Curriculum Prompt Learning with Self-Training for Abstractive Dialogue Summarization

Changqun Li[1], **Linlin Wang**[1]*, **Xin Lin**[1], **Gerard de Melo**[2], **Liang He**[1]
[1] East China Normal University
[2] Hasso Plattner Institute / University of Potsdam
52215901009@stu.ecnu.edu.cn, {llwang,xlin,lhe}@cs.ecnu.edu.cn, gdm@demelo.org

Abstract

Succinctly summarizing dialogue is a task of growing interest, but inherent challenges, such as insufficient training data and low information density impede our ability to train abstractive models. In this work, we propose a novel curriculum-based prompt learning method with self-training to address these problems. Specifically, prompts are learned using a curriculum learning strategy that gradually increases the degree of prompt perturbation, thereby improving the dialogue understanding and modeling capabilities of our model. Unlabeled dialogue is incorporated by means of self-training so as to reduce the dependency on labeled data. We further investigate topic-aware prompts to better plan for the generation of summaries. Experiments confirm that our model substantially outperforms strong baselines and achieves new state-of-the-art results on the AMI and ICSI datasets. Human evaluations also show the superiority of our model with regard to the summary generation quality.

1 Introduction

As billions of people engage in instant messaging and other forms of interaction, there is notable interest in techniques to process and distill recorded dialogues into concise and natural summaries (Gurevych and Strube, 2004; Murray et al., 2006). The inherent challenges of the dialogue make abstractive dialogue summarization particularly difficult. First, the available labeled data is substantially smaller than news summarization data, e.g., 137 meetings in AMI (McCowan et al., 2005) vs. 312K articles in CNN/Daily Mail (Hermann et al., 2015). Second, everyday dialogue involves a dynamic information exchange flow (Sacks et al., 1978) such that salient information is often scattered across multiple utterances by different interlocutors (Li and Choi, 2020).

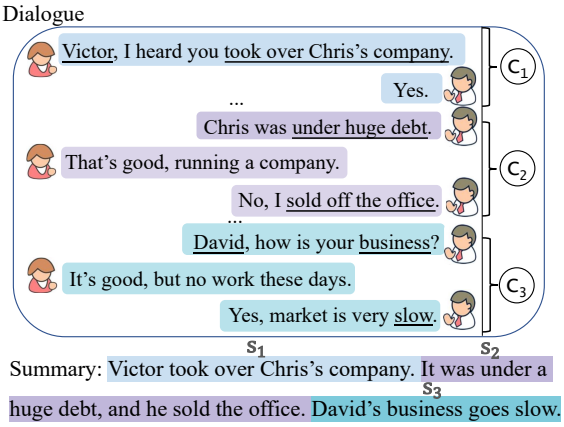

Figure 1: A dialogue and its summary, consisting of three sentences s_1, s_2, s_3. For illustration purposes, we show relevant content snippets C_1, C_2, C_3 in the dialogue that these correspond to, and have also underlined salient information that appears in the summary.

Recently, there is a growing trend of using Pre-trained Language Models (PLMs) for dialogue summarization (Gliwa et al., 2019; Fabbri et al., 2021). However, these models tend to require abundant labeled dialogue, which is difficult to procure for low-resource domains and downstream tasks. To reduce the dependency on labeled dialogue resources, Chen and Yang (2021) explore a self-training method (He et al., 2020; Xie et al., 2020) to include unlabeled data, but the model fails to consider that abundant in-domain unlabeled dialogue remains hard to procure in low-resource settings. In recent years, prompt learning (Li and Liang, 2021; Lester et al., 2021) has become a promising alternative to full model fine-tuning that significantly reduces the amount of parameters to be tuned in few-shot settings. In our pilot experiments, however, prompt learning did not work as expected for dialogue summarization. This is because general prompt learning assigns universal prompt tokens to all inputs in a given task. In contrast, dialogue is a dynamic form of interaction

* Corresponding author. Email: llwang@cs.ecnu.edu.cn.
The first two authors contributed equally to this work.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1096 - 1106
December 7-11, 2022

among multiple participants. We conjecture that this tasks needs to be addressed using prompt learning with a high level of semantic understanding.

Considering a typical example in Figure 1, the dialogue includes discussion of three different matters, which we marked as content snippets C_1, C_2, and C_3, respectively. The summary consists of salient information summarizing each snippet C_i with one sentence s_i. Moreover, the content of s_1 stems from a single utterance, while the content of s_2 and s_3 is scattered across multiple utterances. Additionally, s_3 is more abstract, so it is necessary to understand interaction relationships between utterances to succinctly summarize the C_3 snippet. This phenomenon suggests that a single prompt may be insufficient to model the dialogue adequately.

Considering the inherent challenges of dialogue and the advantage of prompt learning in few-shot settings, in this work, we propose a novel curriculum-based prompt learning method with self-training. The proposed prompts are learned using a curriculum learning strategy that gradually increases the degree of prompt perturbation to obtain a high level of semantic understanding, specifically including soft prompts, perturbed prompts, and interpolated prompts. We further incorporate unlabeled dialogue via self-training (Zoph et al., 2020; He et al., 2020) to optimize the prompts and reduce the reliance on labeled dialogue. Additionally, we investigate topic-aware prompting to better plan for the generation of dialogue summaries. Extensive experiments on diverse benchmark datasets evince the effectiveness of our model for dialogue summarization. To sum up, our contributions are:

- Our curriculum-based prompt learning strategy gradually increases the degree of prompt perturbation to improve the understanding ability of the proposed model on dialogue.
- We further utilize unlabeled dialogue by self-training to alleviate the problem of insufficient training data and investigate topic-aware prompts to better plan for the generation.
- We extensively evaluate our model on three dialogue summarization datasets and obtain new state-of-the-art results on AMI and ICSI.

2 Related Work

2.1 Dialogue Summarization

Dialogue summarization has received increasing attention recently. Current studies typically apply conventional Transformer-based summarization architectures, e.g., BART (Lewis et al., 2020), directly to dialogue scenarios (Gliwa et al., 2019; Fabbri et al., 2021), whereas these models are pre-trained with well-written texts such as news articles that are notably different from dialogue in a number of respects. To achieve better results, subsequent research incorporates diverse distinctive traits of dialogue to boost the performance, including dialogue acts (Goo and Chen, 2018), topic-related multi-modal information (Li et al., 2019), domain terminologies (Koay et al., 2020), commonsense knowledge (Feng et al., 2021a), and dialogue discourse (Feng et al., 2021b). Another line of work solves the challenge of very long sequences in input dialogues with a hierarchical architecture (Zhu et al., 2020) and the Longformer model (Fabbri et al., 2021). However, most existing summarization approaches perform poorly when the annotated dialogues are limited (Chen and Yang, 2021). In this work, we explore a novel strategy that enables the model to utilize abundant relevant signals from unlabeled data, thereby reducing the dependency on labeled dialogue in low-resource settings.

2.2 Prompt Learning and Self-Training

A recent trend in Natural Language Processing (NLP) has been to explore prompt learning as a lightweight alternative to fine-tuning. Prompt learning keeps the parameters of PLMs frozen and optimizes only a small portion pertaining to prompts. This allows few-shot or nearly zero-shot learning for pre-trained models on new tasks with scarce or entirely unlabeled data and it has been demonstrated to be very effective over fine-tuning in a number of tasks. For example, Li and Liang (2021) propose "Prefix-Tuning", which only tunes "soft tokens" (prefix) activation prepended to all Transformer layers, and keeps the PLM parameters frozen. Prompt tuning (Lester et al., 2021) prepends a sequence of prompt tokens to the source text, and only the embeddings of these tokens are optimized. Gu et al. (2022) propose a pretrained prompt tuning framework to boost the performance of existing models in few-shot learning. However, dialogue summarization requires models to understand interactions between multiple utterances to generate succinct summaries, suggesting that existing prompt learning techniques may be insufficient to extract adequate information imparted across multiple turns of the dialogue.

Self-training is a simple and effective pseudo-label semi-supervised learning method (Lee et al., 2013; Chen et al., 2021; Chen and Yang, 2021) that often iteratively performs the process of creating pseudo-labels on unlabeled data with a teacher model, and subsequently applying the combined labeled data to train a student model. Bringing in large amounts of unlabeled data can lead to better-performing models, particularly when labeled data is scarce. Inspired by these developments in prompt-learning and self-training, in this work, we propose a new curriculum-based prompt learning with self-training optimization for better abstractive dialogue summarization.

3 Task Formulation

Given a dialogue X consisting of D utterances from multiple speakers, abstractive dialogue summarization aims to compress the input dialogue X into a concise summary Y, typically maximizing the conditional probability $P(Y|X;\theta,\phi)$ over N instances, where θ here represents prompt-related parameters and ϕ denotes the remaining parameters of the backbone model, such as BART.

4 Approach

To address the task of abstractive dialogue summarization, we propose an encoder–decoder architecture with heterogeneous prompts, which gradually increases the degree of prompt perturbation via a curriculum learning strategy, and conducts the optimization using a self-training technique.

4.1 Model Overview

An overview of our model is given in Figure 2, where the backbone encoder–decoder architecture is an extension of the prominent BART model (Lewis et al., 2020). As depicted, heterogeneous prompts, including curriculum prompts (left) and topic-aware prompts (middle) are incorporated to boost the performance of our model on dialogue summarization. Self-training optimization (right) is further proposed to exploit abundant relevant information from unlabeled dialogue, aiming to alleviate the problem of insufficient training data. In the following, we explain these steps in detail.

4.2 Heterogeneous Prompt Construction

We design two types of prompts, including: (1) curriculum learning based prompts, on which we increase the degree of perturbation so as to improve the generalization ability and obtain better results, and (2) topic-aware prompts, which enable planning the generation of dialogue summarization.

4.2.1 Curriculum Learning based Prompts

Motivated by the cognitive progress of humans when gradually acquiring knowledge from easy to hard, we propose a curriculum learning approach (Bengio et al., 2009) to augment prompt learning by increasing the degree of prompt perturbation gradually. Specifically, we introduce three types of prompts, namely soft prompts, perturbed prompts, and interpolated prompts. Soft prompts serve to learn essential features for the dialogue understanding and modeling, and perturbed prompts aim to improve the generalization of soft prompts via additive perturbations. Interpolated prompts are relevant mixtures of soft prompts with perturbed prompts, which boost the understanding of the inherently rich interactions between utterances.

Soft Prompts Inspired by Lester et al. (2021), we prepend a soft prompt P to the input of our encoder. Here, $P = \{p_1, p_2, ..., p_\rho\}$ is a sequence of trainable token vectors parametrized by θ_P, where $\theta_P \in \mathbb{R}^{\rho \times d}$, ρ is the length of the soft prompt, and d is the hidden state dimensionality. During training, we update the parameters of the soft prompt while freezing all PLMs parameters.

Perturbed Prompts In empirical investigations of dialogue summarization, we observe a common phenomenon that particularly long prompts with more than 200 trainable parameters tend to overfit the training data and become less generalizable. To improve the generalization ability of prompts, we introduce two simple operations for prompt perturbations: (1) random swapping, which breaks the original relations by randomly swapping two tokens in a prompt, such that the perturbed prompt may become $P' = \{p_2, p_i, ..., p_1, ..., p_\rho\}$ (where ρ denotes the length of the perturbed prompt), and (2) cutoff, which consists of both span and token-level cutoff operations, aiming to induce a more severely perturbed prompt. To explain this process, we define two preset coefficients α_1 and α_2 that indicate the ratio between the length/number of removed spans/tokens to that of the prompt. For span-level cutoff, we set the length of a span as $l = \alpha_1 \times \rho$, and randomly sample the starting index s for this span from the index range from 0 to $\rho - l$. Subsequently, the vectors with respect to the prompt tokens from the s-th to $(s + l - 1)$-

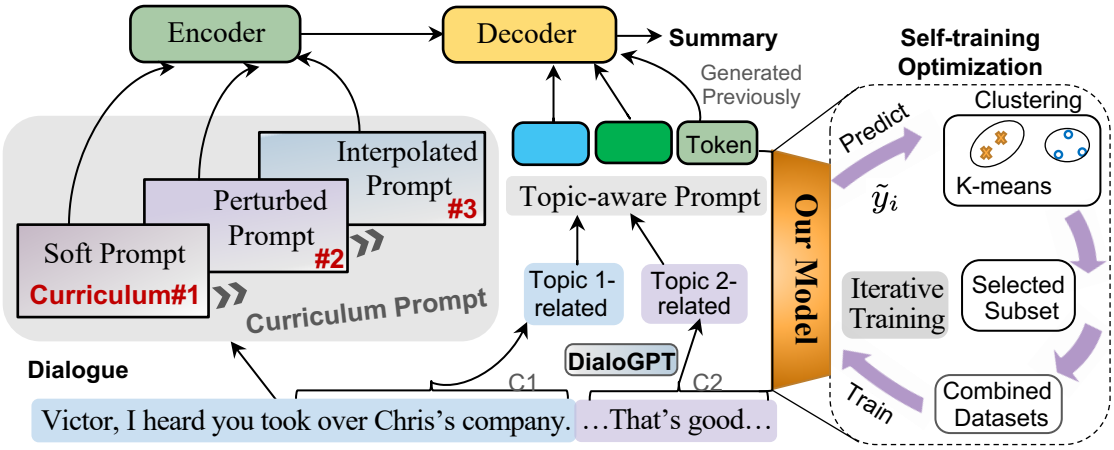

Figure 2: Overview of our proposed model that consists of a Transformer-based encoder–decoder architecture with heterogeneous prompts and self-training optimization. Heterogeneous prompts include **curriculum prompts** with increasing degree of perturbation to improve the generalization ability, and **topic-aware prompts** that aid in planning for the summary generation. Self-training optimization exploits information in unlabeled dialogue.

th positions are all masked, i.e., turned into zero vectors. As for token-level cutoff, the number of masked tokens is set to be $\alpha_2 \times \rho$, and the indexes of masked tokens are randomly sampled as well.

Interpolated Prompts Inspired by the recent success of Mixup (Zhang et al., 2018) and MixText (Chen et al., 2020), we explore interpolation techniques to mix the above two types of prompts, seeking to understand the interaction relationships in multiparty dialogue with interpolated prompts. Given a soft prompt P and its corresponding perturbed prompt P', the Mixup algorithm creates a novel virtual prompt by linear interpolation:

$$\tilde{P} = \lambda P + (1 - \lambda) P', \quad (1)$$

where λ is a scalar mixing ratio that is sampled from a Beta distribution $\mathrm{Beta}(\alpha, \alpha)$ for every batch to perform the interpolation, and α is the hyperparameter to control the distribution of λ. Different from Mixup, we do not need to mix the corresponding labels in the same way. We believe that by mixing the two types of prompts, more attention can be given to understanding and modeling inherently rich interactive relations between utterances.

4.2.2 Topic-aware Prompts for Planning

Decoding with Learned Prompts Unlike regular prose documents, dialogue consists of multiple utterances from two or more participants forming a dynamic information exchange flow (Sacks et al., 1978). The topics being discussed can vary during the progression of a conversation. This trait makes it difficult to summarize dialogue.

Given the structure of dialogue, we investigate how to better control the summary generation process with topic planning. Specifically, we introduce topic-aware prompts P_t, which are constructed by prepending topic-related features from different dialogue segments. As shown in Figure 2, we first leverage DialoGPT (Feng et al., 2021b), an unsupervised dialogue annotator, to capture topic-related information by dividing the dialogue into C topically coherent segments. Therefore, we have $P_t = \{P_t^1, P_t^2, \ldots, P_t^c, \ldots, P_t^{\mathrm{C}}\}$ with parameters θ_{P_t} to be updated, where $c \in \{1, \ldots, \mathrm{C}\}$, each topic-aware prompt $P_t^c = \{p_1^c, p_2^c, ..., p_{\rho_c}^c\}$ corresponds to a topic segment, and ρ_c refers to the length. We thus combine each P_t^c with the corresponding topic segment by calculating the average value of every dimension in P_t^c, and adding this average value to the token embeddings in the corresponding segment. We believe that by prompting based on different topic segments, these combined prompts are able to capture crucial features for different topics in the dialogue. Further, we concatenate these prompts P_t to the decoder inputs. Thus, topic-aware prompts are able to aid in planning for the generation process of dialogue summarization, thereby improving the quality of summaries.

4.3 Prompt Optimization with Self-Training

To further improve the ability to learn from limited labeled dialogues, we combine the proposed prompt learning with self-training (Zoph et al., 2020; He et al., 2020) to harness unlabeled dialogue data. Our prompt optimization with self-training

Algorithm 1: Self-Training Optimization

Input: Labeled dialogue $D^L = \{(x_q, y_q)\}_{q=1:|L|}$, unlabeled dialogue $D^U = \{(\tilde{x}_h)\}_{h=1:|U|}$, and the maximum number of iterations E.

Output: Dialogue summarization model $f(.)$.

while not reaching the maximum iteration steps **do**

Learn a teacher model $f(x_q; \theta^t)$ on D^L, which minimizes the cross-entropy loss:

$$\frac{1}{|L|}\sum_{q=1}^{|L|}\ell(y_q, f(x_q;\theta^t)) \quad (2)$$

Predict pseudo-summaries for D^U with $f(.;\theta^t)$:

$$\tilde{y}_h = f(\tilde{x}_h;\theta^t), \forall h = 1,\ldots,|U| \quad (3)$$

Learn a student model $f(x;\theta^s)$, which minimizes the cross-entropy loss on D^L and $D^{U'}$(a subset selected from D^U via dynamic thresholding):

$$\frac{1}{|L|}\sum_{q=1}^{|L|}\ell(y_q, f(x_q;\theta^s)) + \frac{1}{|U|}\sum_{h=1}^{|U|}\ell(\tilde{y}_h, f(\tilde{x}_h;\theta^s)) \quad (4)$$

end

approach is specified in Algorithm 1.

Self-training refers to the process of creating pseudo-labels on unlabeled data with a teacher model, and then applying the combined data to train a student model. Nevertheless, conventional self-training approaches have some shortcomings. On one hand, self-training needs a sizeable amount of in-domain unlabeled data, whereas, for low-resource tasks, abundant in-domain unlabeled data is often not available. On the other hand, a drawback of previous work (Chen and Yang, 2021) is that it relies on a pre-defined constant threshold and ignores a considerable amount of other unlabeled dialogue, especially for samples that have a greater learning difficulty, which may fail to get selected throughout the entire training process.

To address the first problem, we leverage data augmentation (Wu et al., 2021b) to synthesize unlabeled data and reduce the impact of the data domain. For SAMSum, we first randomly select two dialogues from the dataset without considering the labels, and subsequently concatenate them by adding a special token <SEP> in between, obtaining ample synthetic data as unlabeled resources. For the AMI and ICSI datasets, we leverage DialoGPT[1] to divide the dialogue into topically coherent segments for all training instances, and subsequently randomly sample three topic segments from all segments as synthetic data. To improve the quality of all synthetic data, we further utilize advanced techniques, e.g., masking words, to process all synthetic data, obtaining the final forms of abundant unlabeled data. To resolve the second problem, we further replace the constant threshold with a dynamic thresholding mechanism explained in further detail in the following.

Dynamic Thresholding To perform dynamic thresholding, we first cluster the entire unlabeled data with the K-means algorithm, such that different clusters represent dialogue samples with different learning difficulties. We then use the teacher model to generate pseudo-summaries for every instance in a cluster c_k, and calculate the corresponding BERTScore, which is later used for comparison with the threshold to determine which dialogue samples to add. The proposed dynamic thresholding mechanism can thus dynamically adjust our threshold for every cluster based on the number of instances in the corresponding cluster.

4.4 Training Objective

We formulate the training objective as follows with respect to the model parameters $\widehat{\theta}$:

$$\widehat{\theta^*} = \underset{\widehat{\theta}}{\text{argmax}} \sum_{j=1}^{N} \log p\left(Y^j \mid X^j; \theta_{P_c}, \theta_{P_t}, \phi\right), \quad (5)$$

where $\theta_{P_c} = \{\theta_P; \theta_P^1; \theta_P^2\}$ refers to the parameters of curriculum-based prompts, and θ_P, θ_P^1, θ_P^2 represent the soft, perturbed, interpolated prompt parameters, respectively. In addition, θ_{P_t} stands for topic-aware prompts parameters, and ϕ denotes the remaining parameters of the backbone model. Note that $\widehat{\theta}$ consists of various types of parameters, while prompt learning keeps the parameters of PLMs frozen and optimizes only θ_{P_c} and θ_{P_t}.

5 Evaluation

5.1 Experimental Setup

Datasets We conduct our experiments on three commonly-used benchmark datasets, AMI (McCowan et al., 2005), ICSI (Janin et al., 2003), and SAMSum (Gliwa et al., 2019), comprising English language dialogues from both meeting and daily life chat domains. Detailed statistics are given in Table 1. Additionally, we further study the effectiveness of our model in few-shot scenarios, which

[1] https://github.com/xcfcode/PLM_annotator

Description	AMI	ICSI	SAMSum
Source Domain	Meeting	Meeting	Daily Chat
Number of dialogues	137	59	16,369
Train/Dev./Test	97/20/20	43/10/6	14,732/818/819
Avg. participants	4.0	6.2	2.38
Avg. turns	296.74	398.63	11.08
Avg. dialogue length	5,490.98	11,000.60	83.68
Avg. summary length	286.81	521.53	20.31
Unlabeled data	291	129	29,464

Table 1: Statistics of AMI, ICSI (Meeting), and SAMSum (daily life chat) datasets.

Model	AMI			ICSI		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
TextRank (Mihalcea and Tarau, 2004)	35.19	6.13	15.70	30.72	4.69	12.97
PGN (See et al., 2017)	42.60	14.01	22.62	35.89	6.92	15.67
Sentence-Gated (Goo and Chen, 2018))	49.29	19.31	24.82	39.37	9.57	17.17
$\text{PEGASUS}_{\text{large}}$ (Zhang et al., 2020a)	47.05	16.64	16.03	42.44	9.15	11.10
$\text{BART}_{\text{large}}$ (Lewis et al., 2020)	50.83	17.80	26.77	41.03	9.45	19.85
HMNet (Zhu et al., 2020)	52.36	18.63	24.00	45.97	10.14	18.54
TopicSeg (Li et al., 2019)	51.53	12.23	25.47	–	–	–
TopicSeg+VFOA (Li et al., 2019)	53.29	13.51	26.90	–	–	–
DDAMS+DDADA (Feng et al., 2021b)	53.15	22.32	25.67	40.41	11.02	19.18
$\text{LED}_{\text{large}}$ (Beltagy et al., 2020)	54.20	20.72	25.98	43.13	11.76	19.08
Ours w/o topic planning	54.69	22.38	27.93	45.83	11.91	20.39
Our Architecture	**55.76**	**22.85**	**28.11**	**46.34**	**12.30**	**20.92**

Table 2: Comparison of results on AMI and ICSI datasets.

are constructed by randomly sampling 10, 100, or 1,000 dialogue instances from the original training set of SAMSum.

Automatic Metrics We use standard evaluation metrics that include ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004)[2] to assess the quality of generated summaries, which consider the overlapping uni-grams, bi-grams, and the longest common subsequences, respectively.

Human Metrics Apart from automatic metrics, we conduct human evaluations to assess the quality of generated summaries. Specifically, we randomly sample 150 dialogues from SAMSum, all dialogues from AMI and ICSI, and ask three experts to grade the quality of generated summaries using three criteria: (1) **Fluency** measures how well the generated summaries are readable. (2) **Informativeness** evaluates how well the generated summaries capture more salient information. (3) **Relevance** evaluates how well the generated summaries reflect the input dialogues. We set the range of rating from 1.0 to 5.0 (higher scores indicating a better quality).

Baselines and Experimental Settings A variety of representative models are chosen as competitive baselines for our experiments, ranging from early ranking-based models such as TextRank (Mihalcea and Tarau, 2004) to mainstream Transformer-based approaches (e.g., $\text{BART}_{\text{large}}$). Further comparison details are provided in Section 5.2.

All implementations are built on the top of the Transformers[3] (Wolf et al., 2020) library. During training, we leverage a linear learning rate scheduler and AdamW optimization (Loshchilov and Hutter, 2019). The two coefficients α_1 and α_2 in perturbed prompts are selected from {0.03, 0.07, 0.1} and {0.01, 0.03, 0.05}, respectively. We set different learning rates for three types of prompts in curriculum-based prompt learning by decreasing the rates as the learning process progresses. Specifically, when using the soft, perturbed, and interpolated prompts, we set the learning rates as 7×10^{-5}, 5×10^{-5}, and 4×10^{-5}, respectively. In addition, we set the maximum iteration times as 5, and the batch size to be 16 for SAMSum and 1 for AMI and ICSI. In the decoding phase, we set

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

[3] https://github.com/huggingface/transformers

the beam size to be 4, and assign length normalization to be 0.8 for SAMSum and 0.5 for AMI/ICSI, respectively.

5.2 Main Results

Table 2 provides a comparison of our model with previous approaches on AMI and ICSI. We observe that our model achieves new state-of-the-art results on these two benchmark datasets. For instance, compared with the previous baseline model $\text{LED}_{\text{large}}$ (Beltagy et al., 2020), our model obtains relative gains of 7.4% on ROUGE-1, 4.6% on ROUGE-2, and 9.6% on ROUGE-L on the ICSI dataset. Furthermore, our model significantly outperforms the conventional $\text{BART}_{\text{large}}$ with fine-tuning (Lewis et al., 2020) on both AMI and ICSI, demonstrating the effectiveness of our prompt learning in dialogue summarization.

Model	R-1	R-2	R-L
Baselines			
TextRank (Mihalcea and Tarau, 2004)	29.27	8.02	28.78
PGN (See et al., 2017)	40.08	15.28	36.63
DialoGPT (Zhang et al., 2020b)	39.77	16.58	38.42
$\text{BART}_{\text{base}}$ (Lewis et al., 2020)	49.1	24.29	45.76
CODA (Chen and Yang, 2021)	50.08	24.62	46.89
FinDS (Lei et al., 2021)	52.23	25.91	50.87
CODS (Wu et al., 2021a)	52.65	27.84	50.79
$\text{BART}_{\text{large}}$ (Lewis et al., 2020)	53.36	28.37	50.19
$\text{BART}_{\text{D}_{\text{all}}}$ (Feng et al., 2021c)	53.70	28.79	50.81
CONDIGSUM (Liu et al., 2021)	54.30	29.30	45.20
DiaSumm+Coref (Liu and Chen, 2021)	55.30	31.30	**53.20**
Our Architecture			
-w/ $\text{BART}_{\text{large}}$	**55.97**	**31.67**	52.32

Table 3: Comparison of results on full SAMSum dataset.

Results of the comparison on SAMSum are reported in Table 3. Our model also achieves strong results on this dataset, suggesting the generalization ability of our model across diverse domains. Instead of relying heavily on additional coreference resolution or named entity tagging tools, our model obtains relative improvements of 1.2% on ROUGE-1 and 1.2% on ROUGE-2 compared with the previous state-of-the-art model DiaSumm+Coref (Liu and Chen, 2021).

Few-shot settings To further study the effectiveness of our model in few-shot scenarios, we set different few-shot settings based on SAMSum. Table 4 shows that our model achieves the best results on different sample settings, especially when there are fewer than 100 samples.

Human Evaluation Table 5 shows the mean human ratings of different models on AMI, ICSI and SAMSum. The summaries generated by our model prove preferable across all considered metrics, further confirming the effectiveness of our approach.

6 Quantitative Analysis

6.1 Ablation Study

To better quantify the contributions of different components in our model, we conduct ablation studies with three types of simplified architectures as follows. The first type removes the all curriculum prompts, and the second drops topic-aware prompts. The third variant is to use our model without self-training. Table 6 provides the results of the corresponding ablations on the dev. set of ICSI. We observe that all of the aforementioned components of our model make noticeable contributions. For example, the removal of curriculum prompts causes a relative performance drop of 3.0% on ROUGE-1, confirming the validity of our curriculum-based prompt learning strategy. When self-training is removed, this causes a relative performance drop of 2.6% on ROUGE-1, which shows the effectiveness of self-training with our augmented pseudo-dialogue data.

6.2 Impact of Curriculum Prompts

Curriculum Prompt Ablation In Table 7, we deeply investigate the effect of our curriculum-based prompt learning, where we adopt the same hyperparameters for all experiments. We observe that the model performance degrades with the removal of each curriculum stage, indicating that curriculum-based prompt learning with three stages in our model accounts for more importance.

Comparison with Prompt Variants We further compare the proposed prompt learning strategy with other prompt variants in Table 8. Here, we use the same backbone models for all variants. From this table, we observe that substantial gains are made when going from prompt tuning to our model. For example, our approach outperforms full-model fine-tuning with a relative increase of 7.4% in terms of ROUGE-1 on the ICSI dataset.

6.3 Parameter Sensitivity of Self-training

Iteration We further study the effects of iterative training in our model, adopting the same hyperparameters for all the iterations. As shown in Figure 3, ROUGE scores keep improving at first, achieve the

Training Instances	10	100	1000
Model	**R-1 / R-2 / R-L**	**R-1 / R-2 / R-L**	**R-1 / R-2 / R-L**
$\text{BART}_{\text{large}}$ (Lewis et al., 2020)	30.24 / 10.83 / 31.97	41.53 / 18.84 / 40.85	49.57 / 23.66 / 46.84
Prompt-tuning (Lester et al., 2021)	33.58 / 12.33 / 33.37	40.77 / 18.39 / 40.52	48.98 / 23.43 / 46.38
Ours w/o topic planning	35.75 / 13.31 / 34.60$_{\uparrow 3.69\%}$	43.87 / 19.62 / 41.56	50.22 / 23.96 / 47.12
Ours	**36.21 / 13.98 / 35.38**$_{\uparrow 6.02\%}$	**44.56 / 20.03 / 42.35**	**51.03 / 24.23 / 47.38**

Table 4: Results on SAMSum dataset in different few-shot settings. Our model significantly outperforms $\text{BART}_{\text{large}}$.

Models	**AMI**			**ICSI**			**SAMSum**		
	Flu.	Info.	Rel.	Flu.	Info.	Rel.	Flu.	Info.	Rel.
$\text{BART}_{\text{large}}$	4.21	4.00	4.34	4.12	3.84	4.23	4.46	4.19	4.62
Prompt-tuning	4.02	3.86	4.17	3.84	3.56	3.99	4.33	3.89	4.41
Ours	4.37	4.02	4.64	4.25	3.94	4.47	4.53	4.21	4.76

Table 5: Human evaluation on AMI , ICSI, and SAMSum. "Flu.", "Info.", and "Rel." stand for fluency, informativeness, and relevance, respectively.

Model	**R-1**	**R-2**	**R-L**
Ours			
-w/ $\text{BART}_{\text{large}}$	47.23	12.67	21.14
The Ablations			
- w/o curriculum prompts	45.83	11.91	20.39
- w/o topic-aware prompts	46.41	12.35	20.86
- w/o self-training	46.00	12.12	20.52

Table 6: Ablation study on dev. set of ICSI.

Model	**R-1**	**R-2**	**R-L**
Curriculum Prompts Only			
Ours			
-w/ $\text{BART}_{\text{large}}$	44.18	10.81	19.91
-w/o interpolated prompts	44.00	10.55	19.56
-w soft prompts	42.62	9.57	18.75

Table 7: Ablation study regarding curriculum stategies on dev. set of ICSI.

best performance at iteration 4, and then start to converge. This indicates the effectiveness of iterative training by continually updating the teacher model to generate better pseudo-summaries.

Dynamic Thresholding Figure 4 compares the performance of our dynamic thresholding with the conventional constant threshold for self-training.

Model	**R-1**	**R-2**	**R-L**
Fine-tuning	43.13	11.76	19.08
Prompt-tuning	40.87	9.66	18.04
Prefix-tuning	41.35	10.01	18.34
Ours	46.34	12.30	20.92

Table 8: Comparison with prompt variants on ICSI.

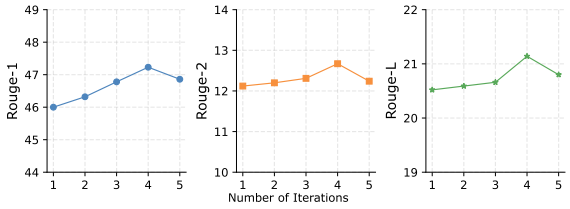

Figure 3: ROUGE scores for different iterations in our model on the ICSI dev. set.

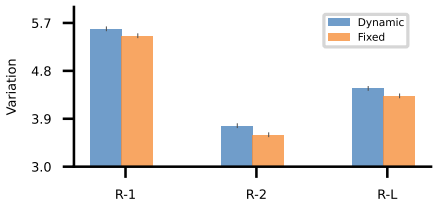

Figure 4: Effect of dynamic vs. fixed thresholding on self-training, where the y-axis represents the change in the ROUGE score.

We observe that our dynamic thresholding mechanism significantly outperforms the fixed thresholds on all metrics. During self-training, hard-to-learn dialogue samples can gradually contribute to the training when we dynamically adjust the threshold. This not only ensures the quality of pseudo-summaries, but also improves the data utilization.

6.4 Case Study

Table 9 presents a sample summary generated by different models for the SAMSum dataset. As the example shows, the summaries generated by our model generally appear more informative and relevant to the given dialogue, consistent with the re-

Dialogue	**[David]** Victor, I heard you took over Chris's company. **[Victor]** Yes. ... **[Victor]** Chris was under huge debt, but he is still working as Director after some changes. **[David]** That's good, running a company. **[Victor]** No, I sold off the office. **[Victor]** David, how is your business? **[David]** It's good, but no work these days. **[Victor]** Yes, market is very slow. **[David]** Expecting it to get better by the end of the year.
Soft Prompt	**Victor** has **taken over Chris's company.** Victor sold off the office and accommodated the employees in his office.
Curriculum	**Victor** has **taken over Chris's company.** Chris was **under huge debt.** Victor **sold off the office** and accommodated them in his office. Victor's business is slow but expected to **get better by the end of the year.**
Heterogeneous	**Victor took over Chris's company.** Chris was **under huge debt**, but he is still **working as Director.** Victor **sold off the office** and accommodated the employees in his office. Victor's business is good but slow.
Reference	Victor took over Chris's company. It was under huge debt, he sold the office and did some changes but Chris still works as Director. David's business goes very slow but he expects it to get better by the end of the year.

Table 9: Example. The first and last rows are a dialogue and its summary. Different colors highlight three topics, which correspond to three sentences in the reference. Important contents generated by different models are in bold.

sults of the human evaluation. Furthermore, we observe that the prompt-tuning model summarizes important information in a single utterance. However, the form of the dialogue determines that salient information is often scattered across multiple utterances, which results in the generated summaries being unable to summarize the entire dialogue. The summary generated by our curriculum prompting covers all salient contents, thanks to the interpolated prompts effectively integrating the contents captured by another two prompts. The summaries generated by our model appear more informative, presumably because combined with self-training to optimize prompts, the model further acquires capabilities for dialogue understanding and modeling.

7 Conclusion

In this work, we propose a novel curriculum-based prompt learning method with self-training that improves the dialogue understanding and modeling capabilities and reduces the dependency on labeled dialogue. We further explore topic-aware prompting to aid in planning for the summary generation. Experiments on diverse datasets with several different settings confirm the effectiveness of our model on abstractive dialogue summarization.

Limitations

Currently, we use a curriculum learning strategy to gradually increase the degree of three types of prompts to accomplish common dialogue summarization tasks. This may be less effective in more complex scenarios. Moreover, additional techniques may be needed for other complex tasks that require specific domain knowledge, complex intents, and fine-grained features. Since our current design provides the model with shared prompts, this approach may fail in future scenarios that require instance-aware prompts for personalized dialogue summary generation. In addition, self-training often requires a large amount of data as well as extra computation efforts with many iterations. Our model may not work well if there exists insufficient labeled data to train an initial base model that misclassifies a certain amount of unlabeled data. In this case, the mislabeled examples may greatly affect the results in subsequent iterations.

Acknowledgements

This work was supported by the National Innovation 2030 Major S&T Project of China (No. 2020AAA0104200 & 2020AAA0104205), National Natural Science Foundation of China (No. 62006077), Shanghai Sailing Program (No. 20YF1411800), the Science and Technology Commission of Shanghai Municipality (No. 21511100100), and Qingpu Scientific Research Project (No. 2021-6).

References

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

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Jiaao Chen and Diyi Yang. 2021. Simple conversational data augmentation for semi-supervised abstractive dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6605–6616.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157.

Yiming Chen, Yan Zhang, Chen Zhang, Grandee Lee, Ran Cheng, and Haizhou Li. 2021. Revisiting self-training for few-shot learning of language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9125–9135.

Alexander Richard Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. Convosumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 6866–6880.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021a. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. In *Chinese Computational Linguistics: 20th China National Conference*, pages 127–142.

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

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021c. Language model as an annotator: Exploring dialogpt for dialogue summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1479–1491.

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.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop*, pages 735–742.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. Ppt: Pre-trained prompt tuning for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8410–8423.

Iryna Gurevych and Michael Strube. 2004. Semantic similarity applied to spoken dialogue summarization. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 764–770.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *International Conference on Learning Representations*.

Karl Moritz Hermann, Tomáš Kočiskỳ, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 1693–1701.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1.

Jia Jin Koay, Alexander Roustai, Xiaojin Dai, Dillon Burns, Alec Kerrigan, and Fei Liu. 2020. How domain terminology affects meeting summarization performance. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5689–5695.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.

Yuejie Lei, Fujia Zheng, Yuanmeng Yan, Keqing He, and Weiran Xu. 2021. A finer-grain universal dialogue semantic structures based model for abstractive dialogue summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1354–1364.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

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.

Changmao Li and Jinho D. Choi. 2020. Transformers to learn hierarchical contexts in multiparty dialogue for span-based question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 5709–5714.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4582–4597.

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

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

Zhengyuan Liu and Nancy Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al. 2005. The ami meeting corpus. In *Proceedings of the 5th international conference on methods and techniques in behavioral research*, volume 88, page 100.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Gabriel Murray, Steve Renals, Jean Carletta, and Johanna D Moore. 2006. Incorporating speaker and discourse features into speech summarization. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 367–374.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55.

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*, pages 1073–1083.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021a. Controllable abstractive dialogue summarization with sketch supervision. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5108–5122.

Xueqing Wu, Yingce Xia, Jinhua Zhu, Lijun Wu, Shufang Xie, Yang Fan, and Tao Qin. 2021b. mixseq: A simple data augmentation methodfor neural machine translation. In *Proceedings of the 18th International Conference on Spoken Language Translation*, pages 192–197.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 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*, pages 270–278.

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.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. 2020. Rethinking pre-training and self-training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 3833–3845.