Abstractive Summarization

李昌群

2022-9-30

目录

- Hallucinations(幻觉), Unfaithful
  - Common Problems Faced by Abstractive Summarization Models

- Prompt-based Domain Adaptation for Abstractive Summarization

Abstractive Summarization

- **Task**: Generating concise, fluent, salient and ***faithful*** to the source document summary;
- **Problem1**: intrinsic and extrinsic hallucinations (unfaithful)

Source: He was re-elected for a second term by the UN General Assembly, unopposed and unanimously, on **21 June 2011**, with effect from 1 January 2012. Mr. Ban describes his priorities as mobilising world leaders to deal with climate change, economic upheaval, pandemics and increasing pressures involving food, energy and water...
Unfaithful Summary: The United Nations Secretary-General Ban Ki-moon was elected for a second term in 2007.
Our Summary: The United Nations Secretary-General Ban Ki-moon was elected for a second term in **21 June 2011**.

这篇文章描述了前联合国秘书长潘基文连任的事件。该模型产生幻觉“2007”，它从未出现在源文档中，导致与所呈现事件的正确日期不一致。

Existing approaches

- Post-processing models
  - Training additional **correction or selection** models by using external resources
- Filtering nonfactual training data
  - Learning factuality directly during fine-tuning by **filtering nonfactual training data**
- FACTPEGASUS
  - Addressing the problem of factuality during pre-training and fine-tuning

Improving Faithfulness with **Contrast Candidate Generation and Selection**

方法/步骤:

1. **Contrast candidate generation**

   将摘要中的实体替换为源文档存在的实体，创建候选摘要的变体。

2. **Selection**

   使用训练的**discriminative model**对候选摘要进行排序，选择得分最高的作为最终的摘要。

Type	%	Ent. %	Num. %
Faithful	23.1	-	-
Ex. Hallucination	73.1	35.9	18.2
In. Hallucination	7.4	1.9	0.5

Table 2: Frequency of extrinsic and intrinsic hallucinations in 500 ground truth summary of the XSum corpus.

Observation

A large fraction of extrinsic hallucinations happen on **named entities and quantities**

实验

Xsum:

Changed Summary 13.3%
Non-existent hallucinated entity 38.4%
Keep the original summary 48.3

评测指标： Rouge, BERTScore

评测生成摘要的fluency, salience

Faithfulness Evaluation:

FEQA, a QA-based metric

Full XSum Test Set			
Method	ROUGE$_L$	BERT	FEQA (%)
BART$_{large}$	**36.95**	**91.57**	-
+ correct	36.70	91.50	-
Changed Summary Only (13.3%)			
BART$_{large}$	**38.63**	**91.61**	22.50
+ correct	36.62	91.10	**25.62**

Table 3: Evaluation with automatic metrics on the summaries generated by the baseline BART$_{large}$ model, plus our post-processing correction method. We report

Entity-level Factual Consistency

- Problem
  - **30%** of the summaries generated suffer from **fact fabrication**
  - **ROUGE** score is **inadequate** to quantify factual consistency

- Method
  - 1. New metrics: 量化生成摘要的实体级事实一致性
  - 2. Data filtering, multi-task learning and joint sequence generation

Method

New Metric

Precision-source ($\mathbf{prec}_s$): $N(h \setminus s) / N(h)$

表明在源文档中找到摘要中出现的命名实体的百分比。越低说明幻觉越严重

	Newsroom			CNNDM			XSUM		
	train	val	test	train	val	test	train	val	test
avg. $\mathcal{N}(t)$	2.08	2.10	2.09	4.36	5.09	4.87	2.08	2.06	2.08
avg. $\mathcal{N}(t \cap s)$	1.88	1.90	1.90	4.21	4.92	4.70	1.64	1.64	1.64
$\mathbf{prec}_s$ (%)	90.6	90.6	90.5	96.5	96.7	96.6	79.0	79.5	79.3

Table 1: Average number of named-entities and the $\mathbf{prec}_s$ scores (%) in the ground truth summary.

在Xsum数据集中指标分数较低，说明在Xsum数据中幻觉较严重

N (*t*) : number of entities in the gold summary

N (*h*) : number of entities in the generated summary

N (*h* \ *s*) : number of entities in gold and generated summary

Entity-based data filtering

1. NER识别实体
2. 不存在匹配项，丢弃

实验

	Newsroom			CNNDM			XSUM		
	train	val	test	train	val	test	train	val	test
original	922,500 (1.58)	100,968 (1.60)	100,933 (1.59)	287,112 (3.90)	13,368 (4.13)	11,490 (3.92)	203,540 (1.0)	11,301 (1.0)	11,299 (1.0)
after filtering	855,975 (1.62)	93,678 (1.64)	93,486 (1.64)	286,791 (3.77)	13,350 (3.99)	11,483 (3.77)	135,155 (1.0)	7,639 (1.0)	7,574 (1.0)

Table 2: Number of examples in three datasets together with the average number of sentences in the ground truth summary (in parentheses) before and after entity-based filtering.

过滤前后的数据统计

	training data	Rouge1	Rouge2	RougeL	macro $\mathbf{prec}_s$	micro $\mathbf{prec}_s$	macro $\mathbf{prec}_t$	micro $\mathbf{prec}_t$	macro $\mathbf{recall}_t$	micro $\mathbf{recall}_t$	macro $F1_t$	micro $F1_t$
Newsroom	original	47.7±0.2	35.0±0.3	44.1±0.2	97.2±0.1	97.0±0.1	65.4±0.3	62.9±0.4	70.8±0.3	68.5 ±0.2	68.0±0.2	65.6±0.3
	+ filtering	47.7±0.1	35.1±0.1	44.1 ±0.1	98.1±0.1	98.0±0.0	66.5±0.1	63.8±0.1	70.2 ±0.2	67.7±0.3	68.3±0.1	65.7±0.1
	+ classification	47.7±0.2	35.1±0.1	44.2±0.2	98.1±0.1	98.0±0.0	67.2±0.4	64.2±0.4	70.3±0.2	67.8±0.4	68.7±0.3	65.9±0.4
	JAENS	46.6 ±0.5	34.3±0.3	43.2±0.3	**98.3**±0.1	**98.3**±0.1	**69.5**±1.6	**67.3**±1.2	68.9±1.5	66.8±1.6	**69.2**±0.1	**67.0**±0.2
CNNDM	original	43.7±0.1	**21.1**±0.1	40.6±0.1	99.5±0.1	99.4±0.1	66.0±0.4	66.5±0.4	74.7±0.7	75.4±0.6	70.0±0.2	70.7±0.3
	+ filtering	43.4±0.2	20.8±0.1	40.3±0.2	**99.9**±0.0	**99.9**±0.0	66.2 ±0.4	66.6±0.3	74.1±0.6	74.9±0.6	69.9±0.2	70.5±0.2
	+ classification	43.5±0.2	20.8±0.2	40.4±0.2	**99.9**±0.0	**99.9**±0.0	**67.0**±0.6	**67.5**±0.5	74.7±0.2	75.5±0.1	70.6±0.3	71.3±0.3
	JAENS	42.4 ±0.6	20.2±0.2	39.5±0.5	**99.9**±0.0	**99.9**±0.0	**67.9**±0.7	**68.4**±0.6	**75.1**±0.7	**76.4**±0.7	**71.3**±0.2	**72.2**±0.2
XSUM	original	**45.6**±0.1	**22.5**±0.1	**37.2**±0.1	93.9±0.1	93.6±0.2	74.1±0.2	73.3±0.2	80.1±0.1	80.3±0.3	77.0±0.1	76.6±0.2
	+ filtering	45.4±0.1	22.2±0.1	36.9±0.1	98.2±0.0	98.2±0.1	77.9±0.2	77.3±0.2	79.4±0.2	79.6±0.2	78.6±0.1	78.4±0.2
	+ classification	45.3±0.1	22.1±0.0	36.9±0.1	98.3±0.1	98.2±0.1	78.6±0.3	**78.0**±0.3	79.5±0.3	79.8±0.4	**79.1**±0.1	**78.9**±0.1
	JAENS	43.4±0.7	21.0±0.3	35.5 ±0.4	**99.0**±0.1	**99.0**±0.1	77.6±0.9	77.1±0.6	79.5±0.6	80.0±0.5	78.5±0.2	78.5±0.1

Main result

Factuality-Aware Pre-training and Fine-tuning

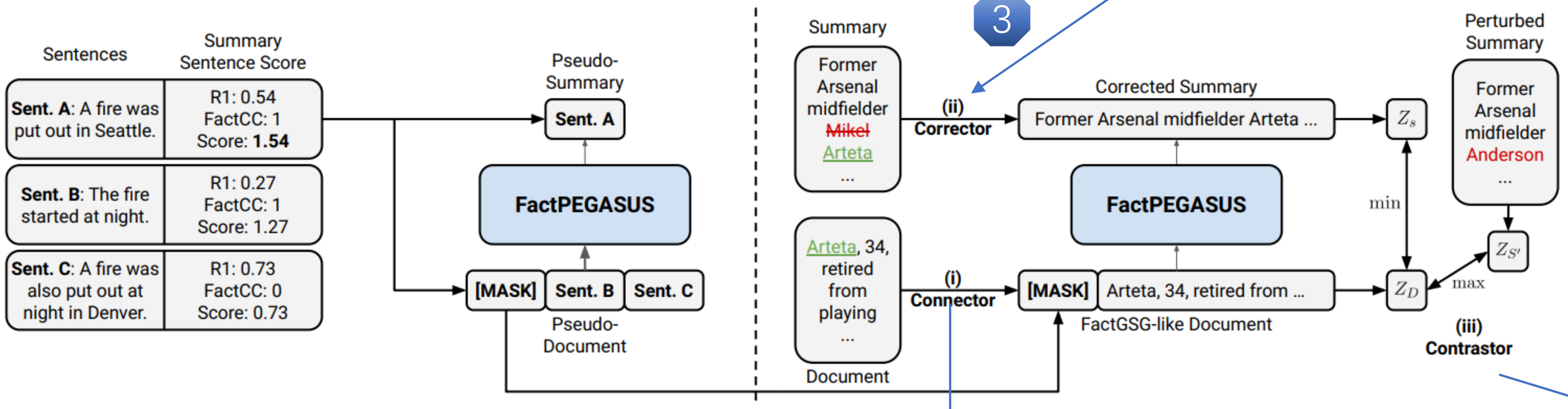
- Pre-training stage
  - Incorporating **factuality into the pre-training objective** of PEGASUS

- Fine-tuning stage
  - **Corrector:** removes hallucinations existing in reference summaries;
  - **Contrastor:** differentiate factual summaries from nonfactual contrastive learning;
  - **Connector:** bridges the gap between the pre-training and finetuning for better transfer of knowledge.

模型架构

Approaches: Replace, Remove and Combined

3



(a) Pre-training

(b) Fine-tuning

4

在文档和摘要之间进行对比学习

1

方法： 结合ROUGE和事实度量FactCC作为选择标准。

目的： 使模型学会生成涵盖输入文档中最重要信息的句子，并对其保持事实一致性

2

将mask token插入到数据集的输入中，从而模拟模型在预训练时的模式，插入的position在验证集进行确定

实验

Factuality Evaluation:

1. FactCC;
2. DEP-Entail: token error and sentence error

Dataset	Model	RL	tok err↓	sent err↓	FactCC
XS	BART-base	**33.78**	12.38	60.70	23.99
	PEGASUS*	33.17	12.33	60.01	24.14
	DAE	31.78	4.79*	35.52*	25.43
	CLIFF	31.40	10.36	53.14	23.77
	FACTPEGASUS	31.17	**6.07**	**38.66**	**34.32**
WH	BART-base	31.81	8.99	45.77	99.09
	PEGASUS*	30.30	9.77	47.28	98.83
	DAE	31.66	4.91*	34.45*	98.87
	CLIFF	**33.82**	13.74	57.42	99.18
	FACTPEGASUS	29.33	**7.86**	**42.40**	**99.41**
GW	BART-base	**35.11**	2.29	19.68	55.66
	PEGASUS*	34.74	2.84	22.66	56.43
	DAE	35.57	0.58*	7.54*	59.61
	CLIFF	34.89	**1.72**	**18.45**	58.53
	FACTPEGASUS	34.23	2.30	19.32	**60.02**

Fine-tuning results

Model	RL	tok err↓	sent err↓	FactCC
factGSG	**32.99**	12.31	59.30	24.94
+ corrector replace	32.48	10.57	55.05	25.06
+ corrector remove	30.37	6.44	39.89	**35.77**
+ corrector combined	31.19	6.10	38.96	33.79
+ contrastor intrinsic	32.14	11.46	57.61	25.26
+ contrastor extrinsic	32.54	11.95	59.10	25.07
+ contrastor + corrector	31.17	6.08	38.92	34.17
FACTPEGASUS	31.17	**6.07**	**38.66**	34.32

Fine-tuning ablation on XSum

Prompt-based Domain Adaptation

- 问题:
  - 在特定的领域，可利用的标注数据较少
  - In-domain数据和out-of-domain数据结合将导致域外数据过拟合
  - 当对话摘要模型应用到新领域时泛化能力较差

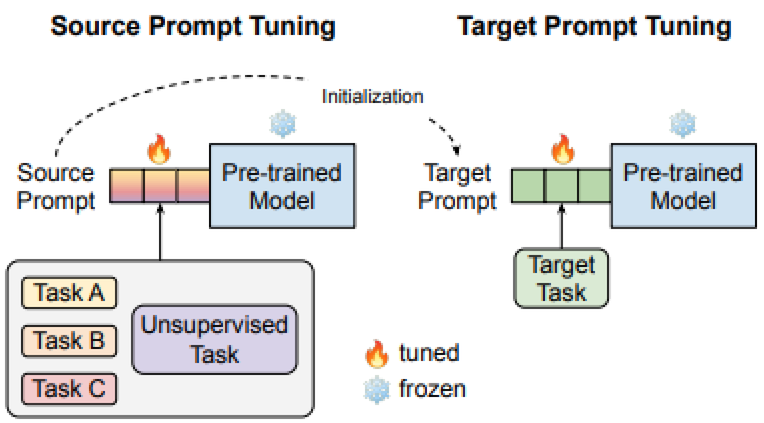
- Domain Adaptation研究的是如何利用通用域的大量的标注数据，来提升目标域的性能。

Domain Adaptation

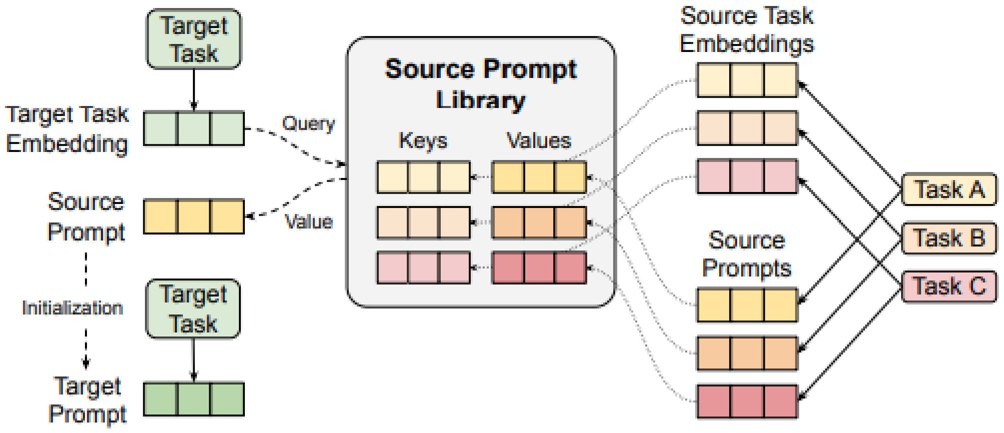
- The key point
  - how to effectively transfer learned knowledge from source domain

- Two aspects
  - **Domain-Invariant** Information (shared knowledge)
  - **Domain-Specific** Information (domain-related features)

Soft Prompt Transfer for Model Adaptation



Soft Prompt Transfer



retrieval approach

任务之间的相似度是一个重要的影响因素

计算流程:

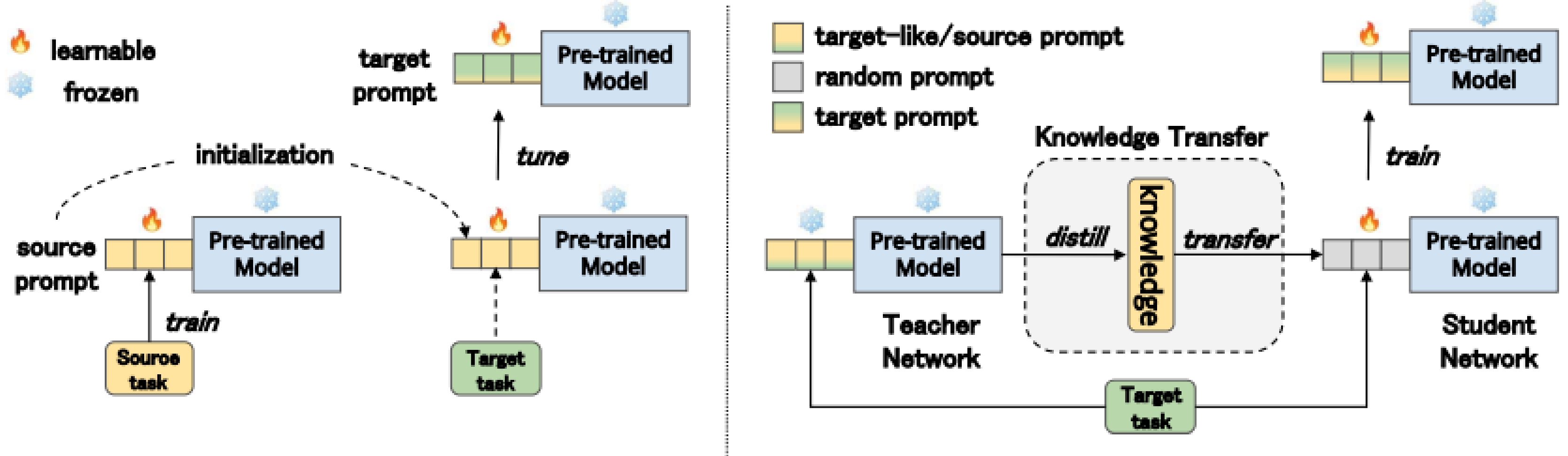
(i)计算一个任务embeddings,
(ii)检索一个最优source prompt,
(iii)将检索到的source prompt用来初始化target prompt。

SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer ACL 2022

Prompt Transfer Meets Knowledge Distillation for Efficient Model Adaptation

问题：

1. be **sensitive** to the similarity between source and target tasks
2. **catastrophic forgetting**



(a) Vanilla Prompt Transfer (b) Our PANDA approach

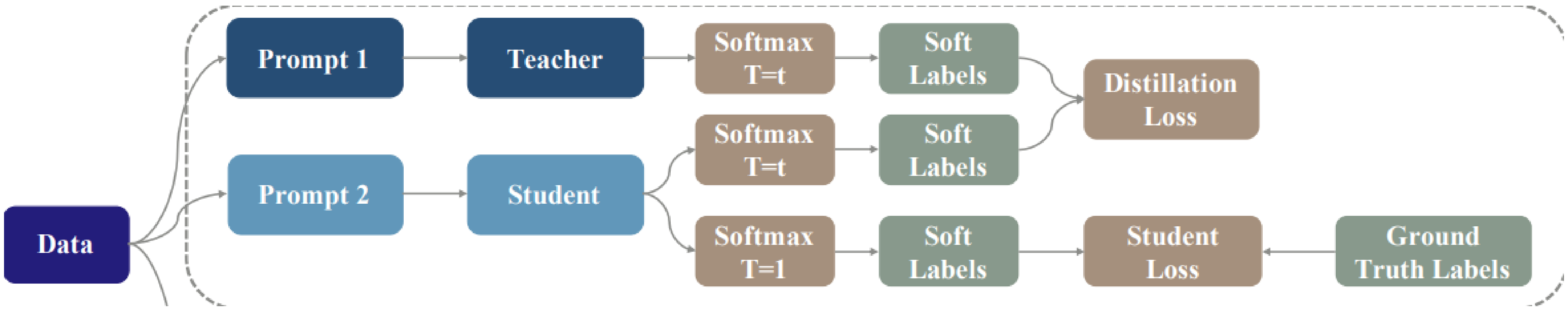
模型架构图

PANDA: Prompt Transfer Meets Knowledge Distillation for Efficient Model Adaptation Under review

损失

$$\mathcal{L}_{all}(u_r, f) = \mathcal{L}_{ce}(u_r, f) \quad (1)$$

The classification loss



Prompt-based Knowledge Distillation

$$\mathcal{L}_{\text{all}}(u_r, f) = \mathcal{L}_{ce}(u_r, f) + \lambda \cdot \mathcal{L}_{kd}(u_r, f) \quad (2)$$

The classification loss + KD loss

KD: Knowledge Distillation

损失

$$\mathcal{L}_{\text{all}}(u_r, f) = \mathcal{L}_{ce}(u_r, f) + \lambda \cdot \text{sim}\left(\hat{h}_s, \hat{h}_t\right) \cdot \mathcal{L}_{kd}(u_r, f) \quad (3)$$

Metric: to measure the prompt similarity

$$\mathcal{D} \xrightarrow[\texttt{[CLS]}]{\mathcal{M}} h_{cls}^{m}, \; e(\mathcal{D}) \xrightarrow[\texttt{[CLS]}]{\overline{\mathcal{M}}, u} h_{cls}^{p}; \quad \hat{h} = h_{cls}^{p} - h_{cls}^{m}; \quad sim(\hat{h}_s, \hat{h}_t) = \left(\frac{\hat{h}_s \cdot \hat{h}_t}{\|\hat{h}_s\| \|\hat{h}_t\|}\right).$$

计算流程:

1. 随机选取部分样本作为代表数据;
2. 输入原始PLM 和经过训练的prompt，得到隐向量;
3. 将两个向量相减作为基于prompt的任务嵌入;
4. 相似性比较

实验结果

Method	CB	COPA	WSC	RTE	WIC	CoLA	MRPC	STSB	Conll_{04}	AVG.
model-tuning	94.6	69.0	68.3	75.8	74.9	60.6	88.0	90.0	85.6	78.5
prompt-tuning	87.5	76.0	64.4	76.2	66.9	63.8	86.8	90.5	85.5	77.5
(a) Transfer with Vanilla Prompt Transfer approach										
MNLI	**96.4**	71.0	**67.3**	**80.9**	66.5	58.9	**88.2**	**91.0**	83.0	**78.1**
QNLI	**89.3**	76.0	**65.4**	76.2	**70.4**	63.7	**88.5**	**90.7**	83.5	**78.2**
Record	78.6	63.0	**65.4**	53.8	51.7	0.0	77.7	85.0	82.7	62.0
SQuAD	87.5	74.0	**66.3**	71.8	51.7	6.0	**87.3**	89.3	82.5	68.5
CoNLL03	73.2	64.0	63.5	60.3	51.9	0.0	71.3	16.4	84.8	53.9
Ontonotes	78.6	65.0	**66.3**	56.7	54.1	59.3	82.4	84.5	**86.1**	70.3
CoNLL05	87.5	65.0	64.4	69.3	**68.3**	61.3	**88.7**	88.4	83.8	75.2
CoNLL12	**89.3**	62.0	**67.3**	63.2	**67.4**	58.7	90.4	88.5	83.6	74.5
SST2	**92.9**	74.0	64.4	71.8	66.8	60.1	**87.0**	89.6	84.3	76.8
(b) Transfer with Our PANDA approach										
MNLI	**92.9**	**77.0**	**67.3**	**78.0**	**68.8**	**66.3**	**88.5**	**90.6**	85.4	$\mathbf{79.4}_{1.3}$
QNLI	**92.9**	**77.0**	**66.3**	**77.3**	**70.8**	63.9	**87.5**	**90.8**	**86.6**	$\mathbf{79.2}_{1.0}$
Record	87.5	76.0	**66.3**	**77.3**	**68.5**	62.4	**87.5**	**90.7**	84.9	$\mathbf{77.9}_{15.9}$
SQuAD	**89.3**	75.0	**66.3**	75.5	**69.3**	63.1	**87.3**	88.9	**85.7**	$\mathbf{77.8}_{9.3}$
CoNLL03	**91.1**	72.0	**68.3**	**76.9**	67.4	63.6	86.5	**90.6**	**85.6**	$\mathbf{78.0}_{24.1}$
Ontonotes	**89.3**	74.0	**66.3**	76.2	**69.1**	**64.2**	**88.0**	**90.8**	**85.7**	$\mathbf{78.2}_{7.8}$
CoNLL05	87.5	**79.0**	**65.4**	**77.6**	**69.6**	**63.7**	**87.5**	**90.8**	84.8	$\mathbf{78.4}_{3.2}$
CoNLL12	87.5	76.0	**66.3**	74.4	**68.5**	**63.7**	**87.5**	**90.8**	85.0	$\mathbf{77.7}_{3.3}$
SST2	**92.9**	**77.0**	**68.3**	**76.5**	**70.1**	**64.8**	**88.5**	**90.7**	**86.3**	$\mathbf{79.5}_{2.7}$

Main Results

实验结论:

1. Prompt-tuning via PANDA approach consistently outperforms model-tuning;

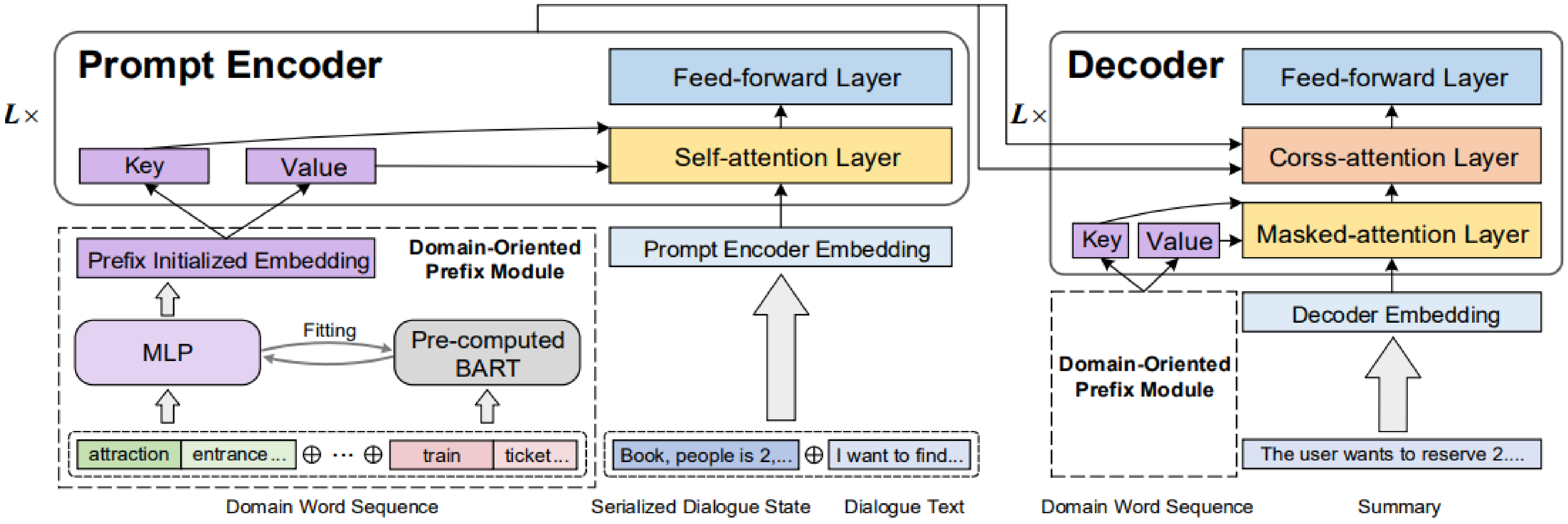
2. Knowledge Distillation helps bridge the gap between different types of tasks

Method	BERT-medium	BERT-tiny
prompt-tuning	70.5	59.1
vanilla PoT	69.16	57.09
PANDA		
-w constant (ones)	71.21	60.18
-w $Eavg$ metric	71.08	60.10
-w ON metric	71.06	60.03
-w Our metric	**71.70**	**60.36**

Scores with different metrics

Domain-Oriented Prefix-Tuning

- 模型架构



1. Utilizing a domain word initialized prefix module

2. Adopting discrete prompts to guide the model

Domain-Oriented Prefix-Tuning: Towards Efficient and Generalizable Fine-tuning for Zero-Shot Dialogue Summarization NAACl 2022

实验

对话数据集

Domains	Size	Dialog.len	Summ.len	DS.len
Train	345	120.67	24.93	18.29
Taxi	435	80.24	29.04	15.80
Restaurant	1,311	105.42	23.04	14.30
Hotel	636	145.16	30.06	21.38
Attraction	150	95.48	22.27	7.92
All	2,877	111.71	25.68	16.24

TODSum

Domains	Size	Dialog.len	Summ.len	QR.len
Academic	312	1,155.78	46.48	8.56
Committee	417	757.68	76.00	14.54
Product	847	971.65	63.96	13.36
All	1,576	951.49	63.68	12.73

QMSum

Models	Train			Taxi			Restaurant			Hotel			Attraction		
	2,332 / 200 / 345			2,242 / 200 / 435			1,366 / 200 / 1,311			2,041 / 200 / 636			2,527 / 200 / 150		
	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
Lead-3	20.36	2.78	16.07	24.20	7.34	20.75	28.27	6.10	23.49	23.86	4.58	18.80	22.76	5.28	19.66
Oracle	39.06	10.04	32.87	38.96	14.06	33.43	45.79	15.57	38.42	39.65	11.28	32.56	41.90	14.18	38.79
BertExt	39.19	9.71	33.24	38.49	13.57	33.36	40.64	12.34	34.43	35.96	9.71	30.10	36.25	11.19	31.41
PGN	32.50	10.47	29.33	32.48	7.79	29.82	33.63	10.78	31.47	32.18	9.36	30.93	32.66	9.95	30.29
Transformer	33.47	10.98	30.28	33.35	8.71	30.57	34.49	11.43	31.99	33.05	10.62	31.63	33.18	10.74	30.91
BertAbs	42.89	16.57	37.32	36.43	14.69	32.15	42.10	18.61	38.87	38.03	13.34	33.22	36.21	14.81	34.67
BART	46.82	18.42	42.06	39.98	15.79	34.41	47.02	22.62	44.93	40.84	14.20	36.83	43.67	20.23	41.44
BART w. DS	49.02	23.80	44.59	43.59	19.56	38.65	49.25	23.57	45.23	43.97	17.02	39.31	47.55	22.62	45.16
Prefix-tuning	45.92	22.70	41.06	41.89	19.47	39.62	47.19	24.20	42.99	43.41	18.75	36.75	44.48	22.43	40.94
DOP (ours)	**52.51**	**25.45**	**47.78**	**47.14**	**24.37**	**42.75**	**51.28**	**32.68**	**47.44**	**48.44**	**24.58**	**41.45**	**52.90**	**30.51**	**49.48**

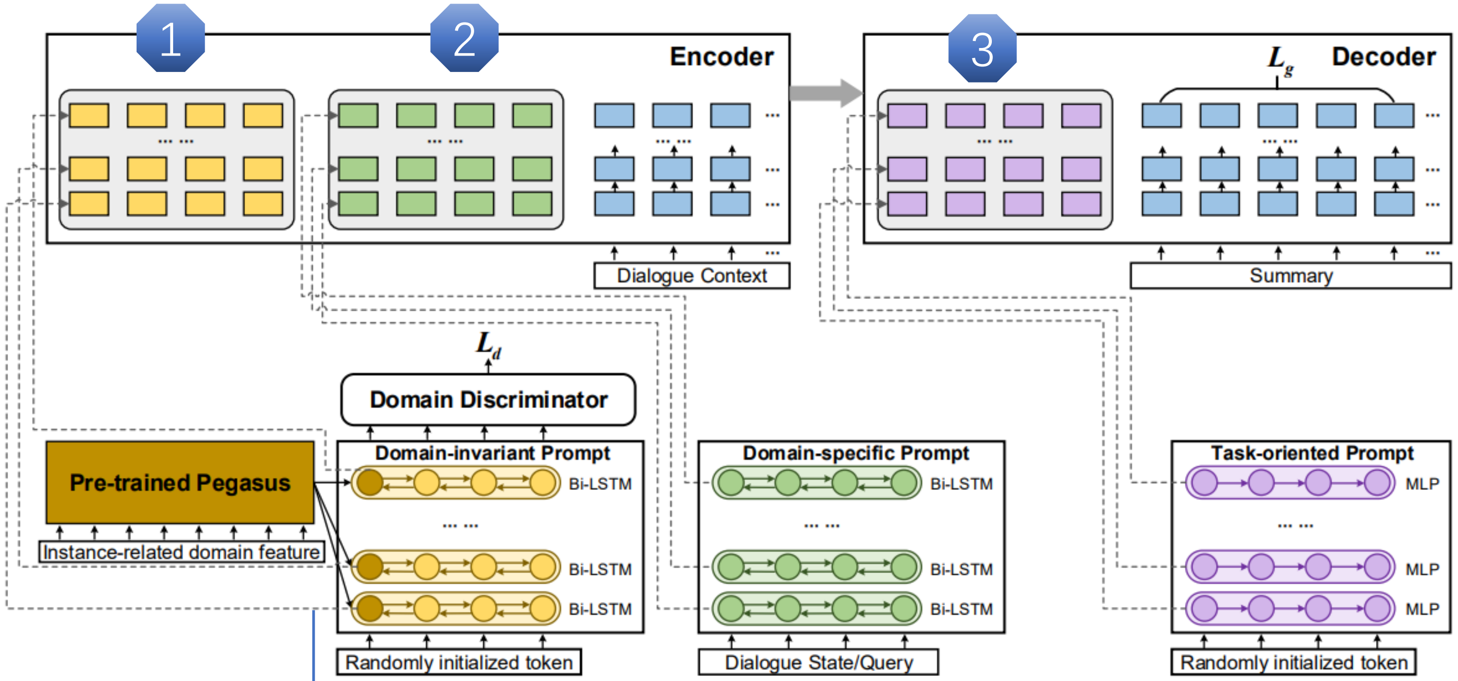
Zero-Shot Experiments Results

Model	R-1	R-2	R-L
DOP (ours)	52.51	25.45	47.78
w/o DW	48.87	23.81	44.52
w/o DS	47.59	23.25	43.41
w/o DW & DS	45.92	22.70	41.06

Ablation study

Table 5: F1 scores of ablation study on *train* domain of TODSum dataset. "DW" denotes domain words and "DS" denotes dialogue states.

Adversarial Prompt-based Domain Adaptation



模型架构

总结：
设计的三个prompt取代了随机初始化，编码了一些特定的信息，从而引出预训练模型中相关的知识。

Three kinds of prompts:

1. Domain-invariant prompt
   Shared knowledge

2. Domain-specific prompt
   Domain-related features

3. Task-oriented prompt

ADPL: Adversarial Prompt-based Domain Adaptation for Dialogue Summarization with Knowledge Disentanglement SIGIR 2022

实验

Models	Train			Taxi			Restaurant			Hotel			Attraction		
	2,332 / 200 / 345			2,242 / 200 / 435			1,366 / 200 / 1,311			2,041 / 200 / 636			2,527 / 200 / 150		
	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
Lead-3	20.36	2.78	16.07	24.20	7.34	20.75	28.27	6.10	23.49	23.86	4.58	18.80	22.76	5.28	19.66
Oracle	39.06	10.04	32.87	38.96	14.06	33.43	45.79	15.57	38.42	39.65	11.28	32.56	41.90	14.18	38.79
BertExt	39.19	9.71	33.24	38.49	13.57	33.36	40.64	12.34	34.43	35.96	9.71	30.10	36.25	11.19	31.41
PGN	32.50	10.47	29.33	32.48	7.79	29.82	33.63	10.78	31.47	32.18	9.36	30.93	32.66	9.95	30.29
Transformer	33.47	10.98	30.28	33.35	8.71	30.57	34.49	11.43	31.99	33.05	10.62	31.63	33.18	10.74	30.91
BertAbs	42.89	16.57	37.32	36.43	14.69	32.15	42.10	18.61	38.87	38.03	13.34	33.22	36.21	14.81	34.67
BART	46.82	18.42	42.06	39.98	15.79	34.41	47.02	22.62	44.93	40.84	14.20	36.83	43.67	20.23	41.44
M-BART	48.62	22.92	43.90	40.37	17.48	36.03	49.23	26.37	45.00	42.47	18.07	38.23	53.65	31.40	50.46
BART w. DS	49.02	23.80	44.59	43.59	19.56	38.65	49.25	23.57	45.23	43.97	17.02	39.31	47.55	22.62	45.16
Prefix-tuning (BART)	45.92	22.70	41.06	41.89	19.47	39.62	47.19	24.20	42.99	43.41	18.75	36.75	44.48	22.43	40.94
Pegasus	52.14	27.19	47.67	48.99	21.94	43.34	54.81	26.00	50.18	48.31	21.11	42.17	53.90	28.12	50.96
Prefix-tuning (Pegasus)	49.77	24.40	45.56	44.62	21.65	40.71	54.93	32.43	50.56	49.11	23.35	41.75	51.94	27.47	47.63
ADPL (ours)	**55.18**	28.03	**52.36**	49.87	**23.86**	**45.62**	**60.01**	**35.97**	**56.37**	**53.45**	**26.78**	**45.16**	**57.28**	**31.69**	51.49

Table 3: Results in terms of ROUGE-1, ROUGE-2, and ROUGE-L on TODSum in the zero-shot setting. All ROUGE scores are reported by averaging three random runs. Here, "DS" denotes the dialogue states. Values in the second row denote the size of train/valid/test set. ($p < 0.01$ under t-test)

Zero-Shot Experiments Results

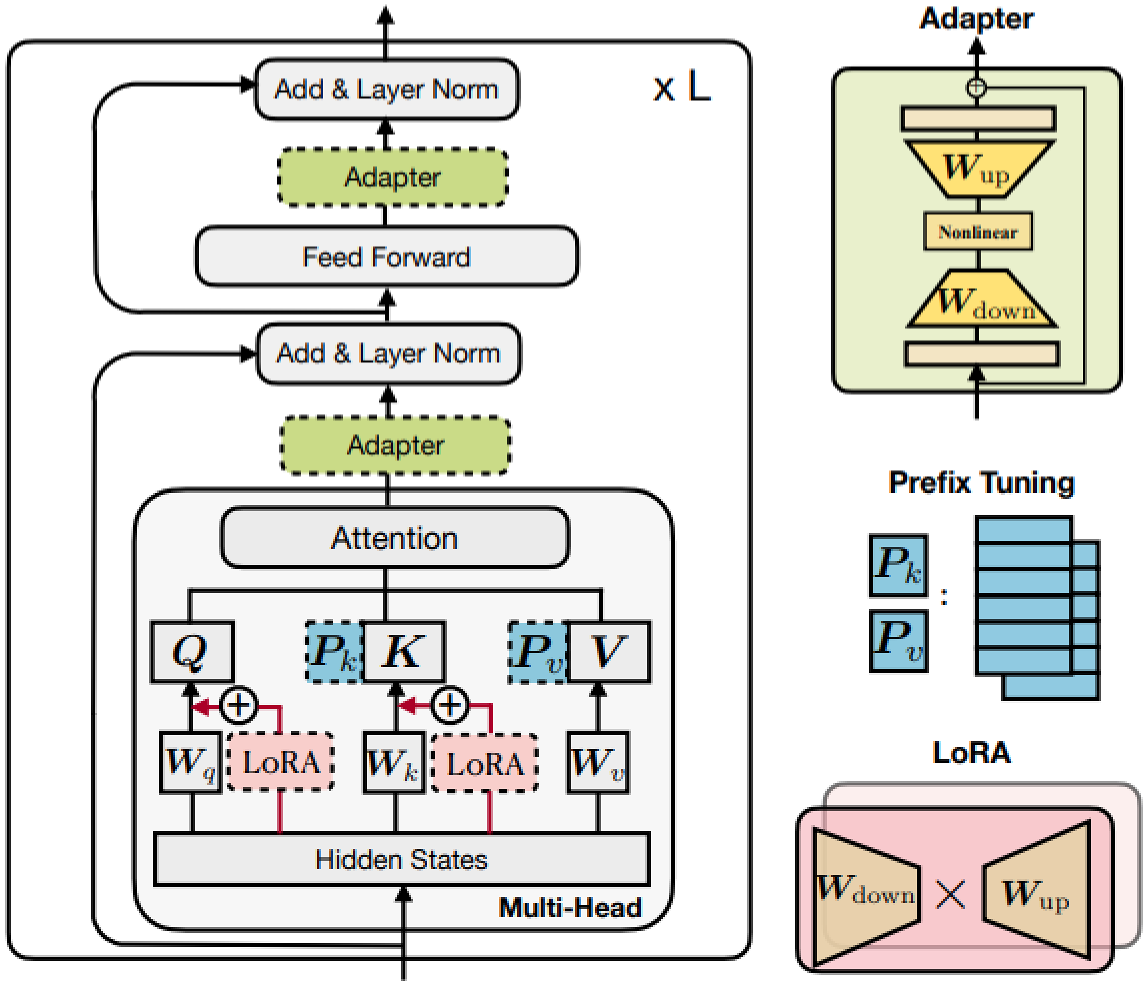
Model	R-1	R-2	R-L
ADPL (ours)	55.18	28.03	52.36
w/o AL	52.37	26.13	48.43
w/o DW	51.63	25.00	46.87
w/o DIP	47.73	21.94	42.12
w/o DSP	49.81	23.80	44.71
w/o DIP & DSP	47.53	21.91	42.32
w/o TOP	50.09	23.65	45.07

Table 5: F1 scores of ablation study on *train* domain of TODSum dataset. "AL" denotes adversarial learning and "DW" denotes domain words. "w/o DIP & DSP" means the removal of encoder prompt.

Ablation study

Thanks

Parameter-efficient Transfer Learning



Existing Methods

Adapter:

$$\mathrm{Adapter}(\mathbf{x}) = \mathbf{W}_u(\mathrm{ReLU}(\mathbf{W}_d\mathbf{x} + \mathbf{b}_d)) + \mathbf{b}_u$$

Prefix-tuning:

$$K_l' = [P_{l,K}; K_l]\ , V_l' = [P_{l,V}; V_l]$$

LoRA:

$$\boldsymbol{h} \leftarrow \boldsymbol{h} + s \cdot \boldsymbol{x}\boldsymbol{W}_{\mathrm{down}}\boldsymbol{W}_{\mathrm{up}},$$

- 这些方法里面关键部分是什么？这些方法之间是否有什么联系？

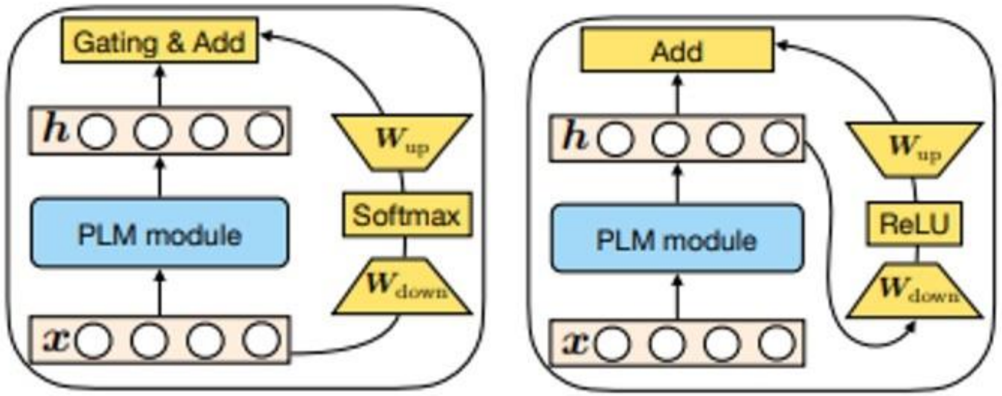
Prefix Tuning:

$$h \leftarrow (1-\lambda(x))h + \lambda(x) f(x W_{\text{down}}) W_{\text{up}}$$

Adapters:

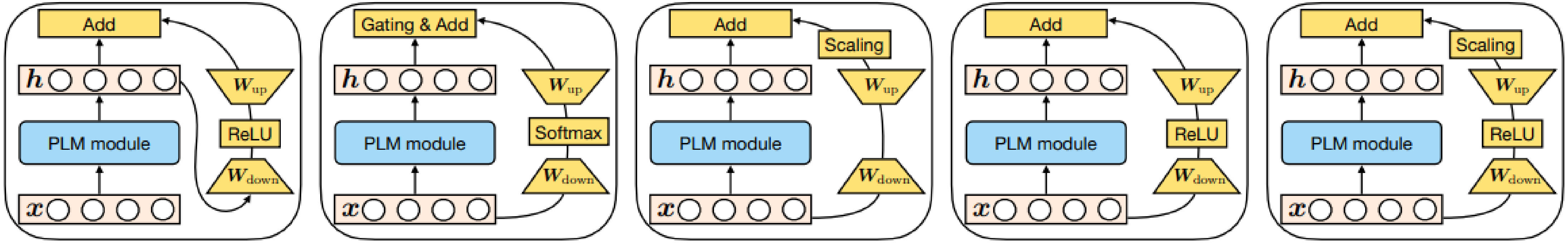
$$h \leftarrow h + f(h W_{\text{down}}) W_{\text{up}}$$

Prefix-tuning是另一种形式的adapter



(b) Prefix Tuning (a) Adapter

计算流程图



(a) Adapter (b) Prefix Tuning (c) LoRA (d) Parallel Adapter (e) Scaled PA

现有方法和提出的变体的图形说明

四个维度:

1. Functional Form
2. Insertion Form
3. Modified Representation
4. Composition Function

实验结论

- Insertion Form: Parallel > Sequential
- Modified Representation: FFN > attn (generally), but multi-head attn is superior with very small parameter budget (0.1% of original parameters)
- Composition: $\boldsymbol{h} \leftarrow \boldsymbol{h} + s \cdot \Delta \boldsymbol{h}$

  (Scaled addition is a good tradeoff between performance and simplicity)

Towards Low-Resource Domain Adaptation for Abstractive Summarization

- A second phase of pre-training under three settings
  - **source domain pre-training** (SDPT) based on a labeled source domain summarization dataset;
  - **domain-adaptive pre-training** (DAPT) based on an unlabeled substantial domain-related corpus;
  - **task-adaptive pre-training** (TAPT) based on an unlabeled smallscale task-related corpus.

实验

Models	Dialog	Email	Movie R.	Debate	Social M.	Science	Average
BART Fine-tuning	39.95	24.71	25.13	24.48	21.76	72.76	34.80
SDPT	42.84	25.16	25.45	25.61	22.43	**73.09**	35.76
w/ RecAdam	**45.23**	**26.97**	**26.06**	25.17	**23.25**	72.60	**36.55**
DAPT	41.22	26.50	24.25	**26.71**	22.95	71.88	35.59
w/ RecAdam	40.05	25.66	25.78	25.01	21.51	72.23	35.04
TAPT	40.15	25.30	25.27	24.59	22.81	73.08	35.20
w/ RecAdam	41.34	25.73	25.65	24.70	23.01	72.80	35.54

Table 2: ROUGE-1 scores on different pre-training methods compared to the baseline BART over all domains.

Domain	Unlabeled Corpus		Labeled data		
	# Tokens	Size	Train	Valid	Test
Dialog	44.96M	212MB	300	818	819
Email	117.54M	705MB	300	1960	1906
Movie R.	11.36M	62MB	300	500	2931
Debate	122.99M	693MB	300	956	1003
Social M.	153.30M	786MB	300	1000	1000
Science	41.73M	291MB	100	350	497

实验发现:

1. 预训练的有效性与预训练数据与目标域任务的相似度相关。
2. 继续进行预训练可能会导致预训练模型的灾难性遗忘

Domain-Agnostic Multi-Source Pretraining

Three procedures:

1. the pretraining of **encoder**

2. the pretraining of **decoder**

3. the pretraining of the **combined encoder-decoder**

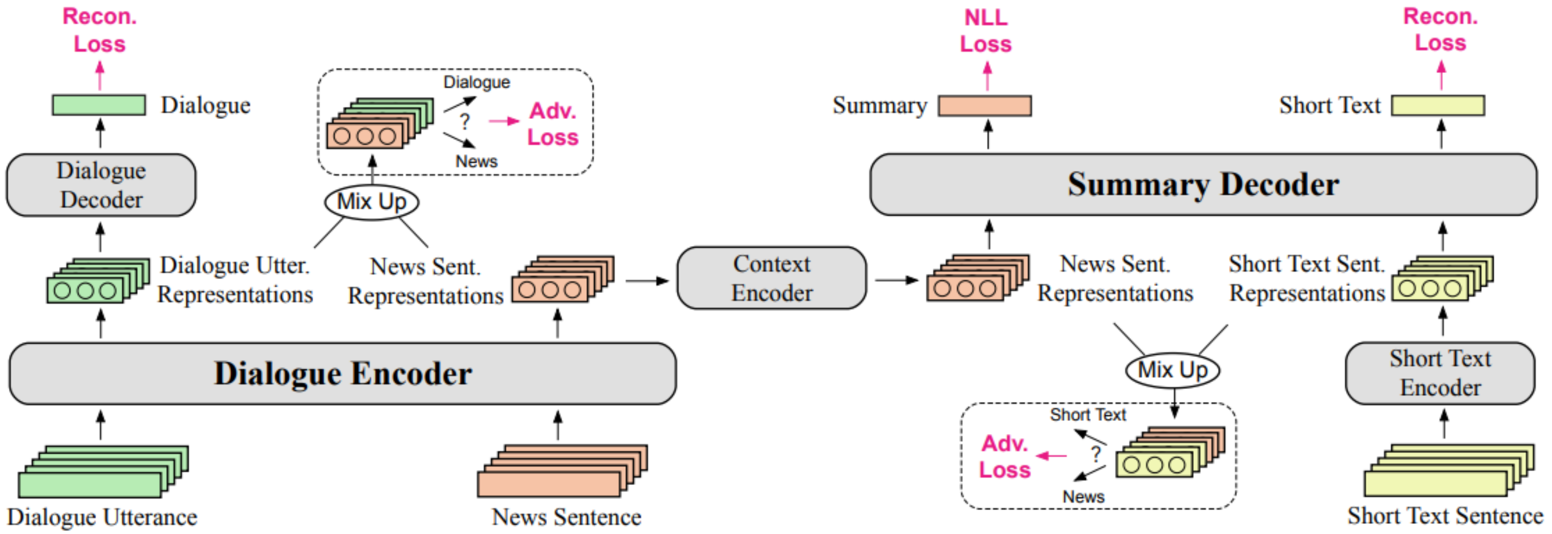


Figure 1: The overall architecture of DAMS. The multi-source pretraining includes: (i) encoder pretraining using dialogues (green); (ii) decoder pretraining using short texts (yellow); (iii) Joint pretraining using general articles with corresponding summaries (orange).

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{gen} + \mathcal{L}_{summ} + \alpha(\mathcal{L}_e^D + \mathcal{L}_g^D). \quad (8)$$

实验

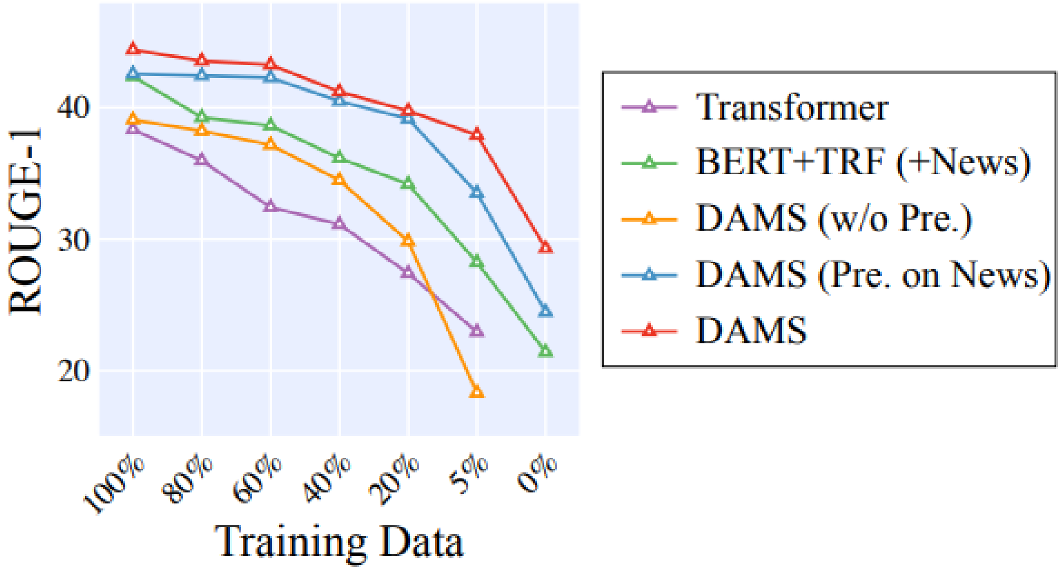


Figure 2: Model performance in low-resource settings.

Few-shot setting

Model	**+News**	**RG-1**	**RG-2**	**RG-L**
Longest-3	-	32.46	10.27	29.92
Seq2Seq+Att	-	29.35	15.90	28.16
Transformer	-	37.27	18.44	32.73
PGNet	-	40.08	15.28	36.63
FastRL	-	40.96	17.18	39.05
FastRL Enhanced	-	41.95	18.06	39.23
D-HGN	-	42.03	18.07	39.56
TGDGA	-	43.11	19.15	40.49
BERT+TRF	-	39.90	17.01	39.12
LightConv	✓	40.29	17.28	36.81
DynamicConv	✓	41.07	17.11	37.27
Transformer	✓	42.37	18.44	39.27
PGNet	✓	37.27	14.42	34.36
FastRL	✓	41.03	16.93	39.05
FastRL Enhanced	✓	41.87	17.47	39.53
BERT+TRF	✓	42.37	17.59	40.73
DAMS (w/o pretrain)	-	39.07	14.59	38.06
DAMS	✓	**44.38**	**19.98**	**43.40**

Table 2: Results of ROUGE-1/2/L on the SAMSum corpus. **+News** means whether the approach exploits external news summary data or not.