Hypernetwork-Assisted Parameter-Efficient Fine-Tuning with Meta-Knowledge Distillation for Domain Knowledge Disentanglement

Changqun Li[1], **Linlin Wang**[1,3,*], **Xin Lin**[1,2,†], **Shizhou Huang**[1], **Liang He**[1,2]
[1] School of Computer Science and Technology, East China Normal University
[2] Shanghai Key Laboraiory of Multdimensional lnfomation Procesing
[3] Shanghai Artificial Intelligence Laboratory
52215901009@stu.ecnu.edu.cn, {llwang,xlin,lhe}@cs.ecnu.edu.cn, huangshizhou@ica.stc.sh.cn

Abstract

Domain adaptation from labeled source domains to the target domain is important in practical summarization scenarios. However, the key challenge is domain knowledge disentanglement. In this work, we explore how to disentangle domain-invariant knowledge from source domains while learning specific knowledge of the target domain. Specifically, we propose a hypernetwork-assisted encoder-decoder architecture with parameter-efficient fine-tuning. It leverages a hypernetwork instruction learning module to generate domain-specific parameters from the encoded inputs accompanied by task-related instruction. Further, to better disentangle and transfer knowledge from source domains to the target domain, we introduce a meta-knowledge distillation strategy to build a meta-teacher model that captures domain-invariant knowledge across multiple domains and use it to transfer knowledge to students. Experiments on three dialogue summarization datasets show the effectiveness of the proposed model. Human evaluations also show the superiority of our model with regard to the summary generation quality.

1 Introduction

Recently, domain adaptation for text summarization has attracted much research interest (Zhang et al., 2020a; Yang et al., 2020; Yu et al., 2021; Zou et al., 2021). Most prior work performs pre-training on large-scale out-of-domain datasets and then adapts to the in-domain summary data. For dialogue summarization, a couple of studies have leveraged large-scale summary data that is fairly distinct from the dialogue domain, $e.g.$, the news domain, to facilitate dialogue summarization (Yu et al., 2021; Zou et al., 2021) in few-shot settings. However, this fails to acknowledge the huge gap between dialogue and general articles, $e.g.$, that dialogue involves a dynamic information exchange flow with multiple interlocutors (Li et al., 2022). Recent work explored prompt-based fine-grained transfer learning between various dialogue domains in zero-shot settings (Zhao et al., 2022a,b). However, these studies did not consider how to transfer knowledge from the source domains to the target domain, and Zhong et al. (2022) pointed out that directly fine-tuning the prompt initialized with the source prompt on target domain might lead to catastrophic forgetting of source knowledge.

Considering a typical example in Figure 1, domain adaptation aims to improve the generalizability of the model from the source domains to the target domain, however, the key challenge is the disentanglement of domain knowledge, whereby various domains contain domain-invariant and domain-specific knowledge which are always entangled. For example, we may take `Academic` and `Product` as source domains, and `Committee` as the target domain, where `Academic` consists of academic meetings, `Committee` contains formal discussions on a wide range of issues ($e.g.$, the energy market), and `Product` focuses on product design in an industrial setting. Although the content discussed in the three domains is different, the key characteristics of the dialogue are the same ($e.g.$, multiple participants, and a dynamic information exchange flow). This phenomenon suggests that the model needs to learn domain-invariant characteristics (that is, characteristics of the dialogue) in the source domains while focusing on what is being discussed in the specific domain.

Inspired by the recent success of performing new tasks through the use of instructions alone (Brown et al., 2020), and considering the inherent problems faced by domain adaptation, in this work, we propose a novel hypernetwork-assisted encoder-decoder based architecture with parameter-efficient fine-tuning, which leverages a

* Corresponding author.
† Corresponding author.

Findings of the Association for Computational Linguistics: NAACL 2024, pages 1681–1695

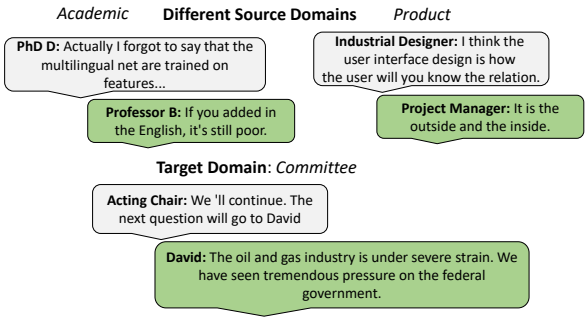

Figure 1: Example of cross-domain text summarization.

hypernetwork instruction learning (HIL) module to generate domain-specific parameters ($e.g.$, decoder adapters) for the underlying pre-trained language model (PLM). Further, to better disentangle and transfer knowledge from source domains to the target domain, we introduce a meta-knowledge distillation strategy to build a meta-teacher model that captures domain-invariant knowledge across multiple domains and use it to transfer knowledge to students. Extensive experiments on different benchmark datasets evince the effectiveness of our model for low/zero-resource dialogue summarization. To sum up, our contributions are:

- Our Hypernetwork Instruction Learning (HIL) module can generate domain-specific parameters by incorporating task and domain-related instructions.
- Our meta-knowledge distillation strategy learns general meta-knowledge on various source domains to learn a good initialization for parameter-efficient fine-tuning and transfers domain-invariant knowledge from the source to target domains with the standard cross-domain knowledge distillation.
- We evaluate our model on three dialogue summarization datasets and obtain new state-of-the-art results in low/zero-resource scenarios.

2 Related Work

Domain Adaptation. Since texts and their summaries across diverse domains might share similarities and benefit from each other, domain adaptation for text summarization has attracted much recent research interest (Zhang et al., 2020a; Yang et al., 2020; Yu et al., 2021; Zou et al., 2021). Most prior work performs pretraining on large-scale external corpora and then adapts to the in-domain summary data. For dialogue summarization, although it is more ideal to perform adaptation from a source dialogue domain to a target dialogue domain (Wang and Cardie, 2013), unfortunately, the inadequacy of available dialogue summaries makes this impossible.

Recent work explored prompt-based domain adaptation for zero-shot dialogue summarization (Zhao et al., 2022b,a). However, this line of work ignores how to transfer knowledge learned from the source domains to the target domain. In this paper, we leverage the knowledge distillation technique to transfer knowledge from source domains to the target domain and effectively alleviate catastrophic forgetting caused by direct fine-tuning.

Parameter-Efficient Fine-Tuning and Hypernetworks. A variety of parameter-efficient methods that only fine-tune a small number of (extra) parameters to attain strong performance have been proposed, including adapters (Houlsby et al., 2019), prefix-tuning (Li and Liang, 2021), and LoRA (Hu et al., 2022). Recent studies have shown the effectiveness of establishing connections between them (Pfeiffer et al., 2021; He et al., 2022a).

Hypernetworks (Ha et al., 2017; Schmidhuber, 1992) have slowly gained popularity in multitask and multilingual setups due to the positive transfer between tasks through the shared hypernetwork while reducing negative transfer by allowing unique generated parameters per task. Several approaches (Tay et al., 2021; Karimi Mahabadi et al., 2021; He et al., 2022b) learn per-task embeddings along with a shared hypernetwork to generate task-specific adapters or soft prompt modules. Inspired by this, we explore hypernetwork-based adaptation methods to learn specific knowledge for cross-domain dialogue summarization.

Meta-Learning and Knowledge Distillation. Meta-Learning, or learning about learning, aims to improve the learning algorithm itself. A prominent meta-learning framework is Model-Agnostic Meta-Learning (MAML), proposed by (Finn et al., 2017). MAML can be applied directly to any learning problem and leads to strong results with a small amount of training data. Recent studies have shown that meta-learning can improve generalization ability across domains (Finn et al., 2017; Pan et al., 2021) and in many few-shot and zero-shot settings (Campagna et al., 2020; Bao et al., 2020).

Knowledge distillation (KD) plays an impor-

tant role in transfer learning in many subfields of NLP (Hahn and Choi, 2019; Ding et al., 2021; Zhong et al., 2022). Recent work has explored meta-learning methods for knowledge distillation. Pan et al. (2021) proposed a framework for model compression by training a meta-teacher across domains and then transferring the knowledge from the meta-teacher to the student. MetaDistil (Zhou et al., 2022) allows the teacher to learn to teach dynamically. Unlike the above methods, we incorporate meta-learning on KD to disentangle and transfer domain-invariant knowledge from source domains and learn a better initialization for parameter-efficient fine-tuning.

3 Approach

Overview. To address the domain knowledge disentanglement, we propose a novel hypernetwork-assisted encoder-decoder based architecture with parameter-efficient fine-tuning. Figure 2 gives an overview of our approach, which leverages a hypernetwork instruction learning (HIL) module to generate domain-specific parameters (decoder adapters) for the underlying pre-trained language model (PLM), and applies meta-knowledge distillation to disentangle and transfer domain-invariant features for parameter-efficient cross-domain learning.

3.1 Hypernetwork-Assisted Architecture

Underlying Model. The underlying model can be any pre-trained encoder-decoder model with additional parameter-efficient submodules ($e.g.$, prefix-tuning, and adapters). In particular, this model is an extension of the prominent MAM model (He et al., 2022a), which is a unified framework that allows for the transfer of design elements across various submodules. Recall that each Transformer layer consists of an attention block and a feed-forward block, each followed by a skip-connection (Vaswani et al., 2017). Specifically, the underlying model further allows prefix-tuning with a small length l to prepend trainable tokens for multi-head attention, and inserts adapter modules with adapter size r after the feed-forward layer of the Transformer.

As depicted in Figure 2, we input a text x to the underlying model, aiming to generate a succinct summary y. The objective is to minimize the negative log-likelihood:

$$\mathcal{L}_{\text{NLL}} = -\sum_{l=1}^{L} \log(p(y_l \mid y_{1:l-1}, x)) \quad (1)$$

where y_l denotes the l-th token in the target summary and $y_{1:l-1}$ are the first $l-1$ tokens.

Hypernetwork Instruction Learning Module. To provide the underlying model with task-specific trainable parameters, our architecture exploits a novel hypernetwork-based network to generate domain-specific adapter parameters for the decoder, which are strongly based on the encoded inputs accompanied by task-related instructions. More precisely, we first create several task instructions with domain-specific descriptions, which are further converted using a pre-trained hyperencoder.

The schema of constructed instructions includes:

- A task instruction for summarization "*To generate a summary in such a way that the context should be present in input*"
- Domain-related instructions for various domains ($e.g.$, "*This input focuses on product design in an industrial setting*" for `Product` domain of QMSum (Zhong et al., 2021))

We construct a single domain-related instruction per domain based on respective dataset descriptions that request the model to summarize input in a custom domain-specific way and format the above two types of instructions by adding placeholders. More examples are provided in Appendix D.

As shown in Figure 2, the hypernetwork instruction learning (HIL) module encodes all instructions using a HyperEncoder, followed by an integration operation to append the encoded inputs that correspond with the encoder output of the underlying model. The HIL module thereby leverages a Parameter Generator at the top layer to generate decoder adapter parameters conditioned on the integrated vector $\mathbf{e} = [\text{Mean}(\mathbf{h_I}); \text{Mean}(\mathbf{h_D})]$, where $\mathbf{h_I}$ is the instruction representation with a HyperEncoder, $\mathbf{h_D}$ is the hidden representation provided by the encoder output of our underlying model, and $\text{Mean}(\cdot)$ refers to the mean pooling operation.

In addition, we further concatenate e with a learnable layer embedding $\mathbf{e}^{\mathbf{l}}$ to ensure diverse adapter parameters at the i-th Transformer layer. Specifically, we use a two-layer Parameter Generator to

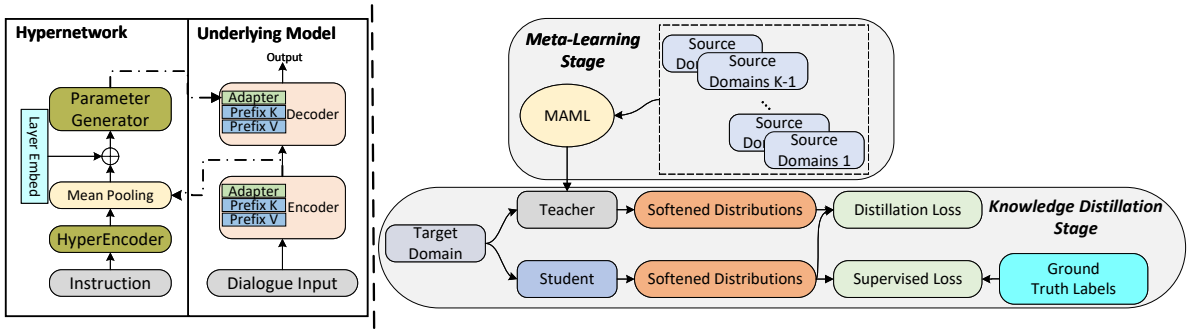

Figure 2: Overview of our proposed approach. **Left**: Hypernetwork instruction learning (HIL) assisted module that generates domain-specific parameters for the underlying pre-trained language model, where the HIL module exclusively generates decoder adapter parameters. **Right**: Pipelines of meta-knowledge distillation (MKD). First, we train our model (shown on the left) with a model-agnostic meta-learning algorithm across different source domains. Then, we leverage the knowledge distillation strategy to transfer the domain-invariant knowledge from the source domains to the target domains for parameter-efficient cross-domain learning.

produce the adapter parameters ϕ according to:

$$\mathbf{h} = \mathrm{ReLU}\left(\mathbf{W}_{i,0}([\mathbf{e};\mathbf{e_I}]) + \mathbf{b_{i,0}}\right) \quad (2)$$

$$\phi_i = \mathbf{W}_{i,1}(\mathbf{h}) + \mathbf{b}_{i,1} \quad (3)$$

where $\mathbf{W}_{i,0}$, $\mathbf{b}_{i,0}$, $\mathbf{W}_{i,1}$, $\mathbf{b}_{i,1}$ are trainable parameters, and the generated parameters ϕ_i are sliced and reshaped to form the adapter parameter [$\mathbf{W}_{iu}$, $\mathbf{W}_{id}$, $\mathbf{b}_{iu}$, $\mathbf{b}_{id}$] at the i-th layer (a brief introduction of the adapter in Appendix A).

3.2 Meta-Knowledge Distillation

To disentangle and transfer domain-invariant knowledge from source domains, we further propose to use meta-knowledge distillation (MKD), a model-agnostic training approach for parameter-efficient cross-domain learning. Most notably, we incorporate meta-learning to disentangle domain-invariant knowledge and learn a better initialization for parameter-efficient fine-tuning based on the general meta-knowledge on various source domains. Subsequently, we transfer crucial domain-invariant knowledge from the source to target domains with standard cross-domain knowledge distillation.

Specifically, we apply meta-learning with a two-step gradient update to learn general meta-knowledge among multiple source domains for better parameter initialization. First, we randomly initialize the parameters θ of our model, which corresponds to our hypernetwork-assisted architecture. We sample n training instances for the k-th source domain S_k to calculate the average loss $\mathcal{L}_{\mathcal{S}_k}(f_{\boldsymbol{\theta}})$, where $f_{\boldsymbol{\theta}}$ refers to the output of the model. Here, we use gradient descent to update parameters and obtain a temporary $\boldsymbol{\theta}'_k$.

$$\boldsymbol{\theta}'_k = \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}_k}(f_{\boldsymbol{\theta}}) \quad (4)$$

Then, we use $\boldsymbol{\theta}'_k$ to recalculate the new corresponding loss and sum up the loss values over all source domains, aiming to accomplish the second-step update. More precisely, we update the parameters of our model by minimizing the meta-learning objective function as follows:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \beta \nabla_{\boldsymbol{\theta}} \sum_{S_k} \mathcal{L}_{S_k}(f_{\boldsymbol{\theta}'_k}) \quad (5)$$

We continue optimizing the model until the validation accuracy of source domains stops increasing.

Similar to the approach of (Pan et al., 2021) for classification tasks ($e.g.$, natural language inference, and sentiment analysis), we exploit standard knowledge distillation to transfer domain-invariant knowledge across domains for summarization tasks. Let $f_{\boldsymbol{\theta_t}}$ and $f_{\boldsymbol{\theta_s}}$ denote teacher and student, respectively. As shown in Figure 2 (right), we rely on a teacher network with transferable knowledge digested across source domains to provide guidance for the student network. As shown in Algorithm 1, the student network is trained on the target domain with guidance from both the meta-teacher and the supervision of ground-truth summaries. Specifically, the student network is trained with the supervision of target summaries, and the softened distributions predicted by the teacher network can be formulated as:

$$\mathcal{L}_{\mathrm{KD}} = \mathbb{E}\left(\sum_{l=1}^{L}(f_{\boldsymbol{\theta_t}}(y_l \mid y_{1:l-1}, x) - f_{\boldsymbol{\theta_s}}(y_l \mid y_{1:l-1}, x))^2\right) \quad (6)$$

Algorithm 1 Meta-Knowledge Distillation

Input: The given dialogues from source domains S and the target domain T; α; β;
Output: An optimal student $f_{\boldsymbol{\theta_s}}$;

Initialize our model with random parameters θ

/* **Meta-Learning Stage**. */
while not done **do**
  for all $S_k \in S$ **do**
    Evaluate $\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}_k}(f_{\boldsymbol{\theta}})$ with respect to samples in S_k;
    Update parameters with gradient descent: $\boldsymbol{\theta}'_k = \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{S}_k}(f_{\boldsymbol{\theta}})$;
  end for
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \beta \nabla_{\boldsymbol{\theta}} \sum_{S_k} \mathcal{L}_{S_k}(f_{\boldsymbol{\theta}'_k})$;
end while

/* **Knowledge Distillation Stage**. */
Initialize the teacher network $f_{\boldsymbol{\theta_t}}$ with the meta-updated θ parameters;
Initialize the student network $f_{\boldsymbol{\theta_s}}$ with random parameters;
Train the student network on the target domain T using the supervised and distillation loss. $\mathcal{L}_{\text{all}} = (1-\lambda)\,\mathcal{L}_{\text{NLL}} + \lambda\,\mathcal{L}_{\text{KD}}$
return $f_{\boldsymbol{\theta_s}}$.

where $f_{\boldsymbol{\theta_t}}(y_l \mid y_{1:l-1}, x)$ and $f_{\boldsymbol{\theta_s}}(y_l \mid y_{1:l-1}, x)$ are model outputs from the teacher and student network, respectively.

The overall loss function $\mathcal{L}_{\text{all}}$ of the student network can be formulated as:

$$\mathcal{L}_{\text{all}} = (1-\lambda)\,\mathcal{L}_{\text{NLL}} + \lambda\,\mathcal{L}_{\text{KD}} \tag{7}$$

where λ is a hyperparameter to balance the influence of each loss, $\mathcal{L}_{\text{NLL}}$ refers to the negative log-likelihood loss in Eq. 1. We use mean squared error between the hidden states of the teacher and the student in Eq. 6. Note that the comparison results of different distillation losses are shown in Appendix E.1.

3.3 Experimental Setup

Datasets. We conduct experiments on three multi-domain summarization datasets, QMSum (Zhong et al., 2021), TODSum (Zhao et al., 2021), and DialogSum (Chen et al., 2021). Specifically, QMSum comprises meeting transcriptions from the `Academic`, `Committee`, and `Product` domains, while TODSum consists of task-oriented dialogues that originate from `Restaurant`, `Hotel`, `Attraction`, `Taxi`, and `Train` domains. DialogSum was collected from diverse daily-life scenarios spanning a wide variety of topics. Detailed dataset statistics are given in Appendix B.1.

Automatic Metrics. To assess the quality of generated summaries, we use standard evaluation metrics ROUGE-1, ROUGE-2, and ROUGE-L, which consider the overlapping uni-grams, bi-grams, and longest common subsequence scores (Lin, 2004)[1], respectively. Furthermore, we report the BERTScore as well, which is highly correlated with human judgement (Zhang et al., 2020b).

Human Evaluation. We conduct a human evaluation along three criteria: (1) **Fluency** evaluates the readability of the generated summaries. (2) **Informativeness** evaluates how well the generated summaries capture more salient information. (3) **Relevance** evaluates how well the generated summaries reflect the input document. Specifically, we randomly sample 200 dialogues for the DialogSum dataset and ask three annotators to rate the quality of generated summaries on a scale of 1.0 to 5.0 using the three criteria (the higher the better). We also regard ChatGPT as a human evaluator and give it evaluation instruction via different prompts. Evaluation results and details are provided in Appendix F.

Baselines and Experimental Settings. We compare our method with several representative baselines including (1) PGN (See et al., 2017), (2) BART (Lewis et al., 2020), (3) Adapter (Houlsby et al., 2019), (4) Prefix-tuning (Li and Liang, 2021), (5) MAM (He et al., 2022a). More comparison details are provided in Appendix B.2.

We use the HuggingFace implementation (Wolf et al., 2020) of the $\text{BART}_{\text{large}}$ model (Lewis et al., 2020). During training, we set the batch size to 16, prefix length l to 30, adapter size r to 400, define the number of training epochs as 30, and leverage AdamW optimization (Loshchilov and Hutter, 2017) together with a linear learning rate scheduler. The hyperparameter α in Eq. 4 is chosen as 5×10^{-5}, β in Eq. 5 is set to be 4×10^{-5}, and λ in Eq. 7 is 0.2. As for decoding, we set the beam size as 6, and the length normalization to be 0.8.

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

Models	Automatic Metrics	Human Ratings
	R-1 / R-2 / R-L / BERTScore	Fluency / Info. / Relevance
PGN	28.74 / 10.56 / 26.17 / 0.25	3.02 / 2.91 / 2.45
ChatGPT	37.86 / 16.36 / 35.51 / 0.34	4.00 / 3.45 / 3.16
$\text{BART}_{\text{large}}$	46.72 / 20.84 / 44.70 / 0.52	4.34 / 4.34 / 4.48
Adapter ($\text{BART}_{\text{large}}$)	43.50 / 19.28 / 42.02 / 0.46	4.00 / 4.06 / 3.82
Prefix-tuning ($\text{BART}_{\text{large}}$)	46.13 / 20.55 / 44.05 / 0.52	4.23 / 4.29 / 4.33
MAM ($\text{BART}_{\text{large}}$)	46.93 / 20.64 / 44.57 / 0.52	4.38 / 4.36 / 4.52
Ours	**47.00 / 20.94 / 45.01**† **/ 0.53**	**4.67 / 4.39 / 4.81**†

Table 1: Comparison of results on DialogSum (the target domain), where the source domain is a mixture of QMSum and TODSum datasets. The ChatGPT results are obtained by In-Context Learning with `gpt-3.5-turbo` API. We report the average of multi-reference results. † indicates a significant difference with the second best result (t-test, p-value$<$0.05).

3.4 Main Results

We first integrate QMSum and TODSum as the source domain and take DialogSum as the target for experiments. Table 1 provides a comparison with previous approaches on DialogSum, which shows that our model achieves new state-of-the-art results. For instance, compared to the previously best-performing model MAM (He et al., 2022a), our model obtains relative gains of 1.5% on ROUGE-2, 1.0% on ROUGE-L, and 1.9% on BERTScore. Simultaneously, it surpasses all baselines in the human evaluation, demonstrating that our model can deliver high-quality summaries. Most importantly, our parameter-efficient model outperforms the BART (Lewis et al., 2020) fine-tuning based architecture, on both automatic and human metrics, confirming the effectiveness of our model-agnostic cross-domain learning strategy. ChatGPT underperforms BART fine-tuning across all metrics. This may be because the responses from ChatGPT are usually more verbose, resulting in lower ROUGE scores.

To further conduct fine-grained cross-domain adaptation, for QMSum and TODSum, we regard each individual domain in the dataset as the target, merging the others into an integrated source domain. These experimental settings are severely challenging since there exists a limited number of training instances in these two datasets (e.g., 158 examples in the `Attraction` domain of TODSum). Table 2 provides a comparison with prior approaches for multi-source cross-domain summarization on TODSum (Top) and QMSum (Bottom), respectively. We can observe that our model achieves state-of-the-art results on these two datasets with limited training instances, suggesting the domain adaptation ability of the proposed approach across diverse domains. For instance, when taking `Restaurant` as the target domain, our approach yields relative improvements of 1.4%, 2.8%, and 1.5% compared with the previous state-of-the-art model BART (Lewis et al., 2020) in terms of ROUGE-1, ROUGE-2, and ROUGE-L scores.

Zero-shot settings. In addition, we explore the performance of our model in zero-shot settings for TODSum and QMSum. The zero-shot setting evaluates the effectiveness of our model with meta-learning. First, we train our model with the model-agnostic meta-learning algorithm on various source domains. Subsequently, we directly transfer the learned domain-invariant knowledge to the target domain for evaluation. Table 3 reports the corresponding results of zero-shot cross-domain summarization. Our model achieves strong results compared with previous approaches, further confirming the adaptation capabilities of our summarization model on unseen domains of dialogue.

4 Quantitative Analysis

4.1 Ablation Study

To verify the effectiveness of different components in our model, we conduct ablation studies by removing each module from our architecture. Table 4 provides the results of these ablations on the `Committee` domain of QMSum, where we observe that all of the components in our model make significant contributions. For instance, the removal of our hypernetwork causes a relative performance drop on all ROUGE scores (e.g., 4.0% on ROUGE-2), confirming the validity of leveraging a hypernetwork to encode domain-related

Target Domain	Train	Taxi	Restaurant	Hotel	Attraction
Models	**R-1 / R-2 / R-L**	**R-1 / R-2 / R-L**	**R-1 / R-2 / R-L**	**R-1 / R-2 / R-L**	**R-1 / R-2 / R-L**
$\text{BART}_{\text{large}}$	89.30 / 83.69 / 88.59	80.83 / 63.94 / 76.57	92.05 / 82.48 / 90.02	89.12 / 81.46 / 88.10	85.56 / 72.55 / 84.76
Adapter	87.28 / 79.71 / 88.05	80.73 / 64.25 / 76.95	92.01 / 81.75 / 89.17	81.87 / 69.95 / 80.70	83.91 / 70.64 / 83.28
Prefix-tuning	88.30 / 81.24 / 88.63	81.87 / 66.94 / 79.02	89.38 / 77.05 / 86.77	88.60 / 79.94 / 87.71	82.61 / 66.87 / 82.30
MAM	87.59 / 80.05 / 86.32	79.37 / 60.84 / 74.36	91.13 / 81.75 / 89.29	89.14 / 81.04 / 88.38	79.01 / 64.14 / 79.42
Ours	**90.39 / 84.43 / 89.32**	**82.07 / 67.01 / 79.51**	**93.36 / 84.82 / 91.35**	**89.95 / 82.30 / 88.81**	**85.74 / 72.94 / 85.29**

Target Domain	Academic	Committee	Product
Models	**R-1 / R-2 / R-L**	**R-1 / R-2 / R-L**	**R-1 / R-2 / R-L**
$\text{BART}_{\text{large}}$	34.22 / 8.79 / 29.01	43.75 / 20.03 / 36.92	**42.16** / 15.51 / 33.17
Adapter	30.11 / 7.39 / 27.28	41.70 / 19.15 / 36.69	39.44 / 14.93 / 32.49
Prefix-tuning	31.46 / 8.83 / 27.76	41.16 / 18.21 / 34.76	38.29 / 14.77 / 32.35
MAM	32.98 / 9.25 / 29.09	42.70 / 19.46 / 36.44	40.52 / 15.32 / 33.10
Ours	**34.31 / 10.49 / 29.95**	**43.85 / 21.54 / 38.54**	41.75 / **16.49 / 33.62**

Table 2: Comparison of results on TODSum (Top) and QMSum (Bottom) datasets, respectively.

Target Domain	Train	Taxi	Restaurant	Hotel	Attraction
Models	**R-1 / R-2 / R-L**	**R-1 / R-2 / R-L**	**R-1 / R-2 / R-L**	**R-1 / R-2 / R-L**	**R-1 / R-2 / R-L**
$\text{BART}_{\text{large}}$	53.59 / 30.57 / 51.00	42.18 / 19.53 / 41.89	53.18 / 26.39 / 52.22	51.10 / 24.69 / 49.69	68.19 / 45.26 / 66.67
Adapter	49.75 / 25.15 / 45.94	47.50 / 24.21 / 45.48	53.91 / 27.05 / 51.96	46.91 / 20.37 / 45.33	64.66 / 40.87 / 63.21
Prefix-tuning	51.16 / 27.86 / 48.88	43.79 / 21.03 / 43.09	53.91 / 27.19 / 54.43	50.99 / 24.15 / 49.68	65.82 / 43.49 / 66.29
MAM	57.24 / 34.18 / 52.89	45.47 / 21.49 / 44.49	53.00 / 24.91 / 52.86	48.18 / 21.67 / 47.46	67.81 / 46.18 / 66.64
Ours	**61.31 / 40.66 / 58.89**	**50.36 / 28.65 / 47.71**	**57.64 / 30.61 / 56.39**	**52.97 / 25.97 / 50.81**	**72.68 / 51.08 / 72.43**

Target Domain	Academic	Committee	Product
Models	**R-1 / R-2 / R-L**	**R-1 / R-2 / R-L**	**R-1 / R-2 / R-L**
ChatGPT	26.84 / 5.11 / 22.70	37.23 / 11.54 / 29.63	35.43 / 9.18 / 27.34
$\text{BART}_{\text{large}}$	32.02 / 7.70 / **26.78**	40.34 / 16.87 / 34.20	**38.63** / 12.70 / 30.94
Adapter	28.99 / 5.94 / 24.69	38.67 / 15.61 / 32.33	35.52 / 11.82 / 29.62
Prefix-tuning	29.28 / 6.73 / 25.08	37.95 / 15.74 / 33.36	35.49 / 12.91 / 30.18
MAM	31.26 / 7.02 / 25.69	40.45 / 16.99 / 34.11	36.24 / 12.19 / 30.49
Ours	**32.25 / 7.79** / 26.58	**40.67 / 17.29 / 34.20**	37.76 / **13.39 / 31.82**

Table 3: Comparison of results on TODSum (Top) and QMSum (Bottom) with zero-shot settings, respectively. We use ChatGPT `gpt-3.5-turbo` for zero-shot settings.

Model	R-1	R-2	R-L
Ours			
– w/ $\text{BART}_{\text{large}}$	**43.85**	**21.54**	**38.54**
Ablations			
– w/o hypernetwork	43.08	$20.68_{\downarrow \mathbf{4.0\%}}$	37.71
– w/o knowledge distillation	42.30	$19.61_{\downarrow \mathbf{9.0\%}}$	36.95
– w/o meta-learning	42.85	$20.28_{\downarrow \mathbf{5.8\%}}$	37.01

Table 4: Ablation study on `Committee` domain of the QMSum dataset.

Model	R-1	R-2	R-L
variants			
without instructions	46.16	20.29	44.15
with simple instructions	46.60	20.73	44.87
with hypernetwork instructions	**47.00**	**20.94**	**45.01**

Table 5: Comparison with instruction variants on Dialogsum (the target domain), where the source domain is a mixture of QMSum and TODSum.

instructions and generate better parameters for the adapters. The ablation of knowledge distillation causes relative performance drops of 3.5%, 9.0%, and 4.1% on ROUGE-1, ROUGE-2, and ROUGE-L, showing the effectiveness of distillation in boosting the cross-domain adaptation abilities. Furthermore, we also conclude that the usage of meta-learning enables our model to learn better initialization parameters during parameter-efficient tuning.

4.2 Multi-source vs. Single-source

We further compare the aforementioned multi-source adaptation with single-source domain adaptation on the QMSum dataset with special procedures. For instance, when taking `Committee` as the target domain, we regard either `Academic` or `Product` as the source for single-source domain adaptation and leverage the mixture of `Academic` and `Product` to serve as the source for the multi-source setting. Table 6 reports the corresponding results, from which crucial conclusions can be drawn from different perspectives. Our model achieves

S→T		BART	Prefix-tuning	Adapter	MAM	Ours
Single-source						
	Similarity			R-1 / R-2 / R-L		
A→C	0.75	38.30/15.81/32.89	30.23/10.88/28.52	34.83/12.90/29.32	37.10/13.95/32.37	**39.70/16.79/34.19**
P→C	0.77	40.67/17.62/35.24	38.12/15.87/33.46	37.92/14.78/31.39	40.85/17.47/34.32	**41.28/18.07/35.12**
Multi-source						
⟶C		43.75/20.03/36.92	41.70/19.15/36.69	41.16/18.21/34.76	42.70/19.46/36.44	**43.85/21.54/ 38.54**
Single-source						
	Similarity			R-1 / R-2 / R-L		
A→P	0.73	35.67/11.69/29.64	33.42/11.34/28.89	37.20/11.81/29.81	36.59/12.36/30.60	35.53/10.65/29.25
C→P	0.77	33.36/10.31/27.77	32.57/10.42/27.10	32.84/10.39/26.67	33.24/10.87/27.57	**36.59/11.56/29.70**
Multi-source						
⟶P		42.16/15.51/33.17	39.44/14.93/32.49	38.29/14.77/32.35	40.52/15.32/33.10	41.75/ **16.49/33.62**

Table 6: Comparison of single and multi-source domain adaptation on QMSum. "S" and "T" refer to source and target domains. "A", "C" and "P" are domain abbreviations for `Academic`, `Committee`, and `Product`.

better results on single-source adaptation with a greater similarity between the source and target domains. In general, multi-source adaptation can yield better results in terms of ROUGE scores compared with single-source domain adaptation.

4.3 Cross-Domain Transferability

We further study the performance of cross-domain transferability with two commonly used metrics, including cosine similarity and the overlapping rate of activated neurons in the network (Su et al., 2022). Figure 3 depicts the comparison results for different models. It can be concluded that our model possesses superior transferability across multiple dialogue domains, surpassing all representative baselines in terms of these two metrics.

4.4 Comparison with Instruction Variants

We additionally investigate the effect of hypernetwork instruction learning in comparison with other variants. As reported in Table 5, the removal of instruction tuning causes a major drop in performance, and our model with hypernetwork-encoded instructions achieves the best results. The variant with simple instructions directly appends the human-written instructions to the input dialogue.

5 Case Study

We conduct a case study with an example from QMSum to illustrate the advantages of our model. Furthermore, we explore applying LLM to a specific domain of dialogue summarization through in-context learning. In Table 7, the summaries generated by our model appear more informative, presumably because it can infer essential dialogue characteristics and focus on domain-specific contents. In contrast, the BART baseline wrongly

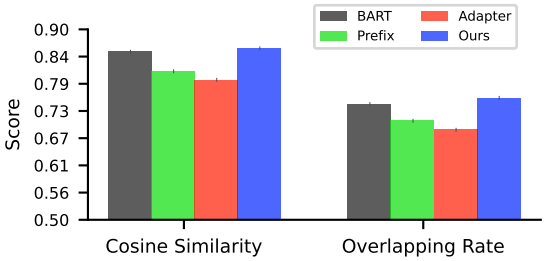

Figure 3: Cross-domain Transferability is measured by two metrics for different models on the QMSum dataset.

predicts the country name "Canada" as "British Columbia", and MAM generates too many incorrect details of dialogue, which makes the summary irrelevant and of poor quality. Indeed, the responses from ChatGPT can become more verbose when the output length is not explicitly limited. However, if the response is explicitly limited in length, there is a possibility that salient information may not be captured adequately. It's important to strike a balance between providing sufficient information and keeping the response concise.

6 Conclusion

In this work, we propose a novel hypernetwork-assisted encoder–decoder architecture with meta-knowledge distillation for domain knowledge disentanglement in cross-domain dialogue summarization. It leverages hypernetwork instruction learning to generate preferable domain-specific adapter parameters and disentangles and transfers domain-invariant features to better improve cross-domain transferability with a model-agnostic distillation strategy. Our model achieves strong results on diverse datasets with several different settings.

Source Domain	**Academic Domain:** PhD D: I forgot to say that multilingual net are trained on features. Professor B: What I we hadn't seen yet was that if you added in the english, it's still poor. **Product Domain:** Industrial Designer: I think the user interface design is he will design how the user will you know the relation between the user and the remote control. User Interface: I think industrial design's, it's the function design.
Target Domain	**Committee Domain:** David: The oil and gas industry is under severe strain... Victor: The federal liberal government's response to the anti-oil lobby was the introduction of the no more pipelines bill, bill c-69, which will prevent any major oil and gas projects from being developed in Canada.
BART	The oil and gas industry was under severe strain due to the anti-oil lobby lobby and the oil shipping ban for the northern coast of British Columbia.
MAM	The oil and gas industry is under stress due to pressure from anti-oil lobby groups. The international oil price war and the covid-19 pandemic caused a huge drop in demand for oil.
ChatGPT	... He then discussed the pressure put on the federal government from anti-oil and gas lobby groups, which resulted in the introduction of Bill C-69 and C-48, both of which had a negative effect on the oil and gas industry and caused over $200 billion of investment to leave Canada.
Ours	The oil and gas industry was under severe strain. The federal liberal government's response to this pressure was the introduction of the no more pipelines bill c-69.
Reference	The oil and gas industry was under severe strain. The federal liberal government's response to the anti-oil lobby was the introduction of the bill c-69, which would prevent any major oil and gas projects from being developed in Canada.

Table 7: Case study for QMSum dataset, where the wrong information is highlighted as pink, and redundant information is highlighted as lime.

Limitations

We leverage a hypernetwork instruction learning module to generate domain-specific parameters that encourage the model to focus on domain-specific content. The limited number of human-written instructions may be less effective in more complex scenarios. When the model is trained in the meta-learning stage, high-quality resources are required to guarantee the high quality of the results. Additionally, the effectiveness of our model is confirmed by experiments on English-language dialogue summarization benchmark datasets. However, whether it can also handle summarization tasks in multiple languages remains unexplored.

Acknowledgements

We are very grateful to the anonymous reviewers for their hard work and valuable comments. This work is supported by National Science and Technology Major Project (2021ZD0111000/2021ZD0111004), the Science and Technology Commission of Shanghai Municipality Grant (No. 21511100101, 22511105901,22DZ2229004). Xin Lin is the corresponding author.

References

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2020. Few-shot text classification with distributional signatures. In *International Conference on Learning Representations*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Giovanni Campagna, Agata Foryciarz, Mehrad Moradshahi, and Monica Lam. 2020. Zero-shot transfer learning with synthesized data for multi-domain dialogue state tracking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 122–132, Online. 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.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F Wong, Dacheng Tao, and Zhaopeng Tu. 2021. Understanding and improving lexical choice in non-autoregressive translation. In *International Conference on Learning Representations*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

David Ha, Andrew Dai, and Quoc V Le. 2017. Hypernetworks. In *In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April*, pages 24–26.

Sangchul Hahn and Heeyoul Choi. 2019. Self-knowledge distillation in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 423–430.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022a. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*.

Yun He, Steven Zheng, Yi Tay, Jai Gupta, Yu Du, Vamsi Aribandi, Zhe Zhao, YaGuang Li, Zhao Chen, Donald Metzler, et al. 2022b. Hyperprompt: Prompt-based task-conditioning of transformers. In *International Conference on Machine Learning*, pages 8678–8690. PMLR.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 565–576, Online. Association for Computational Linguistics.

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.

Changqun Li, Linlin Wang, Xin Lin, Gerard de Melo, and Liang He. 2022. Curriculum prompt learning with self-training for abstractive dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1096–1106, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 4582–4597, 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.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Haojie Pan, Chengyu Wang, Minghui Qiu, Yichang Zhang, Yaliang Li, and Jun Huang. 2021. Meta-KD: A meta knowledge distillation framework for language model compression across domains. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3026–3036, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jürgen Schmidhuber. 1992. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139.

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.

Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, Lei Hou, Maosong Sun, and Jie Zhou. 2022. On transferability of prompt tuning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3949–3969, Seattle, United States. Association for Computational Linguistics.

Yi Tay, Zhe Zhao, Dara Bahri, Donald Metzler, and Da-Cheng Juan. 2021. Hypergrid transformers: Towards a single model for multiple tasks. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Lu Wang and Claire Cardie. 2013. Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405, Sofia, Bulgaria. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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, Online. Association for Computational Linguistics.

Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve. 2020. TED: A pretrained unsupervised summarization model with theme modeling and denoising. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1865–1874, Online. Association for Computational Linguistics.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021. AdaptSum: Towards low-resource domain adaptation for abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5892–5904, Online. Association for Computational Linguistics.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *In 8th International Conference on Learning Representations, ICLR 2020*, pages 26–30.

Lulu Zhao, Fujia Zheng, Keqing He, Weihao Zeng, Yuejie Lei, Huixing Jiang, Wei Wu, Weiran Xu, Jun Guo, and Fanyu Meng. 2021. Todsum: Task-oriented dialogue summarization with state tracking. *arXiv preprint arXiv:2110.12680*.

Lulu Zhao, Fujia Zheng, Weihao Zeng, Keqing He, Ruotong Geng, Huixing Jiang, Wei Wu, and Weiran Xu. 2022a. Adpl: Adversarial prompt-based domain adaptation for dialogue summarization with knowledge disentanglement. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 245–255, New York, NY, USA. Association for Computing Machinery.

Lulu Zhao, Fujia Zheng, Weihao Zeng, Keqing He, Weiran Xu, Huixing Jiang, Wei Wu, and Yanan Wu. 2022b. Domain-oriented prefix-tuning: Towards efficient and generalizable fine-tuning for zero-shot dialogue summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4848–4862, Seattle, United States. Association for Computational Linguistics.

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.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2022. Panda: Prompt transfer meets knowledge distillation for efficient model adaptation. *arXiv preprint arXiv:2208.10160*.

Wangchunshu Zhou, Canwen Xu, and Julian McAuley. 2022. BERT learns to teach: Knowledge distillation with meta learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7037–7049, Dublin, Ireland. Association for Computational Linguistics.

Yicheng Zou, Bolin Zhu, Xingwu Hu, Tao Gui, and Qi Zhang. 2021. Low-resource dialogue summarization with domain-agnostic multi-source pretraining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 80–91, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Parameter-Efficient Methods

Adapter The adapter approach (Houlsby et al., 2019) inserts small modules (adapters) between transformer layers. The adapter layer generally uses a down-projection with $W_{\mathrm{d}} \in \mathbb{R}^{d \times r}$ to project the input h into a lower-dimensional space specified by the bottleneck dimension r, followed by a nonlinear activation function $f()$, and an up-projection with $W_{\mathrm{u}} \in \mathbb{R}^{r \times d}$. These adapters are surrounded by a residual connection, leading to a final form:

$$h \leftarrow h + f(hW_{\mathrm{d}} + b_d)W_{\mathrm{u}} + b_u \tag{8}$$

Houlsby et al. (2019) places two adapters sequentially within one layer of the transformer, one after the multi-head attention and one after the FFN sub-layer. Pfeiffer et al. (2021) have proposed a more efficient adapter variant that is inserted only after the FFN "add & layer norm" sub-layer.

Prefix-tuning Inspired by the success of textual prompting methods, prefix-tuning (Li and Liang, 2021) prepends l tunable prefix vectors to the keys and values of multi-head attention at every layer. Specifically, two sets of prefix vectors $P_k, P_v \in \mathbb{R}^{l \times d}$ are concatenated with the original key K and the value V.

$$K' = [P_k; K], V' = [P_v; V] \tag{9}$$

Then multi-head attention is performed on the new prefixed keys and values.

B Experimental Details

B.1 Dataset Statistics

Table 8 provides statistical information for DialogSum, QMSum, and TODSum datasets.

Datasets	Domain	Train/Dev/Test	Dialog.len	Summ.len
DialogSum	-	12460/500/500	131.6	21.0
QMSum	Academic	259 / 54 / 58	1562.95	77.92
	Committee	308 / 73 / 72		
	Product	690 / 145 / 151		
TODSum	Train	327 / 30 / 31	188.16	44.91
	Taxi	312 / 54 / 51		
	Restaurant	1268 / 53 / 66		
	Hotel	660 / 61 / 72		
	Attraction	158 / 11 / 13		

Table 8: Statistics of Dialogsum, QMSum, and TODSum datasets.

B.2 Baselines

We describe baselines in detail as follows.

PGN This method was proposed by (See et al., 2017). It contains a pointer mechanism and a copy mechanism and solves the Out-Of-Vocabulary (OOV) problem in abstractive summarization.

BART This model was proposed by (Lewis et al., 2020). It is a state-of-the-art abstractive summarization model pre-trained with a denoising autoencoding objective.

MAM This method was proposed by (He et al., 2022a). It provides a mix and match of the favorable designs of prefixes and adapters, allowing fewer parameters to be tuned than by previous methods while being more effective.

C LLM Evaluation

We further regard ChatGPT as a human evaluator and give it evaluation instruction via different prompts. Each prompt should specify (1) which NLG task ($e.g.$, summarization) needs to be evaluated and (2) which aspect ($e.g.$, fluency) of the generation result should be assessed currently. Evaluation criteria include: (1) **Fluency** evaluates the readability of the generated summaries. (2) **Informativeness** evaluates how well the generated summaries capture more salient information. (3) **Relevance** evaluates how well the generated summaries reflect the input document. Detailed prompts are provided in Table 12. Specifically, we randomly sample 200 dialogues for the DialogSum dataset, and ask ChatGPT to rate the quality of generated summaries on a scale of 1.0 to 5.0 using the three criteria (the higher the better).

Table 9 shows the mean LLM ratings of different models on DialogSum. The summaries generated by our model prove preferable across all three evaluation dimensions, further confirming the effectiveness of our approach.

Model	**Fluency**	**Info.**	**Relevance**
PGN	3.00	2.89	2.40
$\mathrm{BART}_{\mathrm{large}}$	4.44	4.34	4.50
Adapter ($\mathrm{BART}_{\mathrm{large}}$)	3.96	4.01	3.78
Prefix-tuning ($\mathrm{BART}_{\mathrm{large}}$)	4.13	4.19	4.23
MAM ($\mathrm{BART}_{\mathrm{large}}$)	4.36	4.34	4.50
Ours	**4.57**	**4.39**	**4.76**

Table 9: LLM evaluation on DialogSum (the target domain), where the source domain is a mixture of QMSum and TODSum datasets.

Model	R-1	R-2	R-L
CE	43.35	21.43	38.24
KL	43.72	21.41	38.25
MSE	**43.85**	**21.54**	**38.54**

Table 10: Comparison with different knowledge distillation losses on `Committee` domain (the target domain), where the source domain is a mixture of `Academic` and `Product`.

D Task and Domain-Related Instruction

Table 11 shows the task instructions for dialogue summarization and the corresponding domain instructions for the QMSum, TODSum, and DialogSum datasets.

E Quantitative Analysis

E.1 Impact of Different Knowledge Distillation Losses

We examine the impact of various knowledge distillation losses, such as KL-divergence (KL), cross-entropy (CE), and Mean Squared Error (MSE), on our model. We conduct experiments on the QMSum dataset, which contains three different domains, *i.e.*, the `Academic`, `Committee`, and `Product`. Table 10 shows the detailed results. The differences between multiple loss functions are relatively small, particularly when comparing KL-divergence and cross-entropy. Moreover, we can observe that our model with MSE achieves the best results.

E.2 Impact of Parameter-efficient Tuning

Prefix Length. We further investigate the effect of different lengths of prefix. Figure 4 (left) depicts the corresponding results when the `Committee` domain of QMSum serves as the target. As we can observe, when varying the prefix length from 20 to 100, all ROUGE scores keep improving at first, achieving the best performance at 30, and then starting to decrease. This indicates the need to leverage the prefix of appropriate lengths.

Adapter Size. We further study the effects of different adapter sizes varying from 200 to 512. In Figure 4 (right), we observe that performance is improving initially, reaching the best result at the adapter size 400, and then starting to degrade. This suggests that the learning ability of our model can be improved by increasing the size of adapters, while an excessive parameter count for adapters may be counterproductive.

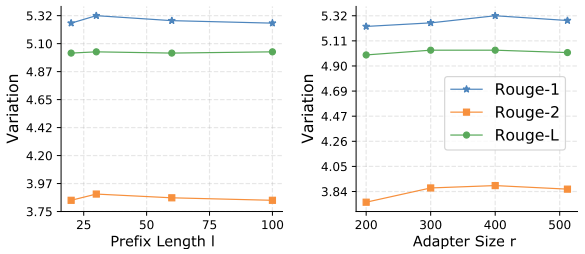

Figure 4: Output quality with different prefix lengths and adapter sizes on the `Committee` domain of QMSum.

F Evaluation Instruction

Following (Wang et al., 2023), when evaluating dialogue summarization models in terms of **Fluency**, **Informativeness** and **Relevance**, the prompt is given in Table 12.

Task	Task Instruction
Summ.	In this task, you are given a conversation, and your task is to generate a summary from the information present in the given conversation. Generate a summary in such a way that the context should be present in the conversation. It should cover the complete context of the conversation.
Domain	Domain-Related Instruction
QMSum	This dataset is a query-based meeting dataset, and your task is to summarize the contents that users are interested in and query.
Academic	These conversation focus on academic meeting, the contents of meetings are specific to the discussions about research among students.
Committee	These conversation focus on the formal discussions on a wide range of issues (e.g., the reform of the education system, public health, etc.
Product	These conversation focus on product design in an industrial setting.
TODSum	This dataset is a task-oriented dataset, and the main questions discussed are attractions, taking a taxi, or booking a restaurant / train tickets / hotel.
Train	These conversations mainly talked questions related to booking train tickets, while also asking questions related to travel.
Taxi	These conversations mainly talked questions related to taking a taxi, and users want to know the color of the car.
Restaurant	These conversations mainly talked questions related to booking a restaurant, and price-related descriptions are usually mentioned.
Hotel	These conversations mainly talked questions related to booking hotel, and users will mention the star rating.
Attraction	These conversations mainly talked questions related to attractions.
DialogSum	This dataset focuses on diverse real-life scenarios such as schooling, work, medication, shopping, leisure, travel.

Table 11: Domain-Related Instruction of QMSum, TODSum, and DialogSum.

	Evaluation Instruction
Format	Score the following [task-ins] with respect to [aspect] with one to five stars, where one star means "[ant-aspect]" and five stars means "perfect [aspect]". Note that [aspect] measures [aspect-ins]. Input: [Dialogue] Output: [Generated Summary] Stars:
Aspect	Evaluation Instruction
Fluency	Score the following dialogue summarization given the corresponding dialogue with respect to fluency with one to five stars, where one star means "disfluency" and five stars means "perfect fluency". Note that fluency measures the quality of individual sentences, are they well-written and grammatically correct. Consider the quality of individual sentences. Input: [a dialogue] Output: [one generated summary] Stars:
Informative	Score the following dialogue summarization given the corresponding dialogue with respect to informative with one to five stars, where one star means "uninformative" and five stars means "perfect informative". Note that informative measures the extent to which information is conveyed effectively and meaningfully. Consider the quality of individual sentences. Input: [a dialogue] Output: [one generated summary] Stars:
Relevance	Score the following dialogue summarization given the corresponding dialogue with respect to relevance with one to five stars, where one star means "irrelevance" and five stars means "perfect relevance". Note that relevance measures the degree to which something is applicable, pertinent, or connected to a particular context, topic, or situation. Consider the quality of individual sentences. Input: [a dialogue] Output: [one generated summary] Stars:

Table 12: Evaluation Instruction of **Fluency**, **Informative** and **Relevance** for dialogue summarization, where [task-ins] and [aspect-ins] are the instructions of the task-specific and aspect-specific, respectively. [aspect] and [ant-aspect] denote the evaluated aspect and its antonym, respectively.