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

Ming Zhong\textsuperscript{1}, Yang Liu\textsuperscript{2}, Yichong Xu\textsuperscript{2}, Chenguang Zhu\textsuperscript{2}, Michael Zeng\textsuperscript{2}

\textsuperscript{1}University of Illinois at Urbana-Champaign
\textsuperscript{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. Source code and all the pre-trained models are available on our GitHub repository (https://github.com/microsoft/DIALOGLM).

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; Kooy et al. 2021). 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 2019; Wang et al. 2020) 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 com-
posed of multiple turns contains more coherent and informative 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), 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. 2020) 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 2019; 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 information and retain part of full attention for global information. We adopt the last two approaches in this paper to reduce...
In this section, we first introduce the pre-training task for DialogLMS: window-based denoising and five types of
dialogue-inspired noises. Then we describe the overall ar-
chitecture of our pre-trained model.

Window-based Denoising
A long conversation usually involves a core theme and mul-
tiple 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
may portray the speaking style of different people, e.g., prod-
uct 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 co-
herent 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 DialogLMS: 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 contain-
ing 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 re-
store this selected window \( W \) by modeling the condi-
tional 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 mask-
ning 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 be-
ing unable to recover it. On the other hand, unlike docu-
ments, a large number of individual turns in a conversa-
tion 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 succes-
sive turns is more likely to contain meaningful and coher-
ent 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 de-
sign the following five types of noise (see Table 1):

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>Original Dialogue</th>
<th>Noisy Dialogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker Mask</td>
<td>Tom: The weather is good today!</td>
<td>[MASK]: The weather is good today!</td>
</tr>
<tr>
<td>Turn Splitting</td>
<td>Tom: The weather is good today!</td>
<td>[MASK]: Do you have any plans?</td>
</tr>
<tr>
<td>Text Infilling</td>
<td>Tom: The weather is good today!</td>
<td>[MASK]: Do you have any plans?</td>
</tr>
<tr>
<td>Turn Permutation</td>
<td>Tom: The weather is good today!</td>
<td>[MASK]: How about we go to play basketball?</td>
</tr>
</tbody>
</table>

Table 1: Dialogue-related noise for generating a noisy window. The bold [MASK] token means that a speaker name is masked,
and the unbolded [MASK] token indicates that a text span in the utterance is masked.
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. OpenSubtitles1 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: DIALOGLM is obtained by further pre-training UniLM-base with the window-based denoising method. Its maxi-

---

`1This corpus is crawled from http://www.opensubtitles.org`
After pre-training, we apply DIALOG 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 model, which utilizes a relational graph encoder to explicitly model the interaction between utterances in a meeting by modeling different discourse relations.
- **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 DIALOG 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. The variant model Longformer-BART-arg (Fabbri et al. 2021) is initialized with the BART-large-CNN 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.

As illustrated in Table 3, DIALOGM achieves state-of-the-art results on most metrics across datasets. Compared

### 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.

<table>
<thead>
<tr>
<th>Datasets</th>
<th># Dialogues</th>
<th># Turns</th>
<th># Speakers</th>
<th># Len. of Dialo.</th>
<th>Task</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMI</td>
<td>137</td>
<td>535.6</td>
<td>4.0</td>
<td>5,570.4</td>
<td>Summarization &amp; TS</td>
<td>Meeting</td>
</tr>
<tr>
<td>ICSI</td>
<td>50</td>
<td>819.0</td>
<td>6.3</td>
<td>8,567.7</td>
<td>Summarization &amp; TS</td>
<td>Meeting</td>
</tr>
<tr>
<td>QMSum</td>
<td>232</td>
<td>556.8</td>
<td>9.2</td>
<td>12,026.3</td>
<td>Summarization &amp; QA &amp; TS</td>
<td>Meeting</td>
</tr>
<tr>
<td>ForeverDreaming</td>
<td>4,348</td>
<td>330.7</td>
<td>-</td>
<td>7,653.5</td>
<td>Summarization</td>
<td>Screenplay</td>
</tr>
<tr>
<td>TVMegaSite</td>
<td>22,503</td>
<td>327.0</td>
<td>-</td>
<td>6,466.6</td>
<td>Summarization</td>
<td>Screenplay</td>
</tr>
<tr>
<td>MediaSum</td>
<td>463,596</td>
<td>30.0</td>
<td>6.5</td>
<td>1,927.2</td>
<td>Pre-training Data</td>
<td>Interview</td>
</tr>
<tr>
<td>OpenSubtitles</td>
<td>138,086</td>
<td>760.0</td>
<td>-</td>
<td>5,615.3</td>
<td>Pre-training Data</td>
<td>TV show</td>
</tr>
</tbody>
</table>

2The 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).
Table 3: Experimental results on meeting-style datasets. Results with * indicate that ROUGE-L is calculated without sentence splitting.

<table>
<thead>
<tr>
<th>Models</th>
<th>Meeting Summarization</th>
<th>Abstractive QA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AMI</td>
<td>ICSI</td>
</tr>
<tr>
<td></td>
<td>R-1</td>
<td>R-2</td>
</tr>
<tr>
<td>PGNET ($l = 2,048$)</td>
<td>42.60</td>
<td>14.01</td>
</tr>
<tr>
<td>HMNET ($l = 8,192$)</td>
<td>53.02</td>
<td>18.57</td>
</tr>
<tr>
<td>DDAms ($l = 15,000$)</td>
<td>53.15</td>
<td><strong>22.32</strong></td>
</tr>
<tr>
<td>BART-large ($l = 3,072$)</td>
<td>51.77</td>
<td>18.83</td>
</tr>
<tr>
<td>HAT-BART ($l = 3,072$)</td>
<td>52.27</td>
<td>20.15</td>
</tr>
<tr>
<td>Longformer ($l = 8,192$)</td>
<td>54.20</td>
<td>20.72</td>
</tr>
<tr>
<td>Longformer-BART-arg ($l = 8,192$)</td>
<td>54.47</td>
<td>20.83</td>
</tr>
<tr>
<td>UNILM-base ($l = 5,120$)</td>
<td>51.92</td>
<td>18.42</td>
</tr>
<tr>
<td>UNILM-CP ($l = 5,120$)</td>
<td>52.67</td>
<td>19.33</td>
</tr>
<tr>
<td>DIALOGLM ($l = 5,120$)</td>
<td><strong>54.49</strong></td>
<td>20.03</td>
</tr>
<tr>
<td>DIALOGLM-sparse ($l = 8,192$)</td>
<td>53.72</td>
<td>19.61</td>
</tr>
</tbody>
</table>

Table 4: Experimental results on dialogue segmentation task.

<table>
<thead>
<tr>
<th>Models</th>
<th>AMI</th>
<th>QSum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-1</td>
<td>R-2</td>
</tr>
<tr>
<td>RANDOM</td>
<td>0.47</td>
<td>0.65</td>
</tr>
<tr>
<td>EVEN</td>
<td>0.52</td>
<td>0.72</td>
</tr>
<tr>
<td>UNILM-base ($l = 5,120$)</td>
<td>0.44</td>
<td>0.57</td>
</tr>
<tr>
<td>UNILM-CP ($l = 5,120$)</td>
<td>0.43</td>
<td>0.50</td>
</tr>
<tr>
<td>DIALOGLM ($l = 5,120$)</td>
<td>0.38</td>
<td>0.39</td>
</tr>
<tr>
<td>DIALOGLM-sparse ($l = 8,192$)</td>
<td>0.36</td>
<td><strong>0.34</strong></td>
</tr>
</tbody>
</table>

Results on Screenplay Domain In this domain, plot details emerge indirectly in character dialogues and are scattered throughout the full transcript. To succinctly summarize, 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 has no dialogue-oriented pre-training and sparse attention. With merely dia-

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 (ICI 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 turn, and utilize the hidden state of [CLS] as turn-level representation for classification.

We use standard metrics $P_k$ (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.
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 grasp the structure and content of conversations. Moreover, we present a hybrid attention approach to strengthen the capacity of the general-purpose pre-trained framework, i.e., window-based denoising, greatly improves the performance of the model. The capacity of DIALOGLM to better grasp the structure and content of conversations is largely responsible for this improvement. 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 and faithful even in the absence of a specific conversation. The reference summaries provide context even if they are not accurate. In the future, we plan to extend the model to allow more dialogue content and storylines, leading to further improvement in a variety of circumstances.

### 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.
Acknowledgements

We thank Shaohan Huang for providing guidance on pre-training UniLM-CP. We would also like to thank annotators such as Liang Chen for their hard work and reviewers for their valuable comments.

References


Linformer: Self-attention with linear complexity.


