174.

Hsueh, P.-Y.; Moore, J. D.; and Renals, S. 2006. Automatic segmentation of multiparty dialogue. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Huang, L.; Cao, S.; Parulian, N.; Ji, H.; and Wang, L. 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*, 1419–1436.

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

Kitaev, N.; Kaiser, L.; and Levskaya, A. 2020. Reformer: The Efficient Transformer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Koay, J. J.; Roustai, A.; Dai, X.; and Liu, F. 2021. A Sliding-Window Approach to Automatic Creation of Meeting Minutes. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 68–75.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 7871–7880. Association for Computational Linguistics.

Li, J.; Zhang, Z.; Zhao, H.; Zhou, X.; and Zhou, X. 2020. Task-specific objectives of pre-trained language models for dialogue adaptation. *arXiv preprint arXiv:2009.04984*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lison, P.; and Tiedemann, J. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 923–929.

Liu, Y.; and Lapata, M. 2019. Text Summarization with Pretrained Encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3730–3740.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Pevzner, L.; and Hearst, M. A. 2002. A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*, 28(1): 19–36.

Rohde, T.; Wu, X.; and Liu, Y. 2021. Hierarchical Learning for Generation with Long Source Sequences. *arXiv preprint arXiv:2104.07545*.

Roy, A.; Saffar, M.; Vaswani, A.; and Grangier, D. 2021. Efficient Content-Based Sparse Attention with Routing Transformers. *Transactions of the Association for Computational Linguistics*, 9: 53–68.

Tay, Y.; Bahri, D.; Yang, L.; Metzler, D.; and Juan, D.-C. 2020. Sparse sinkhorn attention. In *International Conference on Machine Learning*, 9438–9447. PMLR.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, S.; Li, B. Z.; Khabsa, M.; Fang, H.; and Ma, H. 2020a. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Wang, S.; Zhou, L.; Gan, Z.; Chen, Y.; Fang, Y.; Sun, S.; Cheng, Y.; and Liu, J. 2021. Cluster-Former: Clustering-based Sparse Transformer for Question Answering. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, 3958–3968. Association for Computational Linguistics.

Wang, Y.; Rong, W.; Zhang, J.; Ouyang, Y.; and Xiong, Z. 2020b. Knowledge Grounded Pre-Trained Model For Dialogue Response Generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Wu, C.-S.; Hoi, S. C.; Socher, R.; and Xiong, C. 2020. TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 917–929.

Zaheer, M.; Guruganesh, G.; Dubey, K. A.; Ainslie, J.; Alberti, C.; Ontañón, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; and Ahmed, A. 2020. Big Bird: Transformers for Longer Sequences. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zhang, H.; Gong, Y.; Shen, Y.; Li, W.; Lv, J.; Duan, N.; and Chen, W. 2021a. Poolingformer: Long Document Modeling with Pooling Attention. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 12437–12446. PMLR.

Zhang, J.; Zhao, Y.; Saleh, M.; and Liu, P. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, 11328–11339. PMLR.

Zhang, Y.; Ni, A.; Yu, T.; Zhang, R.; Zhu, C.; Deb, B.; Celikyilmaz, A.; Hassan, A.; and Radev, D. 2021b. An Exploratory Study on Long Dialogue Summarization: What Works and What's Next. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Findings*.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, W. B. 2020b. DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 270–278.

Zhong, M.; Yin, D.; Yu, T.; Zaidi, A.; Mutuma, M.; Jha, R.; Hassan, A.; Celikyilmaz, A.; Liu, Y.; Qiu, X.; et al. 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*, 5905–5921.

Zhu, C.; Liu, Y.; Mei, J.; and Zeng, M. 2021. MediaSum: A Large-scale Media Interview Dataset for Dialogue Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5927–5934.

Zhu, C.; Xu, R.; Zeng, M.; and Huang, X. 2020. A Hierarchical Network for Abstractive Meeting Summarization with Cross-Domain Pretraining. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 194–203.

Zou, Y.; Lin, J.; Zhao, L.; Kang, Y.; Jiang, Z.; Sun, C.; Zhang, Q.; Huang, X.; and Liu, X. 2021a. Unsupervised Summarization for Chat Logs with Topic-Oriented Ranking and Context-Aware Auto-Encoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14674–14682.

Zou, Y.; Zhao, L.; Kang, Y.; Lin, J.; Peng, M.; Jiang, Z.; Sun, C.; Zhang, Q.; Huang, X.; and Liu, X. 2021b. Topic-Oriented Spoken Dialogue Summarization for Customer Service with Saliency-Aware Topic Modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14665–14673.