SPECTRUM: Speaker-Enhanced Pre-Training for Long Dialogue Summarization

Anonymous ACL submission

Abstract

Multi-turn dialogues are characterized by their extended length and the presence of turn-taking conversations. Traditional language models often overlook the distinct features of these dialogues by treating them as regular text. In this paper, we propose a speaker-enhanced pre-training method for long dialogue summarization, which leverages the inherent structure of multiple-turn dialogues. To support our study, we curate a diverse dataset that includes transcripts from real-world scenarios, movie or TV show transcripts, and dialogues generated by a Large Language Model. We then perform a pre-training, which encompasses the detection of speaker changes, and masked utterance generation. Experimental results of fine-tuned models demonstrate that our model achieves state-of-the-art performance on downstream benchmarks with long context, surpassing baseline models and highlighting the effectiveness of our approach. Our findings highlight the importance of curating pre-training datasets that exhibit diversity and variations in length distribution to ensure effective alignment with downstream datasets.

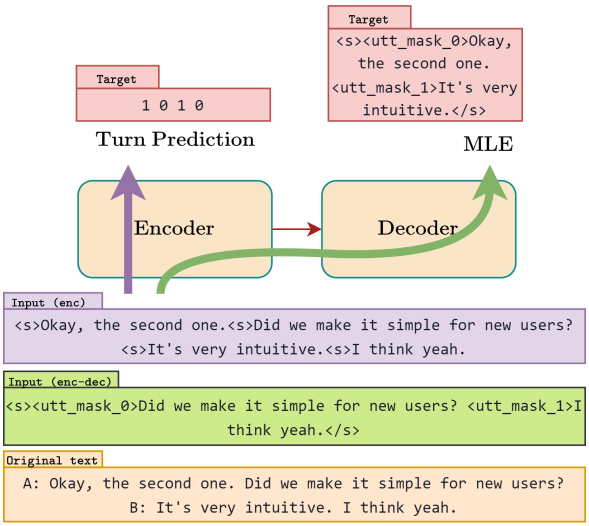

Figure 1: The proposed approach implements a speaker-enhanced learning objective within an encoder-decoder model. Two distinct forward paths are incorporated to enhance dialogue understanding. The first path involves only the encoder component, which predicts speaker turn switches. The second path integrates both the encoder and decoder components to generate masked sentences within dialogues, further improving dialogue comprehension.

1 Introduction

Dialogue summarization has been a challenging task in natural language processing. First, dialogue inherently includes redundant verbal expressions and often excludes certain mentions from the dialogue history. Consequently, dialogue summarization typically requires a broader context window to effectively handle complex long dependencies. Second, to gain a thorough comprehension of the content in a conversation, it is essential to take into account the details of the speaker and the signals from the sequential turns of utterances, as these factors significantly influence the interpretation of the information being shared. Thirdly, dialogue data varies greatly in terms of quality, length, and the presence or absence of speaker information. Finding a method to utilize all these signals is indeed a challenge.

To address the aforementioned issues, we introduce *SPECTRUM*, a pre-training approach that incorporates speaker information and leverages the inherent structure of multi-turn dialogues, aiming to enhance the effectiveness of models in interpreting dialogues. Unlike traditional language models that treat speaker information as plain text, our methodology incorporates the turn-based structure inherent to dialogues. Integrating speaker information improves the model's ability to focus on speaker-specific traits, allowing for enhanced understanding of long-context dialogues. To address long-context challenges, we utilize sparse attention networks as our backbone models, specifically LED (Beltagy et al., 2020) and PEGASUS-X (Phang et al., 2022). We collect multi-turn dia-

logues from diverse sources including two newly collected datasets from YouTube and IMDB Movie Quotes as well as existing public datasets for effective pre-training. We then classify them based on the presence or absence of turn-taking information and perform 2-stage pre-training. The second stage of pre-training is further tailored based on the diverse length distributions in the datasets.

Overall, our work contributes to the advancement of dialogue understanding by introducing a speaker-enhanced pre-training method and offering a curated dataset for pre-training. We thoroughly analyze the experimental outcomes and provide insights into best practices for pre-training across different downstream datasets. The empirical results validate the efficacy of our approach, showcasing its potential to enhance the performance of dialogue understanding models in various natural language processing applications. Our contributions are summarized as follows:

- We propose a two-stage multi-task pre-training approach to leverage the inherent multi-turn structure of dialogues.
- We curate pre-training data from diverse sources and explore the effects of varying sequence length distribution in the data.
- We achieve state-of-the-art performance compared to models of similar sizes on various datasets and elucidate effective pre-training practices for diverse downstream datasets.

2 Related Work

Speaker Information in Dialogue. Incorporating speaker information in dialogues has been explored in numerous studies. Leveraging contextual models to accommodate distinct speaker roles is a practical approach for enhancing the recognition of role patterns and aiding in the comprehension of dialogues (Chi et al., 2017; Liu et al., 2020a,b). Furthermore, Lei et al. (2021) utilizes the hierarchical model, Chi et al. (2017) is based on the contrastive learning method, and Gu et al. (2020) uses another speaker embedding representation to discern the relationships between speakers and their respective utterances. In contrast to solely designing enhanced contextual models, various pre-training methods are employed to improve dialogue comprehension and generation. For instance, Bao et al. (2020) utilizes multi-task pre-training objectives for dialogue response generation and selection, while Zhong et al. (2022) employs diverse masking schemes to effectively capture the dynamics within dialogues.

Efficient Transformers. Transformer (Vaswani et al., 2017) was purposed to solve the sequence-to-sequence tasks. However, it requires higher computation for longer sequences due to the self-attention operation being quadratic. Efficient transformers (Tay et al., 2020) are purposed to reduce the computation complexities while remain the same performance. In specific, ETC (Ainslie et al., 2020), BigBird (Zaheer et al., 2021), and Longformer (Beltagy et al., 2020) use local attentions with attached global positions to reduce the computation. The PEGASUS-X (Phang et al., 2022) uses block-wised attention. The LongT5 (Guo et al., 2022), on the other hand, implements Transient global attention for the problem. In this work, we employ Longformer-Encoder-Decoder(LED) (Beltagy et al., 2020) and PEGASUS-X (Phang et al., 2022) for handling long dialogues.

Pre-training Training Objectives. The training objectives and the corresponding singles are crucial for pre-training the models. For encoder-only models, BERT (Devlin et al., 2019) first introduces Masked Language Model Loss(MLM). Subsequently, various models utilize distinct self-supervision signals to enhance the understanding of contextualized semantics in documents. For example, ERNIE (Zhang et al., 2019) incorporates entity knowledge, while SpanBERT (Joshi et al., 2020) utilizes span boundaries. For the decoder-only model, GPT (Radford and Narasimhan, 2018; Radford et al., 2019) uses language model loss, which predicts the next tokens conditional on all previous tokens. Similar to the decoder-only model, encoder-decoder models employ conditional language model loss, in which the output token is also conditional on input tokens. More specifically, BART (Lewis et al., 2020) uses denoising learning objectives to let the model reconstruct the original text from the polluted input text. PEGASUS (Zhang et al., 2020), introduces sentence masking loss on the decoder side. T5 (Raffel et al., 2020) uses a complementary mask for the input and output. In this paper, we combined turn prediction and masked sentence generation as learning objectives for pre-training.

Long Dialogue Summarization. Typical Summarization dataset like SAMSum (Gliwa et al., 2019) contains only 94 input words on average. Similarly, DialogSum (Chen et al., 2021) contains only 137

input words on average. Unlike typical dialogue summarization datasets, long dialogue summarization data is way harder to collect. Thus long dialogue summarization data is usually on a small scale. QMSum (Zhong et al., 2021) contains 1,808 summaries for meetings with an average of 9,161 input words. AMI (McCowan et al., 2005) contains 137 summaries with around 6 thousand input words. ICSI (Janin et al., 2003) contains 59 summaries with more than 13 thousand averaged input words. SumScreen (Chen et al., 2022) contains an average of 5 to 6 thousand input tokens. In this study, our primary focus is on dialogues or meetings that extend beyond a few thousand words.

3 Pre-training Objectives

We formulate three learning objectives: Speaker Turn Prediction, Masked Language Modeling, and Masked Sentence Generation. Speaker Turn prediction enhances overall dialogue understanding, while MLM of coreferences, named entities, and verbs aims to enhance entity-level contextual representations in dialogues. Masked Sentence Generation serves the dual purpose of facilitating dialogue understanding and fostering the ability to generate coherent contextual content. Our pre-training methodology comprises two pathways: the first exclusively updates the encoder part of the model, whereas the second updates the entire encoder-decoder model. This pre-training approach improves the model's understanding of contextual dialogues and its ability to generate high-quality sequences aligned with the provided context.

3.1 Sentence Generation

We explore two masking strategies in pre-training for models to learn the context of dialogues: word-level and sentence-level. Initial experiments using the masking methods are conducted and the results are presented in Table 6.3.2. Based on the observation of the superior sentence-level masking strategy and drawing inspiration from the gap sentence generation method (Zhang et al., 2020), we choose to employ sentence-level masking from a single speaker. In Figure 1, represented by the green arrow, randomly selected sentences are replaced by `<utt_mask>` tokens. The target sequence comprises sequences of mask tokens preceding the selected sentences. Given the dialogues with masked tokens m, the objective is to generate the token sequence in the original sentence $t = [w_1, w_2, \dots]$ for m, where w_i is the i-th token in t. The loss of the generation step is computed as the negative log-likelihood of the conditional language generation:

$$L_{gen} = -\sum_i \log \Pr(w_i \mid w_1, \cdots, w_{i-1}; m) \quad (1)$$

3.2 Speaker Turn Prediction

Dialogues, being a natural form of conversation driven by multiple speakers, present an opportunity to enhance the model's understanding of the contextual nuances inherent in such interactions. In this study, we aim to teach the model to develop a more comprehensive understanding of dialogue context by explicitly providing speaker-turn information. By leveraging insights from previous works (Liu et al., 2022; Sangwoo Cho and Yu, 2022), which have demonstrated the benefits of segmenting articles for the summarization task, our model learns to predict turn switches as a means of enhancing dialogue comprehension. In Figure 1, illustrated by the purple arrow, we prepend `<s>` tokens to each sentence and utilize the resulting contextualized representations for turn switch prediction. To facilitate this learning objective, a binary classification approach is employed, leveraging the representations from the top encoder layer.

$$\hat{y}_i = \sigma(\mathbf{w}^\top \mathbf{h}_i + b) \quad (2)$$

where $\{\mathbf{h}_i\}_{i=1}^N$ is contextualized output representations for N sentences in the input dialogues. $\hat{y}_i$ is the turn prediction score and y_i is the ground-truth turn switch label, which can be readily obtained from the dialogues.

$$\mathcal{L}_{turn} = -\frac{1}{N}\sum_{i=1}^{N}\Big(y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\Big) \quad (3)$$

This objective can be effectively applied to data containing speaker turns, available in numerous real-world transcripts.

3.3 Masked Language Modeling (MLM)

We follow the practice of the BERT (Devlin et al., 2019) methodology. It is worth noting that dialogues often contain a higher frequency of pronouns, and important information can often be inferred from named entities or verbs. To utilize this observation, we preprocess the datasets by annotating coreferences, named entities, and verbs

	Dataset	Domain	Turn	#Turn	#Wds(Total)	#Wds	#Insts	#Wds (M)	#Insts (M)
S1	OpenSubtitles	Movie, TV show	X	-	2.857B	-	-	3146	908,068
	Gutenburg-dialogue	Books	X	-	295M	130	2,279,927	3400	86,882
S2	MediaSum	Interview	O	28.7	689M	1554	443,596	1088	633,547
	YouTube-news	News	O	16.4	50M	759	66,312	923	54,495
	YouTube-shows	Talk shows	O	42.4	26M	870	30,128	889	29,461
	Movie Quotes	Quotes	O	2.9	45M	38	1,181,196	918	49,398
	Soda	Social dialogue	O	7.6	146M	123	1,191,132	955	153,306

Table 1: Statistics of pre-training datasets. `S1` and `S2` represent datasets with and without turns, respectively. **#Wds** is the abbreviation for *number of words*. **#Insts** is the abbreviation for *number of samples*. **(M)** indicates the corresponding statistics after packing within each dataset. The max number of tokens is 4096 for S1 and 1400 for S2.

Dataset	Max Target Tks	#Tks-avg	min	max
S1	4096	4088.8	4080	4096
S2-short	1400	489.7	23	1400
S2-long	1400	1388.5	1381	1400
S2-both	1400	724.2	23	1400

Table 2: Statistics of various pre-training datasets. **#Tks** indicates the number of tokens.

as candidate mask tokens[1]. Initially, we apply MLM with two additional pre-training objectives. Nevertheless, following preliminary experiments (Sec. 6.3.1), we observe that MLM results in detrimental performance in downstream tasks. Consequently, we opt to exclude MLM from the final pre-training objective for our model, *SPECTRUM*.

The final pre-training objective is as follows:

$$\mathcal{L}(\Theta) = \mathcal{L}_{gen} + \beta \mathcal{L}_{turn} \quad (4)$$

where β is a coefficient that balances sentence-level cross-entropy losses and masked sentence generation and Θ is the model parameters.

4 Datasets

4.1 Pre-training Corpus

For our pre-training process, we utilize two sets of datasets: Stage1 (S1) without turn information; and Stage2 (S2) with turn information, as outlined in Table 1. The S1 data comprises the following two datasets:

OpenSubtitles (Lison and Tiedemann, 2016) This dataset consists of subtitles extracted from movies and TV shows. The English portion contains an extensive corpus of over 2.8 billion words.

Gutenberg-dialogue (Csaky and Recski, 2021) This dataset consists of dialogues extracted from public-domain books available through Project Gutenberg.

The S2 dataset consists of five diverse datasets, including:

[1] `en_core_web_sm` from SpaCy (Honnibal and Montani, 2017) with `neuralcoref` (Clark and Manning, 2016)

MediaSum (Zhu et al., 2021) This dataset comprises transcripts from media interviews, specifically from CNN and NPR. Speaker information is provided alongside corresponding utterances.

YouTube-news and YouTube-shows We have collected the datasets by crawling a variety of YouTube channels, such as news, sports news for YouTube-news; and variety shows, and talk shows for YouTube-shows. Turn switch signals are obtained from the transcripts.

Movie Quotes We have collected a dataset of 1.2 million memorable movie quotes from IMDB. This dataset includes the characters, utterances, and stage directions for each quote.

Soda (Kim et al., 2022) This dataset consists of socially-grounded dialogues distilled from GPT-3.5 (Ouyang et al., 2022), a large language model.

To enhance the model's capability to process longer inputs, we pack instances in each dataset up to a maximum of 4096 tokens for S1 and 1400 tokens (the average token count across all S2 datasets) for S2. The packed S2 dataset is referred to as S2-long. Nevertheless, we posit that consolidating multiple contexts from different dialogues into a single sequence may introduce confusion to the model. Therefore, we create a S2-short dataset concatenating all S2 datasets without packing to attain coherent context in each sequence. Finally, we construct S2-both by concatenating S2-short with the long-context S2-long dataset to incorporate diverse input sequence lengths. The statistics of each pre-training dataset are presented in Table 2. Note that we limit the max length of S2 datasets constrained to 1400, leading to the truncation of sequences in S2-short and S2-both that surpass this limit. Since the S1 dataset lacks turn information, the speaker turn prediction objective is exclusively applied to the S2 dataset. The experimental results of various pre-training datasets on downstream datasets are presented in Sec. 6.3.1.

Dataset	Domain	#Dialogues	#Wds	#Wds (S)
QMSum	Meetings	1,808	9161.1	69.7
AMI	Meetings	137	6007.7	296.6
ICSI	Meetings	59	13317.3	488.5
SS/FD	TV show	4,348	5930.2	101.2
SS/TMS	TV show	22,503	4599.6	344.2

Table 3: Statistics of summarization datasets based on dialogues. **#Wds** is the abbreviation for *number of words*. **(S)** indicates words in target summary. The input and target length are averaged across each dataset.

4.2 Summarization Datasets

Table 3 shows the statistics for the downstream datasets. We select long dialogue datasets in different domains given their inherent complexity and need for broad contextual understanding, suitably evaluating our pre-training strategy's effectiveness.

AMI and ICSI (McCowan et al., 2005; Janin et al., 2003) The meeting scripts used in datasets were generated by an Automatic Speech Recognition (ASR) system. The AMI corpus was collected from product design meetings in a company, while the ICSI corpus was collected from academic group meetings in a school.

QMSum (Zhong et al., 2021) is a dataset specifically designed for query-based meeting summarization. It encompasses meetings from three domains: AMI, ICSI, and committee meetings of the Welsh Parliament and Parliament of Canada. Each query and sample is written by domain experts.

SummScreen (Chen et al., 2022) consists of community-contributed transcripts of television show episodes obtained from The TVMegaSite, Inc. (TMS) and ForeverDream (FD). Each transcript is accompanied by a summary, which is either a recap sourced from TMS or a recap of the FD shows sourced from Wikipedia and TVMaze.

5 Experimental Setup

In our study, we employ sparse Transformer models, specifically LED (Beltagy et al., 2020) and PEGASUS-X (Phang et al., 2022), to handle long input dialogues efficiently. Considering the considerable computational requirements for pre-training, we employ the pre-trained checkpoint from PEGASUS-X (568M) and LED-large model (459M). The experiments are performed using the PEGASUS-X model unless stated otherwise.

5.1 Pre-training

In our pre-training process, we follow the general recipe outlined in PEGASUS (Zhang et al., 2020) for sentence generation. Our pre-training involves two stages, each utilizing distinct datasets. In the first stage, we utilize the S1 data, which contains approximately 3.3 times more words than the S2 data. This facilitates effective adaptation of the pre-trained models to dialogue data. The maximum lengths for input and output sequences are set at 4096 and 512 tokens, respectively, with an utterance masking ratio of 12% for effective training. We allocate 1% of the data for the validation set to monitor the training progress.

The second stage involves dual-path training, one for the encoder-only and the other for the encoder-decoder, to compute the joint loss (Eq. 4) on the S2 datasets. This approach allows for effective learning by improving the model's ability to understand dialogue sequences. In this stage, we employ a maximum of 1400 input tokens and 256 output tokens, maintaining a masking ratio of 18% to ensure that the ratio remains below 20%. We employ 1% data from the YouTube and Movie Quotes datasets for the validation set, while employing a standard validation set for the remaining datasets.

5.2 Fine-tuning

We evaluate our pre-trained models by fine-tuning on the downstream summarization datasets as listed in Table 3. Considering the average word counts for each dataset, we fine-tune *SPECTRUM* with an input length of up to 16,384 tokens for QMSum and ICSI, and 8,192 tokens for the remaining datasets. Due to the limited maximum input length (4096 for S1 and 1400 for S2) used in pre-trining, the parameters of positional embedding are repeated up to the respective maximum input lengths for fine-tuning. We configure the output length to 256 tokens for QMSum and SummScreen/FD, and 512 tokens for the other datasets. The turn prediction loss is not employed during the fine-tuning process. During inference, we employ a greedy search method to produce all summaries, thereby eliminating the necessity to introduce new variables during the decoding process.

We report the Rouge evaluation metric (Lin, 2004) by using the `rouge-score` library[2]. Our evaluation results include Rouge-1 (R-1), Rouge-2 (R-2), Rouge-L (R-L)[3], and the geometric average of the three Rouge scores (R-G).

[2] https://pypi.org/project/rouge-score/

[3] Rouge-L is computed without considering newlines. The summary-level Rouge-L score, which considers sentence splitting, is reported in other systems.

Model	*AMI*				*ICSI*				*QMSum*			
	R-1	**R-2**	**R-L**	**R-G**	**R-1**	**R-2**	**R-L**	**R-G**	**R-1**	**R-2**	**R-L**	**R-G**
PGNet	42.60	14.01	22.62	23.81	35.89	6.92	15.67	15.73	28.74	5.98	25.13*	-
BART+SCL	51.40	17.81	25.30	28.50	-	-	-	-	-	-	-	-
HMNet	52.36	18.63	24.00	28.61	45.97	10.14	18.54	20.52	32.29	8.67	28.17*	-
HAT-BART	52.27	20.15	50.57*	-	43.98	10.83	41.36*	-	-	-	-	-
DDAMS	53.15	**22.32**	25.67	31.23	40.41	11.02	19.18	20.44	-	-	-	-
Summ^N	53.44	20.30	51.39*	-	48.87	12.17	46.38*	-	34.03	9.28	29.48*	-
DialogLED	54.80	20.37	52.26*	-	50.11	13.23	47.25*	-	34.50	9.92	30.27*	-
DialogLED†	54.45	20.20	26.03	30.58	**50.28**	13.27	**20.16**	23.78	34.09	9.91	20.34	19.01
PGS-X (568M)	(50.57)	(19.53)	(25.14)	(29.17)	(48.65)	(13.40)	(18.26)	(22.83)	33.2	9.6	21.6	19.02
LT5$_{LG}$ (770M)	-	-	-	-	-	-	-	-	35.1	12	23.3	21.41[4]
LT5$_{XL}$ (3B)	-	-	-	-	-	-	-	-	34.9	11.8	23.5	21.31[4]
SPECTRUM	**55.12**	21.72	**26.73**	**31.75**	50.04	**14.5**	19.81	**24.31**	**35.86**	**12.34**	**23.53**	**21.84**

Table 4: ROUGE scores on meeting summarization datasets: AMI, ICSI, and QMSum. **R-G** is the geometric average of R-1, R-2 and R-L. * denote the ROUGE-L scores with sentence split. † indicates Rouge scores computed using outputs from the author's repository. The numbers in parentheses represent the fine-tuned results using PEGASUS-X. Our results are significantly better than the baseline models ($\rho < 0.05$). *SPECTRUM*is the same size of PGS-X.

Model	*SummScreen/FD*				*SummScreen/TMS*			
	R-1	**R-2**	**R-L**	**R-G**	**R-1**	**R-2**	**R-L**	**R-G**
BART-large	33.82	7.48	29.07*	-	43.54	10.31	41.35*	-
UniLM	33.29	6.74	28.21*	-	44.07	9.96	41.73*	-
Summ^N	32.48	6.12	27.14*	-	44.64	11.87	42.53*	-
DialogLED	36.70	8.68	31.38*	-	**45.22**	11.69	42.86*	-
DialogLED†	36.45	8.57	19.09	18.13	44.29	10.86	18.74	20.81
PGS-X	35.7	9.1	20.6	18.84	-	-	-	-
LT5$_{LG}$	35.6	9.2	21.2	19.07	-	-	-	-
LT5$_{XL}$	**35.8**	**9.6**	21.1	**19.35**	-	-	-	-
SPECTRUM	35.77	9.54	**21.22**	**19.35**	45.01	**12.21**	**19.11**	**21.90**

Table 5: ROUGE scores on SummScreen: FD (ForeverDreaming) and TMS (TV MegaSite).

5.3 Implementation Details

We implement our code with Huggingface-transformers [5] and Pytorch-lightning [6].

During pre-training, we train PEGASUS-X for 20 epochs with stage 1 and 2 strategies: S1 with sentence masking objective, S2 with joint objectives. We utilize 64 V100 GPUs with a total batch size of 256. We use Adam (Kingma and Ba, 2014) as our optimizer with a linearly decreased learning rate schedule with 5% of warmup. The maximum learning rate is $2e^{-4}$. We pre-train models with fp32 for stable training and use gradient-checkpointing for only stage 1 to save GPU memory. For the second stage of pre-training, we fix β=3. We experiment with an adaptive method by updating the parameter β at the end of each epoch, based on the average loss ratio between L_{gen} and L_{turn} of all steps. However, this approach underperforms in our experiments.

During fine-tuning, we keep the same settings of hyper-parameters except for the optimizer. For saving GPU memory, we use Adafactor (Shazeer and Stern, 2018) and apply a cosine learning rate decay schedule with 5% of warmup steps.

6 Results and Analysis

6.1 Baselines

We compare *SPECTRUM* with various baselines.

PGNet (See et al., 2017) uses copy mechanism to balance between accurate and novel words for the summarization task. **BART+SCL** (Geng et al., 2022) utilizes token, turn, and global-level of supervised contrastive learning on short and long dialogue benchmarks. **HMNet** (Zhu et al., 2020) employes hierarchical attention structure. It utilizes cross-domain pre-training for meeting summarization. **HAT-BART** (Rohde et al., 2021) introduces a new hierarchical attention Transformer-based architecture that outperforms standard Transformers. **DDAMS** (Feng et al., 2020) uses a relation graph to model the utterance interactions in a meeting by using different discourse relation models. **Summ^N** (Zhang et al., 2022) proposes a simple multi-stage summarization framework by generating a coarse summary in multiple stages and finally producing a fine-grained summary. **UniLM** (Dong et al., 2019) is a pre-trained model that employs a shared Transformer and utilizes self-attention masks to control the prediction context. **DialogLED** (Zhong et al., 2022) proposes pre-training strategies for dialogues based on sparse attention and LED models. **PGS-X** (Phang et al., 2022) utilizes block-based local attentions for computational efficiency based on PE-

[4] https://www.scrolls-benchmark.com/leaderboard

[5] https://huggingface.co/docs/transformers/index

[6] https://www.pytorchlightning.ai/index.html

Data & Mask & Loss	QMSum R-1	R-2	R-L	R-G
S1-word masking (L_{gen})	27.49	6.73	18.96	15.19
S1-sentence masking (L_{gen})	29.94	7.42	19.42	16.28
S2-long (L_{gen})	29.50	7.45	19.19	16.16
S2-long (L_{gen}+L_{turn})	30.24	7.47	19.39	16.36
S2-long (L_{gen}+L_{turn}+L_{MLM})	26.96	6.52	18.39	14.79

Table 6: ROUGE scores on QMSum with LED-large, comparing word and sentence masking strategies and combinations of learning objectives (L_{gen}, L_{turn}, L_{MLM}).

GASUS (Zhang et al., 2020) to incorporate long input. **LT5** (Guo et al., 2022) integrates the extended efficient attention architecture from ETC (Ainslie et al., 2020) and pre-training strategies from PEGASUS (Zhang et al., 2020).

6.2 Results on Summarization Tasks

Meeting Summairzation Table 4 presents a performance comparison of *SPECTRUM*, which is based on PGS-X, on three long meeting summarization tasks: AMI, ICSI, and QMSum. The performance of *SPECTRUM* is reported from the best results based on different pre-trained checkpoints. A more detailed analysis of different pre-trained checkpoints is provided in Section 6.3.1. In all datasets, we observe a significant improvement in Rouge scores with *SPECTRUM* compared to PGS-X. The R-G gains for AMI, ICSI, and QMSum are 2.58, 1.48, and 2.82, respectively. This highlights the effectiveness of our pre-training strategy and the carefully curated data that aligns with the downstream tasks, resulting in improved performance. Furthermore, our *SPECTRUM* model outperforms other pre-trained models such as HMNet and DialogLED by a large margin. Notably, *SPECTRUM* even surpasses larger-sized models with different efficient Transformers, LongT5-large and LongT5-xl, on the QMSum task. This highlights the superior performance and efficiency of our approach in long meeting summarization tasks.

Screenplay Summairzation In Table 5, we present the performance of the *SPECTRUM* models on the SummScreen dataset. Our observations indicate that *SPECTRUM* surpasses all other models in R-G on both subsets of the dataset. Additionally, *SPECTRUM* demonstrates competitive performance when compared to the 3B-size model, LongT5-xl. These results highlight again the effectiveness of *SPECTRUM* in achieving strong performance on the long dialogue summarization dataset.

CKPT	QMSum R-1	R-2	R-L	R-G
PGS-X	33.2	9.6	21.6	19.02
S1	35.54	11.74	23.05	21.26 (↑2.24)
S2-short	34.84	10.81	21.54	20.09 (↑1.07)
S2-short-ckpt	35.23	11.84	23.29	21.34 (↑2.32)
S2-long	35.08	11.73	22.72	21.07 (↑2.05)
S2-long-ckpt	36.06	11.52	22.95	21.20 (↑2.18)
S2-both-ckpt	35.86	12.34	23.53	21.84 (↑**2.82**)

Table 7: ROUGE scores on QMSum using different pre-trained checkpoints.

CKPT	ICSI R-1	R-2	R-L	R-G
PGS-X	48.65	13.40	18.26	22.83
S1	49.00	13.82	18.56	23.25 (↑0.42)
S2-short	49.82	14.16	19.38	23.91 (↑1.08)
S2-short-ckpt	49.00	12.63	17.75	22.23 (↓0.60)
S2-long	47.31	13.29	18.34	22.59 (↓0.24)
S2-long-ckpt	50.04	14.50	19.81	24.31 (↑**1.48**)
S2-both-ckpt	48.01	12.52	17.89	22.07 (↓0.76)

Table 8: ROUGE scores on ICSI using different pre-trained checkpoints.

6.3 Analysis

6.3.1 Effect of Pre-training Objectives

A series of ablation studies are conducted to analyze the masking strategies of the generation learning objective and identify the optimal combination of three pre-training objectives. The effectiveness of various pre-trained models using LED-large is demonstrated through fine-tuning on QMSum, as illustrated in Table 6. The first two rows indicate that sentence-level masking outperforms word-level masking with the L_{gen} loss pre-trained on S1. The Subsequent rows show the result of the complementarity of each learning objective. We employ the sentence-masking strategy for the L_{gen} loss and set β to 1 in the experimental setup. Notably, the joint computation of the turn prediction learning objective (L_{turn}) leads to enhanced performance. We additionally investigate the impact of the MLM loss (L_{MLM}) on downstream performance. The last row illustrates a notable degradation in performance. We hypothesize that masking tokens in the encoder input introduce confusion, impeding the model's accurate prediction of turns and exacerbating the learning process within the encoder path. Consequently, L_{MLM} is excluded from our final model.

6.3.2 Effect of Pre-training Corpus

Table 7, 8 shows the fine-tuned results on QMSum and ICSI datasets using different pre-trained check-

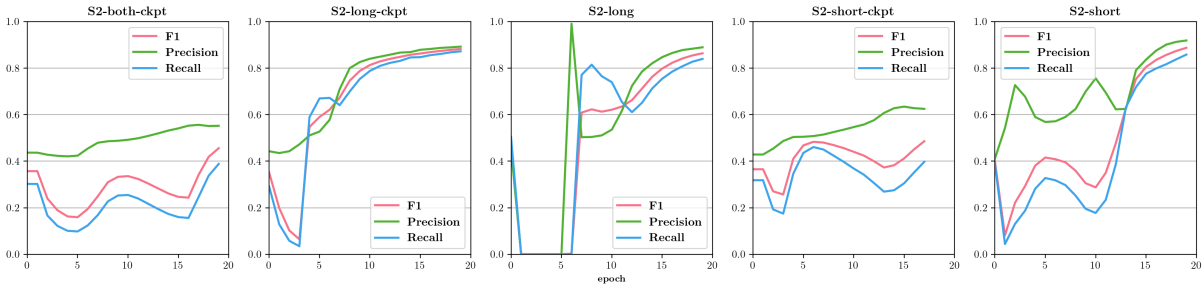

Figure 2: F1, precision, recall scores for the turn switch prediction. Each figure shows the performances with the corresponding S2 pre-training data with ('-ckpt') or without parameter initialization pre-trained on S1.

points. PGS-X refers to the original PEGASUS-X checkpoint, S1 denotes pre-training on S1 with L_{gen}, and S2 corresponds to pre-training on different S2 data with or without initialization from the S1 checkpoint. Significant performance improvement is evident across all pre-trained checkpoints on QMSum. However, performance gains are comparatively limited on ICSI. We postulate that this phenomenon is attributed to the small size of the ICSI dataset, comprising only 43 training instances, which poses a challenge for the model to align effectively. In contrast, the QMSum dataset contains 1,257 training samples, allowing the model to better adapt during fine-tuning. This observation aligns with the insights proposed by (Zhang et al., 2020), emphasizing the significance of aligning the domain of pre-trained models to facilitate more effective transfer to downstream tasks.

6.3.3 Effect of Speaker Turn Prediction

Figure 2 presents the F1, precision, and recall scores for turn switch prediction concerning various S2 data with the initialization of the S1 checkpoint. Several noteworthy observations can be derived from the results. Firstly, it is observed that stage2 pre-training attains high turn prediction performance even without initializing the model with the S1 checkpoint, as evidenced by the S2-short and S2-long cases. Furthermore, the S2-long-ckpt configuration exhibits notably high F1 scores. We posit that the consistent length of the S1 data aligns effectively with the characteristics of the S2-long data, thereby contributing to the observed high prediction performance.

On the other hand, S2-short-ckpt and S2-both-ckpt, which involve different data distributions in terms of length, show degraded prediction performance during pre-training. This suggests that the deviation of data length in two-stage training adversely affects the model's ability to accurately predict turns. Importantly, the model's ability to predict turn switches aligns closely with the downstream results on the ICSI dataset (Table 8), which has a limited number of training samples. The parameter initialization with a checkpoint trained on a different data length distribution can impair the fine-tuning performance on scarce data, contrasting with the opposite scenario observed with a sufficiently large dataset (Table 7). Taken together, these findings highlight the importance of pre-training methodologies, dataset characteristics, and the alignment between the pre-training and fine-tuning stages.

7 Conclusion

In this paper, we have introduced *SPECTRUM*, a speaker-enhanced pre-training method that specifically addresses the challenges associated with dialogue summarization. Our two-stage pre-training approach leverages the natural structure of dialogues with multiple turns, allowing the model to better comprehend the intricacies of dialogue interactions. To enable robust training of models, we utilized a comprehensive dataset comprising real-world transcripts, including dialogues from movies, TV shows, and dialogues generated by large language models. This diverse dataset with turn-taking information provides solid signals for training dialogue understanding models. Our experimental results demonstrate the effectiveness of our speaker-enhanced pre-training method and show significant performance gains over baseline models. By capturing and leveraging speaker and turn-based information, our approach enhances the ability of models to generate accurate and coherent summaries of dialogues. Moving forward, future research endeavors can further explore and refine our pre-training methods.

8 Limitation

While our work introduces novel learning objectives and exhibits promising results, it is not without limitations.

Although our model is designed for long dialogue summarization, it still encounters difficulties when dealing with exceptionally long dialogues. With the increase in length, the complexity of the dialogue also grows, and it becomes harder to maintain the semantic coherence and continuity of the summarization output. This is a common problem in summarization tasks, but its impact is amplified in long dialogue summarization due to the increased amount of data the model must process and understand.

The current evaluation metrics we employ have their inherent limitations. While ROUGE scores have been widely used in summarization tasks, they may not fully capture the quality of generated summaries, such as the preservation of speaker-specific nuances or shifts in conversation topics. Also, given the lengthy dialogues, conducting human evaluation is both challenging and potentially inaccurate. The constraint on context length in Large Language Models often poses challenges for evaluating long dialogue summarization, and LLMs may potentially exhibit hallucination problems. Therefore, further research on better representative evaluation metrics is needed.

Another aspect of the proposed model is open to many potential risks such as loss of context in conversation, contextual inconsistency, biased summary in dialogue, and ethical concerns in social interactions. Addressing these risks involves continuous improvement of the summarization models, careful consideration of ethical implications, and user feedback to enhance the system's performance in handling diverse and dynamic dialogues. Regular updates and refinements to the underlying models can contribute to mitigating these challenges over time.

Lastly, the training and deployment of our model demand significant computational resources. This may limit the use of our method in real-world applications, particularly in scenarios where computational resources are constrained.

Future work will aim to address these limitations. We plan to refine our approach to better handle lengthy and complex dialogues, improve its ability to generalize across various domains, and explore new ways to quantify the performance of our model more effectively.

References

Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. ETC: Encoding long and structured inputs in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 268–284, Online. Association for Computational Linguistics.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

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

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

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Ta-Chung Chi, Po-Chun Chen, Shang-Yu Su, and Yun-Nung Chen. 2017. Speaker role contextual modeling for language understanding and dialogue policy learning. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 163–168, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas. Association for Computational Linguistics.

Richard Csaky and Gábor Recski. 2021. The Gutenberg dialogue dataset. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 138–159, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation.

Xiachong Feng, Xiaocheng Feng, Bing Qin, Xinwei Geng, and Ting Liu. 2020. Dialogue discourse-aware graph convolutional networks for abstractive meeting summarization. *ArXiv*, abs/2012.03502.

Zhichao Geng, Ming Zhong, Zhangyue Yin, Xipeng Qiu, and Xuanjing Huang. 2022. Improving abstractive dialogue summarization with speaker-aware supervised contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6540–6546, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Adam L. Janin, Don Baron, Jane Edwards, Daniel P. W. Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The icsi meeting corpus. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, 1:I–I.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, et al. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Yuejie Lei, Yuanmeng Yan, Zhiyuan Zeng, Keqing He, Ximing Zhang, and Weiran Xu. 2021. Hierarchical speaker-aware sequence-to-sequence model for dialogue summarization. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7823–7827.

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.

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.

Pierre Lison and Jörg Tiedemann. 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)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Longxiang Liu, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. 2020a. Filling the gap of utterance-aware and speaker-aware representation for multi-turn dialogue. In *AAAI Conference on Artificial Intelligence*.

Yafei Liu, Hongjin Qian, Hengpeng Xu, and Jinmao Wei. 2020b. Speaker or listener? the role of a dialog agent. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4861–4869, Online. Association for Computational Linguistics.

Yang Liu, Chenguang Zhu, and Michael Zeng. 2022. End-to-end segmentation-based news summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 544–554, Dublin, Ireland. Association for Computational Linguistics.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Melissa Kronenthal, Guillaume Lathoud,

Mike Lincoln, Agnes Lisowska, Wilfried Post, Dennis Reidsma, and Pierre D. Wellner. 2005. The ami meeting corpus.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Jason Phang, Yao Zhao, and Peter J. Liu. 2022. Investigating efficiently extending transformers for long input summarization. *ArXiv*, abs/2208.04347.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

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

Xiaoyang Wang Fei Liu Sangwoo Cho, Kaiqiang Song and Dong Yu. 2022. Toward unifying text segmentation and long document summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

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.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020. Efficient transformers: A survey. *CoRR*, abs/2009.06732.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big bird: Transformers for longer sequences.

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

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

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization.

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

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 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*, pages 5927–5934, 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.