2024年度暑期强化课程

Prompt Paradigm：Pre-train, Prompt, and Predict

授课人：曹亚男

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

Prompt Paradigm

NLP领域发展的四个阶段

- **非神经网络时代的完全监督学习 (Fully Supervised Learning, Non-Neural Network)**
  - 特征工程（Feature Engineering）：人工设计和定义特征模板
- **基于神经网络的完全监督学习 (Fully Supervised Learning, Neural Network)**
  - 结构工程（Architecture Engineering）：神经网络解放手动配置特征模板所需要的人力，但需要探究最适配下游任务的结构偏置
- **预训练，微调范式 (Pre-train, Fine-tune)**
  - 目标函数挖掘（Objective Engineering）：引入额外的目标函数到预训练语言模型上，以便让其更适配下游任务，网络结构的挖掘非此阶段的主旋律
- **预训练，提示，预测范式 (Pre-train, Prompt, Predict)**
  - Prompt挖掘工程：将微调改为“Prompt+预测”，通过设计合适的prompt实现对下游任务建模方式的重新定义

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

范式、任务和模型

- **任务：取决于输入和输出** $D=\{x_i,y_i|1\leq i\leq N\}$
- **范式：具有一种特定形式的任务的模型框架**
- **范式和任务之间的关系**
  - ✓ **一个任务可以通过将其转换成不同的形式，用多个范式来解决**
  - ✓ **一个范式也可以通过将不同任务转换为一种形式，来解决多种任务**

Sun et al.. *Paradigm Shift in Natural Language Processing*. Arxiv. 2021.9.26 (复旦 邱锡鹏).

NLP中的七种基础范式

- 分类
- 匹配
- 序列标注
- 阅读理解
- Seq2Seq生成
- Seq2ASeq序列决策
- (M)LM

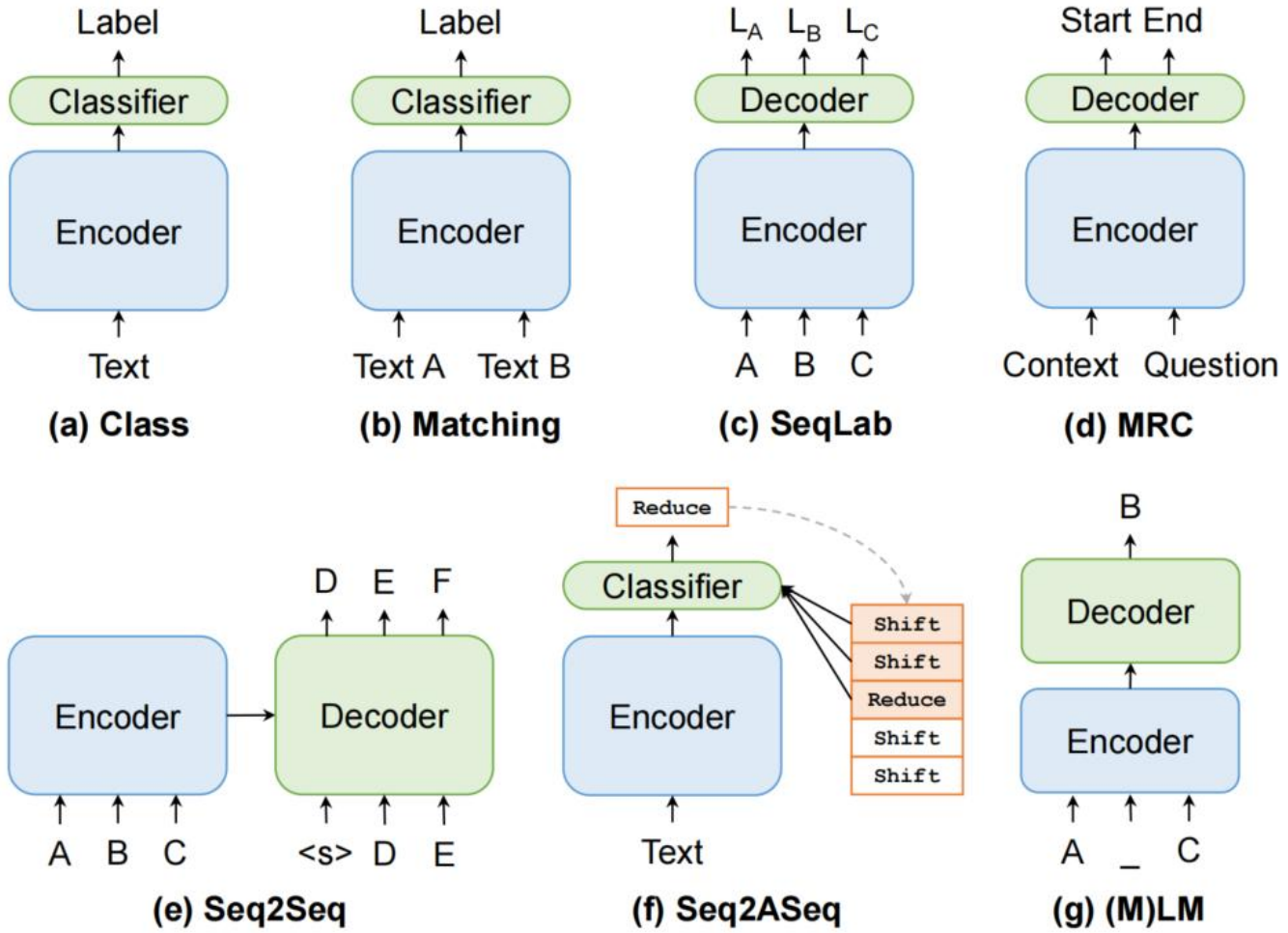
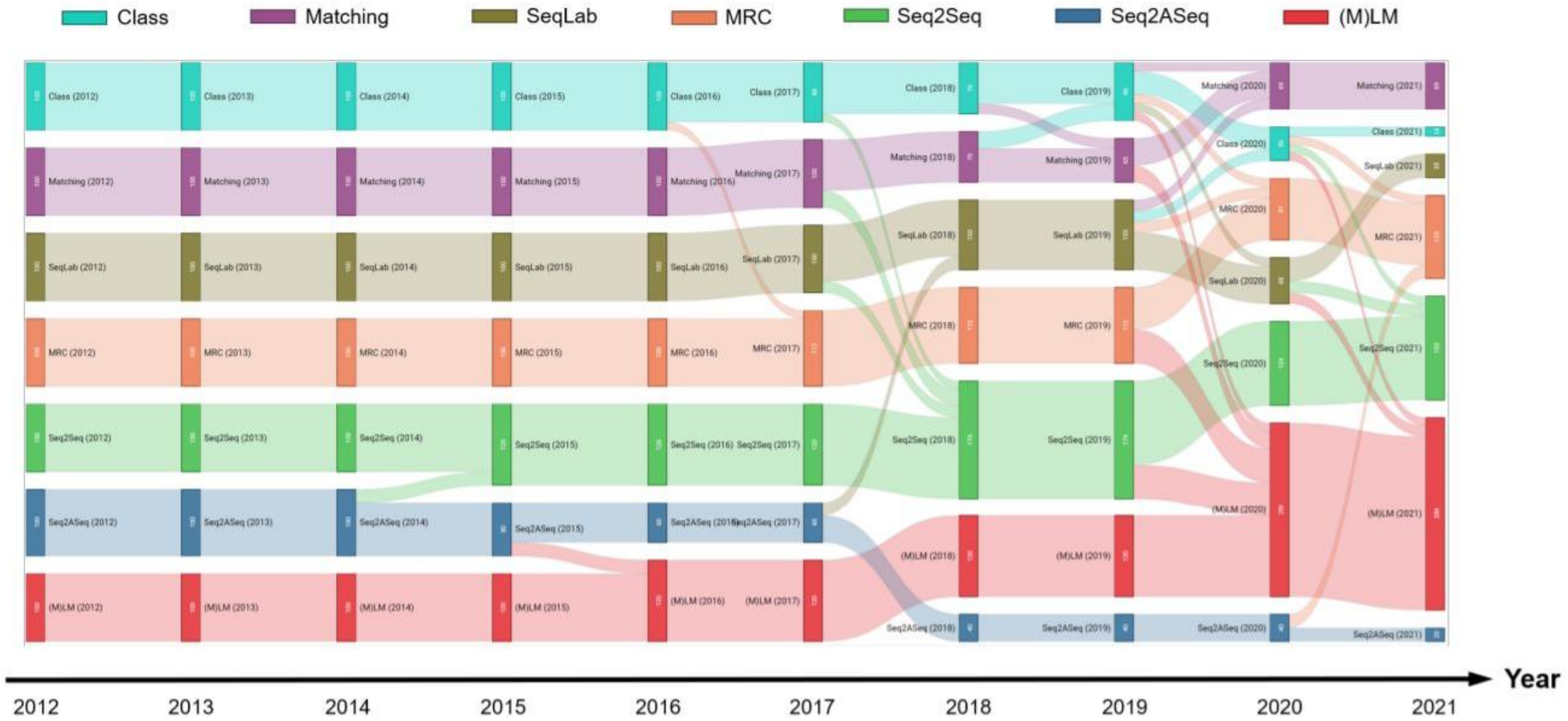


Figure 1: Illustration of the seven mainstream paradigms in NLP.

Sun et al.. *Paradigm Shift in Natural Language Processing*. Arxiv. 2021.9.26 (复旦 邱锡鹏).

NLP Paradigm Shift

- 一个任务可采用多个范式，从Class、SeqLab转向Matching、MRC、Seq2Seq、(M)LM



NLP Paradigm Shift

- 一个任务可采用多个范式，从Class、SeqLab转向Matching、MRC、Seq2Seq、(M)LM

Task		Class	Matching	SeqLab	MRC	Seq2Seq	Seq2ASeq	(M) LM
TC	Input	$\mathcal{X}$	$\mathcal{X}, \mathcal{L}$			$\mathcal{X}$		$f_{prompt}(\mathcal{X})$
	Output	$\mathcal{Y}$	$\mathcal{Y} \in \{0, 1\}$			$y_1, \cdots, y_m$		$g(\mathcal{Y})$
	Example	Devlin et al. (2019)	Chai et al. (2020)			Yang et al. (2018a)		Schick and Schütze (2021a)
NLI	Input	$\mathcal{X}_a \oplus \mathcal{X}_b$	$\mathcal{X}_a, \mathcal{X}_b$			$f_{prompt}(\mathcal{X}_a, \mathcal{X}_b)$		$f_{prompt}(\mathcal{X}_a, \mathcal{X}_b)$
	Output	$\mathcal{Y}$	$\mathcal{Y}$			$\mathcal{Y}$		$g(\mathcal{Y})$
	Example	Devlin et al. (2019)	Chen et al. (2017b)			McCann et al. (2018)		Schick and Schütze (2021a)
NER	Input	$\mathcal{X}_{span}$		$x_1, \cdots, x_n$	$\mathcal{X}, \mathcal{Q}_y$	$\mathcal{X}$	$(\mathcal{X}, \mathcal{C}_t)_{t=0}^{m-1}$	
	Output	$\mathcal{Y}$		$y_1, \cdots, y_n$	$\mathcal{X}_{span}$	$(\mathcal{X}_{ent_i}, \mathcal{Y}_{ent_i})_{i=1}^{m}$	$\mathcal{A} = a_1, \cdots, a_m$	
	Example	Fu et al. (2021)		Ma and Hovy (2016)	Li et al. (2020)	Yan et al. (2021b)	Lample et al. (2016)	
ABSA	Input	$\mathcal{X}_{asp}$	$\mathcal{X}, \mathcal{S}_{aux}$		$\mathcal{X}, \mathcal{Q}_{asp}, \mathcal{Q}_{opin\&sent}$	$\mathcal{X}$		$f_{prompt}(\mathcal{X})$
	Output	$\mathcal{Y}$	$\mathcal{Y}$		$\mathcal{X}_{asp}, \mathcal{X}_{opin}, \mathcal{Y}_{sent}$	$(\mathcal{X}_{asp_i}, \mathcal{X}_{opin_i}, \mathcal{Y}_{sent_i})_{i=1}^{m}$		$g(\mathcal{Y})$
	Example	Wang et al. (2016)	Sun et al. (2019)		Mao et al. (2021)	Yan et al. (2021a)		Li et al. (2021)
RE	Input	$\mathcal{X}$			$\mathcal{X}, \mathcal{Q}_y$	$\mathcal{X}$		$f_{prompt}(\mathcal{X})$
	Output	$\mathcal{Y}$			$\mathcal{X}_{ent}$	$(\mathcal{Y}_i, \mathcal{X}_{sub_i}, \mathcal{X}_{obj_j})_{i=1}^{m}$		$g(\mathcal{Y})$
	Example	Zeng et al. (2014)			Levy et al. (2017)	Zeng et al. (2018)		Han et al. (2021)
Summ	Input		$(\mathcal{X}, \mathcal{S}_{cand_i})_{i=1}^{n}$	$\mathcal{X}_1, \cdots, \mathcal{X}_n$		$\mathcal{X}, \mathcal{Q}_{summ}$		$\mathcal{X}$, Keywords/Prompt
	Output		$\hat{\mathcal{S}}_{cand}$	$\mathcal{Y}_1, \cdots, \mathcal{Y}_n \in \{0, 1\}^n$		$\mathcal{Y}$		$\mathcal{Y}$
	Example		Zhong et al. (2020)	Cheng and Lapata (2016)		McCann et al. (2018)		Aghajanyan et al. (2021)
Parsing	Input			$x_1, \cdots, x_n$	$\mathcal{X}, \mathcal{Q}_{child}$	$\mathcal{X}$	$(\mathcal{X}, \mathcal{C}_t)_{t=0}^{m-1}$	$(\mathcal{X}, \mathcal{Y}_i)_{i=1}^{k}$
	Output			$g(y_1, \cdots, y_n)$	$\mathcal{X}_{parent}$	$g(y_1, \cdots, y_m)$	$\mathcal{A} = a_1, \cdots, a_m$	$\hat{\mathcal{Y}}$
	Example			Strzyz et al. (2019)	Gan et al. (2021)	Vinyals et al. (2015)	Chen and Manning (2014)	Choe and Charniak (2016)

Table 1: Paradigms shift in natural language processing tasks.

NLP Paradigm Shift

- **Matching Paradigm**

{x: I love this movie, y: positive} ⟶ *{x: I love this movie [SEP] This is a great movie, y: entailment}*

- **MRC Paradigm**
  - 在原始输入后，拼接question，从原始输入中抽取对应的span

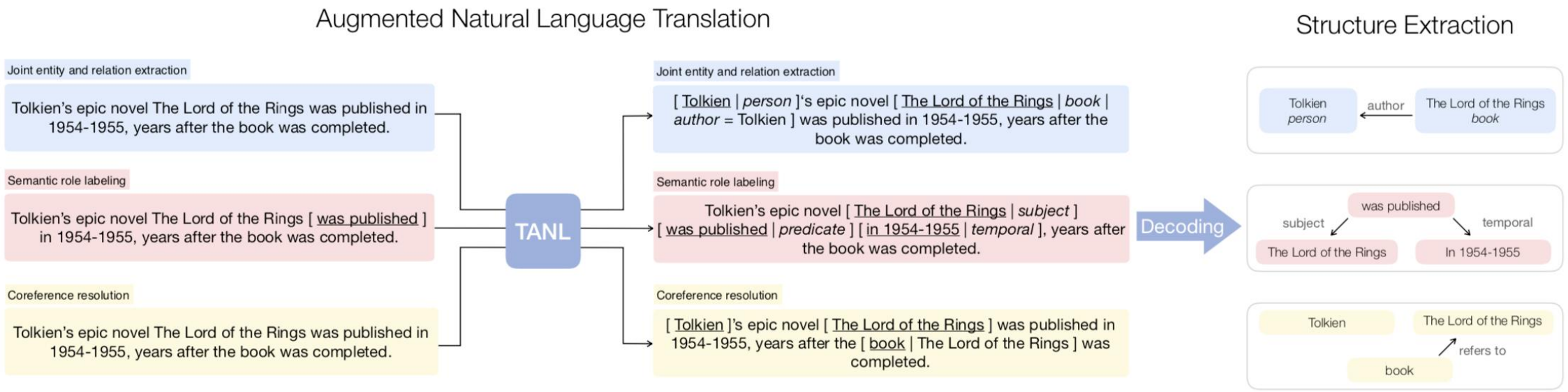
{x: Google was founded in 1998} ⟶ *{x: Google was founded in 1998 [SEP] Find organizations in the text, including companies, agencies and institutions}*

  - 应用任务：***entity-relation extraction***、***coreference resolution***、***entity linking***、***dependency parsing***、***dialog state tracking***、***event extraction***、***aspect-based sentiment analysis***

Sun et al.. *Paradigm Shift in Natural Language Processing*. Arxiv. 2021.9.26 .

NLP Paradigm Shift

- **Seq2Seq Paradigm**



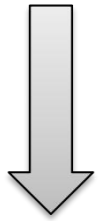
- 应用任务：***joint entity and relation extraction*** 、***named entity recognition***、***relation classification***、***semantic role labeling***、***coreference resolution***、***event extraction***、***event extraction***、***dialogue state tracking***

Paolini et al. *Structured Prediction as Translation between Augmented Natural Languages*. ICLR 2021 (Amazon).

NLP Paradigm Shift

- **MLM Paradigm (Prompt Paradigm)**

{x: I love this movie, y: positive}



{x: I love this movie. It was [MASK].}

Sun et al.. *Paradigm Shift in Natural Language Processing*. Arxiv. 2021.9.26 .

Prompt Paradigm

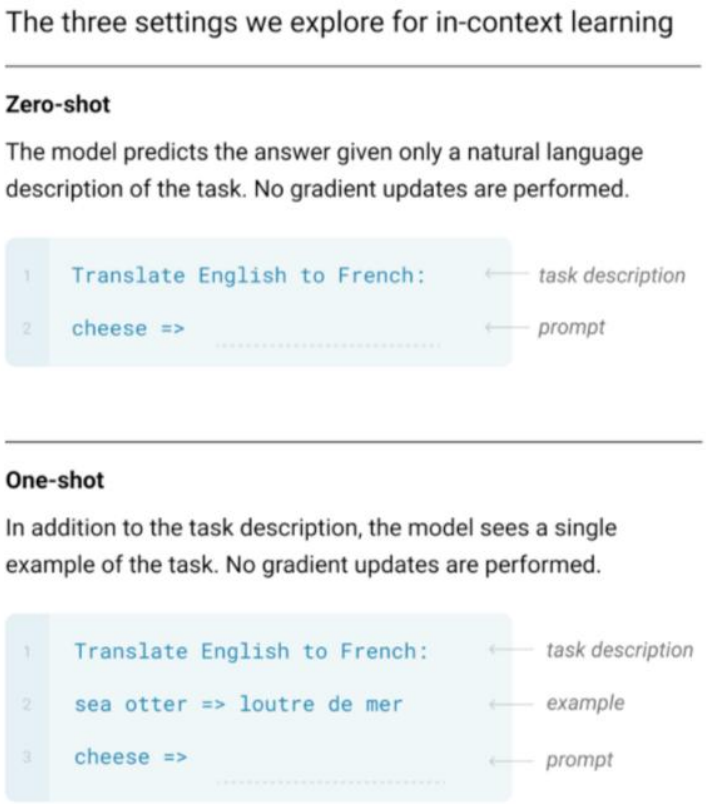
Prompt的由来与兴起：GPT-3

■ Fine-tuning问题

- 标注数据需求和硬件需求
- 精细化的预训练目标
- 预训练目标与下游任务目标不一致，PLM无法直接适配

■ GPT-3新的发现

- 无参数更新过程，直接将任务描述信息，样例及提示信息拼接后输入到Transformer模型的Decoder中，解码生成预测结果，在小样本/零样本等任务中效果很好
- **启发**：给预训练模型的一些线索和提示，让其更好地理解下游任务



Brown et al. *Language Models are Few-Shot Learners*. Arxiv. 2020.7.22.

Prompt定义

- Supervised Learning
  - 给定输入x，预测输出y，监督学习模型：$P(\boldsymbol{y}|\boldsymbol{x};\theta)$
  - 情感分析示例：

Input x: *I love this movie.*

Output y: ++(very positive) Y={++, +, ~, -, --}

- Prompting
  - 从学习$P(\boldsymbol{y}|\boldsymbol{x};\theta)$改为学习$P(\boldsymbol{x};\theta)$，再去预测y
  - 通过引入模板将输入x调整为完形填空格式的x'，调整后的输入中包含一些空槽，利用语言模型 $P(\boldsymbol{x};\theta)$ 预测槽值进而推断出y

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

Prompt定义

- Prompt Addition：将输入x 转换为$promt$ $x' = f_{\text{prompt}}(x)$
  - 采用模板转换，模板包括一个输入槽[X]和一个答案槽[Z]，将输入x填入槽[X]

Source Input x: *I love this movie.*

Template: *[X] Overall, it was a [Z] movie*

Prompt Input x': *I love this movie. Overall, it was a [Z] movie*

  - 模板可以是自然语言token（hard prompt）或非自然语言token（soft prompt）
  - 答案[Z]可在句中（cloze prompt，NLU常用)或句末（prefix prompt，NLG常用)
  - slot的数量可以为任意个

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

Prompt定义

■ Answer Search

- 将x’输入到语言模型$P(x;\theta)$，从Z中搜索使得语言模型得分最高的候选槽值

$$\hat{z} = \underset{z \in Z}{\text{search}}\, P(f_{\text{fill}}(x', z); \theta).$$

- Z可以包括词表中所有的token（生成任务），也可以是一个特定标签集合（如分类任务）
- **示例：** Z={“excellent”，“good”，“OK”，“bad”，“horrible”}
  Y={++, +, ~, -, --}

■ Answer Map

- 将得到的答案*z*与对应任务的标签*y*做1-1或*N*-1映射
- *z*与*y*的映射函数也称为*Verbalizer*

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

Prompt术语形式化表示

Name	Notation	Example	Description
Input	$\boldsymbol{x}$	I love this movie.	One or multiple texts
Output	$\boldsymbol{y}$	++ (very positive)	Output label or text
Prompting Function	$f_{\text{prompt}}(\boldsymbol{x})$	`[X]` Overall, it was a `[Z]` movie.	A function that converts the input into a specific form by inserting the input $\boldsymbol{x}$ and adding a slot `[Z]` where answer $\boldsymbol{z}$ may be filled later.
Prompt	$\boldsymbol{x}'$	I love this movie. Overall, it was a `[Z]` movie.	A text where `[X]` is instantiated by input $\boldsymbol{x}$ but answer slot `[Z]` is not.
Filled Prompt	$f_{\text{fill}}(\boldsymbol{x}', \boldsymbol{z})$	I love this movie. Overall, it was a bad movie.	A prompt where slot `[Z]` is filled with any answer.
Answered Prompt	$f_{\text{fill}}(\boldsymbol{x}', \boldsymbol{z}^*)$	I love this movie. Overall, it was a good movie.	A prompt where slot `[Z]` is filled with a true answer.
Answer	$\boldsymbol{z}$	"good", "fantastic", "boring"	A token, phrase, or sentence that fills `[Z]`

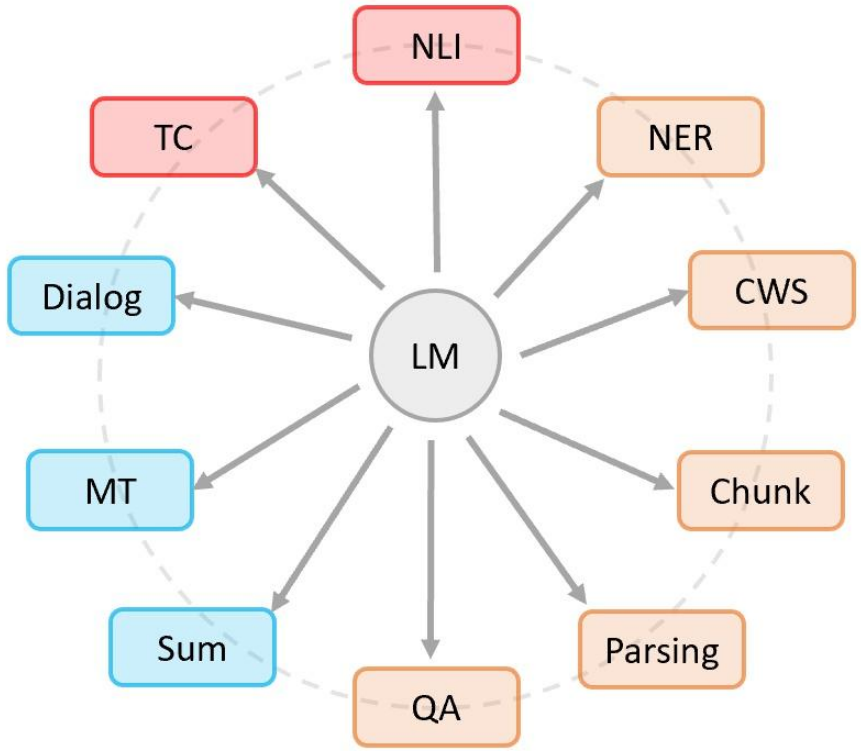
Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

不同任务的Prompt

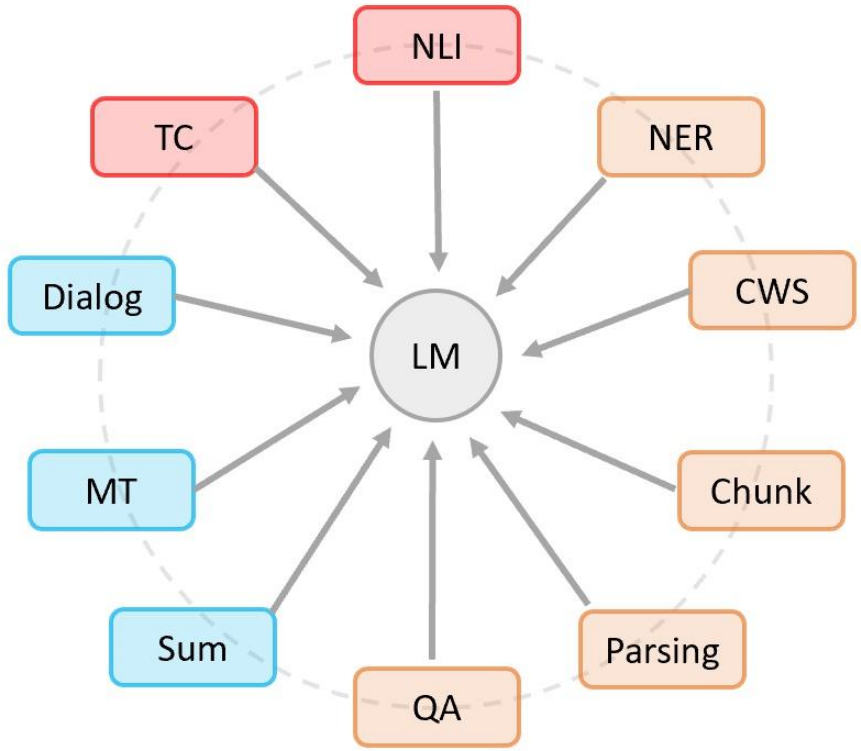
Type	Task	Input ([X])	Template	Answer ([Z])
Text CLS	Sentiment	I love this movie.	[X] The movie is [Z].	great fantastic ...
	Topics	He prompted the LM.	[X] The text is about [Z].	sports science ...
	Intention	What is taxi fare to Denver?	[X] The question is about [Z].	quantity city ...
Text-span CLS	Aspect Sentiment	Poor service but good food.	[X] What about service? [Z].	Bad Terrible ...
Text-pair CLS	NLI	[X1]: An old man with ... [X2]: A man walks ...	[X1]? [Z], [X2]	Yes No ...
Tagging	NER	[X1]: Mike went to Paris. [X2]: Paris	[X1][X2] is a [Z] entity.	organization location ...
Text Generation	Summarization	Las Vegas police ...	[X] TL;DR: [Z]	The victim ... A woman ... ...
	Translation	Je vous aime.	French: [X] English: [Z]	I love you. I fancy you. ...

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

Fine-tuning和Prompting区别



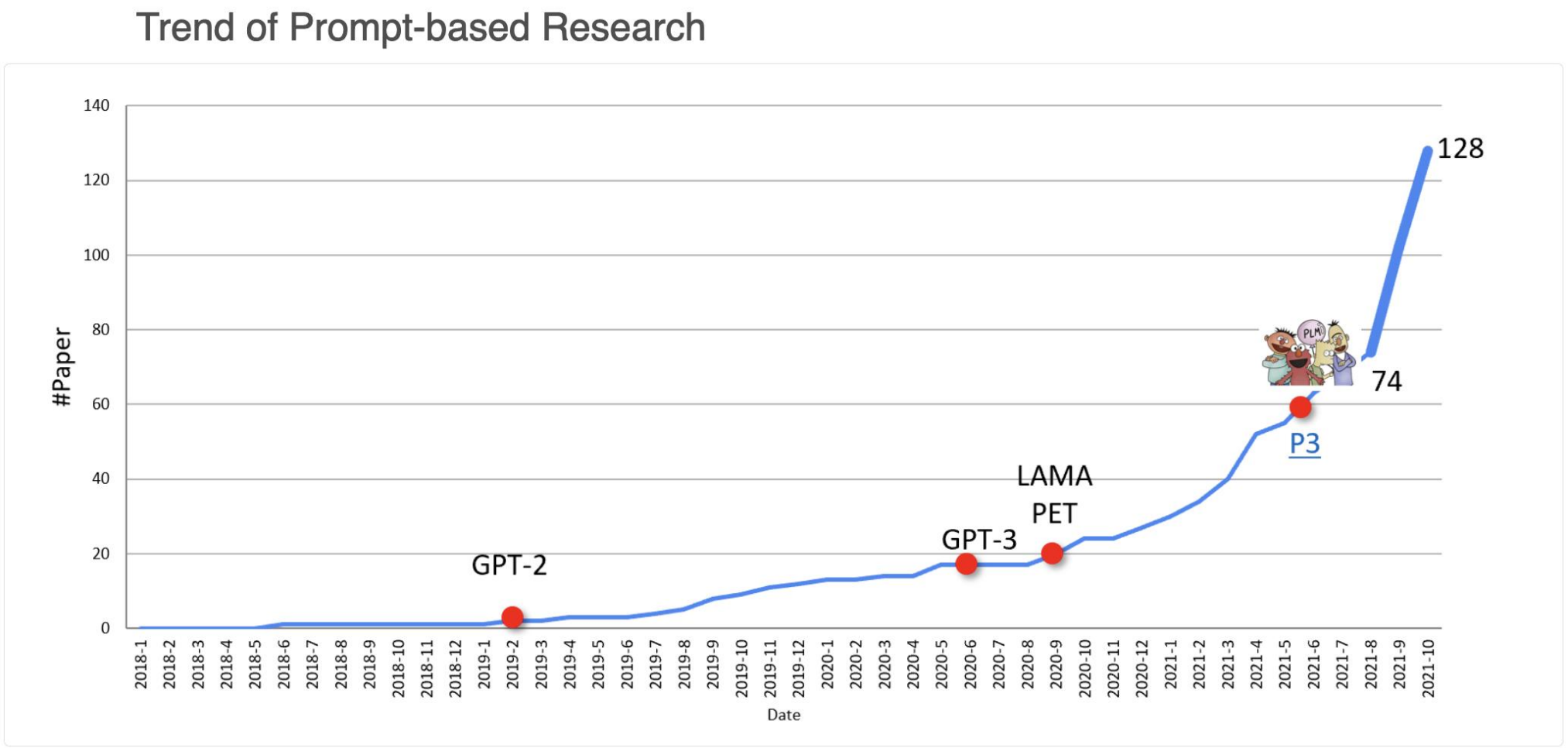
Fine-tuning



Prompting

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

Prompt研究逐步升温



Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

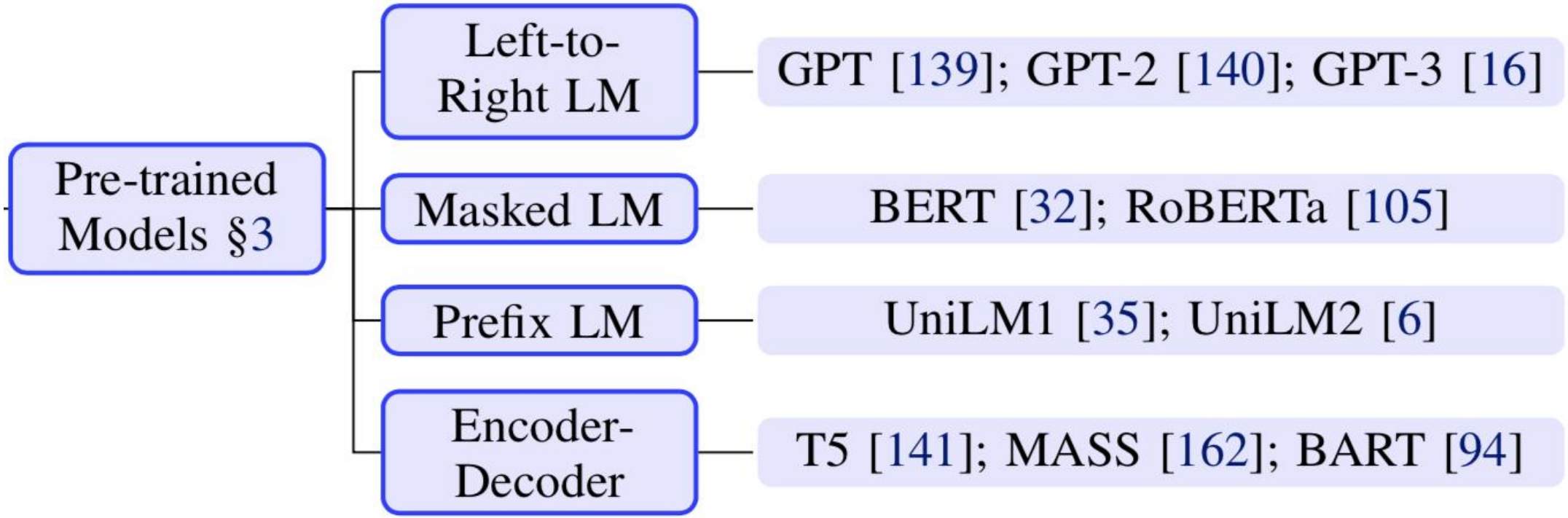
Prompt Paradigm

Design Considerations

- **P1: 预训练模型选择**
  - PLM模型结构多样，同时预训练任务也各有特色，选择与特定任务匹配的PLM模型十分必要
- **P2: Prompt Engineering**
  - 针对不同任务，选择对应合适的Prompt输入至关重要
- **P3: Answer Engineering**
  - 主要涉及模型输出的粒度形式（token,span,sentence,etc.）及模型输出到实际标签的映射方式
- **P4: Multi-Prompt Learning**
  - 相对于Single Prompt, Multiple Prompt学习能在一定程度上提升最终效果
- **P5: Training Strategies**
  - 训练Prompt的方法多种多样，LM参数及Prompt参数可进行多种组合训练

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

P1: 预训练模型选择



Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

P2: Prompt Engineering

- **Prompt Engineering是将原始输入X转化为带有Prompt提示输入的过程**

Name	Notation	Example	Description
Input	$\boldsymbol{x}$	I love this movie.	One or multiple texts
Output	$\boldsymbol{y}$	++ (very positive)	Output label or text
Prompting Function	$f_{\text{prompt}}(\boldsymbol{x})$	`[X]` Overall, it was a `[Z]` movie.	A function that converts the input into a specific form by inserting the input $\boldsymbol{x}$ and adding a slot `[Z]` where answer $\boldsymbol{z}$ may be filled later.
Prompt	$\boldsymbol{x}'$	I love this movie. Overall, it was a `[Z]` movie.	A text where `[X]` is instantiated by input $\boldsymbol{x}$ but answer slot `[Z]` is not.
Filled Prompt	$f_{\text{fill}}(\boldsymbol{x}', \boldsymbol{z})$	I love this movie. Overall, it was a bad movie.	A prompt where slot `[Z]` is filled with any answer.
Answered Prompt	$f_{\text{fill}}(\boldsymbol{x}', \boldsymbol{z}^*)$	I love this movie. Overall, it was a good movie.	A prompt where slot `[Z]` is filled with a true answer.
Answer	$\boldsymbol{z}$	"good", "fantastic", "boring"	A token, phrase, or sentence that fills `[Z]`

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

P2: Prompt Engineering

- **Prompt Shape**
  - **Cloze Prompt**
    - 填充文本中的空白字符
    - 适用于使用掩码(Mask)LM解决的任务

      Template: *[X] Overall, it was a [Z] movie*

  - **Prefix Prompt**
    - 添加在原始字符串末尾
    - 适用于生成任务或使用标准自回归 LM 解决的任务

      Template: *[X] Please continue to write: [Z]*

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

P2: Prompt Engineering

- **人工构造模板**
  - 优势：可解释性强，简单
  - 劣势：可能非最优

- **自动构造模板**
  - 离散型Prompt
    - 生成的Prompt为真实的文本字符串
  - 连续型Prompt
    - 生成的Prompt为连续的embedding向量

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

P2: Prompt Engineering

- **自动构造的离散型Prompt (Hard Prompt)**
  - **D1: Prompt Mining**
    - 给定input X and output Y，从大量文本中（Wikipedia）中挖掘同时包含X和Y的频繁句式
  - **D2: Prompt Paraphrasing**
    - 在现有的种子prompt（手动构建或自动挖掘得到的）基础上，将其转述为其他语义相近但描述不同的candidate prompt，并从候选中选择在目标任务上效果最好的prompt
    - Paraphrasing的方式可以是 1）将源语言翻译成另一种语言，然后再翻译回来；2）使用同义词表进行短语替换；3）基于定制化的神经转义模型；

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

P2: Prompt Engineering

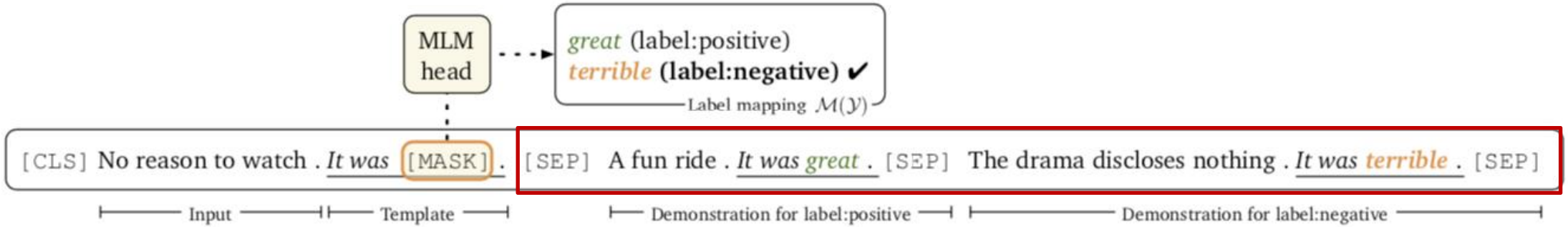
- **自动构造的离散型Prompt (Hard Prompt)**
  - **D3: Gradient-based Search**
    - 在候选词中选择部分词作为prompt并参与训练，利用梯度下降法不断搜索词的排列组合
  - **D4: Prompt Generation**
    - 使用专门的文本生成模型（如T5，BART）来生成prompt (LM-BFF模型)
  - **D5: Prompt Scoring**
    - 手工制作一组模板作为候选，然后填充输入和答案，并用LM模型对候选模板进行打分，选择分数最高的prompt，值得注意的是可以为每个输入自动选择不同的prompt

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

LM-BFF

■ Better Few-shot Fine-tuning of Language Models

- 使用比GPT-3稍小规模的LM模型（如BERT/Roberta）在Few-shot上进行实验
- 受GPT-3“in-context learning”启发，拼接部分样例到当前输入后，每个类别只选取一个语义最相似的样例



(c) Prompt-based fine-tuning with demonstrations (our approach)

Gao et al. *Making Pre-trained Language Models Better Few-shot Learners*. Arxiv. 2021.6.2(陈丹琦, 普林斯顿大学).

LM-BFF

■ 方法简介

- 不增加新的参数，只微调LM原有的参数，称 *Prompt-based Fine-tuning*
- 用T5模型生成<X>与<Y>对应的span，从而生成候选模板

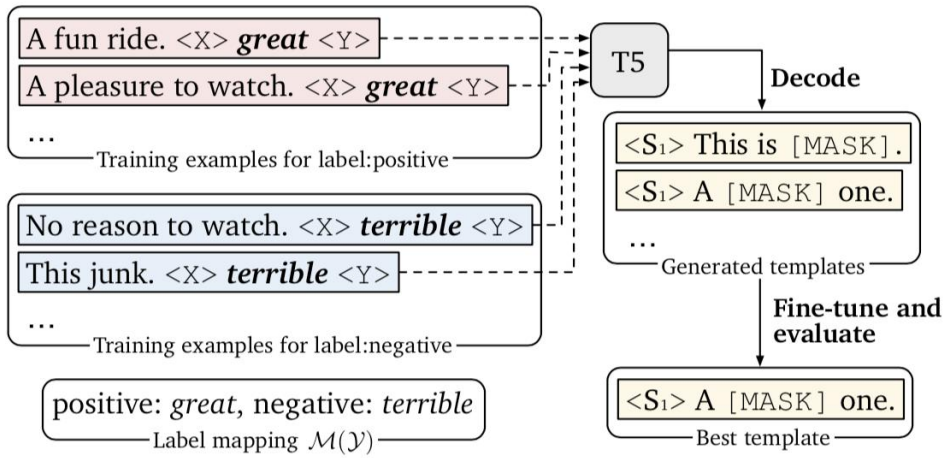


Figure 2: Our approach for template generation.

Template	Label words	Accuracy
SST-2 (positive/negative)		mean (std)
$<S_1>$ It was [MASK] .	great/terrible	**92.7 (0.9)**
$<S_1>$ It was [MASK] .	good/bad	92.5 (1.0)
$<S_1>$ It was [MASK] .	cat/dog	91.5 (1.4)
$<S_1>$ It was [MASK] .	dog/cat	86.2 (5.4)
$<S_1>$ It was [MASK] .	terrible/great	83.2 (6.9)
Fine-tuning	-	81.4 (3.8)
SNLI (entailment/neutral/contradiction)		mean (std)
$<S_1>$? [MASK] , $<S_2>$	Yes/Maybe/No	**77.2 (3.7)**
$<S_1>$. [MASK] , $<S_2>$	Yes/Maybe/No	76.2 (3.3)
$<S_1>$? [MASK] $<S_2>$	Yes/Maybe/No	74.9 (3.0)
$<S_1>$ $<S_2>$ [MASK]	Yes/Maybe/No	65.8 (2.4)
$<S_2>$? [MASK] , $<S_1>$	Yes/Maybe/No	62.9 (4.1)
$<S_1>$? [MASK] , $<S_2>$	Maybe/No/Yes	60.6 (4.8)
Fine-tuning	-	48.4 (4.8)

Table 2: The impact of templates and label words on prompt-based fine-tuning ($K = 16$).

Gao et al. *Making Pre-trained Language Models Better Few-shot Learners*. Arxiv. 2021.6.2(陈丹琦, 普林斯顿大学).

LM-BFF

- **实验效果**

在16个任务上进行测试

	SST-2 (acc)	SST-5 (acc)	MR (acc)	CR (acc)	MPQA (acc)	Subj (acc)	TREC (acc)	CoLA (Matt.)
Majority†	*50.9*	*23.1*	*50.0*	*50.0*	*50.0*	*50.0*	*18.8*	*0.0*
Prompt-based zero-shot‡	83.6	35.0	80.8	79.5	67.6	51.4	32.0	2.0
"GPT-3" in-context learning	84.8 (1.3)	30.6 (0.9)	80.5 (1.7)	87.4 (0.8)	63.8 (2.1)	53.6 (1.0)	26.2 (2.4)	-1.5 (2.4)
Fine-tuning	81.4 (3.8)	43.9 (2.0)	76.9 (5.9)	75.8 (3.2)	72.0 (3.8)	90.8 (1.8)	88.8 (2.1)	**33.9** (14.3)
Prompt-based FT (man)	92.7 (0.9)	47.4 (2.5)	87.0 (1.2)	90.3 (1.0)	84.7 (2.2)	91.2 (1.1)	84.8 (5.1)	9.3 (7.3)
+ demonstrations	92.6 (0.5)	**50.6** (1.4)	86.6 (2.2)	90.2 (1.2)	**87.0** (1.1)	**92.3** (0.8)	87.5 (3.2)	18.7 (8.8)
Prompt-based FT (auto)	92.3 (1.0)	49.2 (1.6)	85.5 (2.8)	89.0 (1.4)	85.8 (1.9)	91.2 (1.1)	88.2 (2.0)	14.0 (14.1)
+ demonstrations	**93.0** (0.6)	49.5 (1.7)	**87.7** (1.4)	**91.0** (0.9)	86.5 (2.6)	91.4 (1.8)	**89.4** (1.7)	21.8 (15.9)
Fine-tuning (full)†	*95.0*	*58.7*	*90.8*	*89.4*	*87.8*	*97.0*	*97.4*	*62.6*
	MNLI (acc)	**MNLI-mm** (acc)	**SNLI** (acc)	**QNLI** (acc)	**RTE** (acc)	**MRPC** (F1)	**QQP** (F1)	**STS-B** (Pear.)
Majority†	*32.7*	*33.0*	*33.8*	*49.5*	*52.7*	*81.2*	*0.0*	-
Prompt-based zero-shot‡	50.8	51.7	49.5	50.8	51.3	61.9	49.7	-3.2
"GPT-3" in-context learning	52.0 (0.7)	53.4 (0.6)	47.1 (0.6)	53.8 (0.4)	60.4 (1.4)	45.7 (6.0)	36.1 (5.2)	14.3 (2.8)
Fine-tuning	45.8 (6.4)	47.8 (6.8)	48.4 (4.8)	60.2 (6.5)	54.4 (3.9)	76.6 (2.5)	60.7 (4.3)	53.5 (8.5)
Prompt-based FT (man)	68.3 (2.3)	70.5 (1.9)	77.2 (3.7)	64.5 (4.2)	69.1 (3.6)	74.5 (5.3)	65.5 (5.3)	71.0 (7.0)
+ demonstrations	**70.7** (1.3)	**72.0** (1.2)	**79.7** (1.5)	**69.2** (1.9)	68.7 (2.3)	77.8 (2.0)	**69.8** (1.8)	73.5 (5.1)
Prompt-based FT (auto)	68.3 (2.5)	70.1 (2.6)	77.1 (2.1)	68.3 (7.4)	**73.9** (2.2)	76.2 (2.3)	67.0 (3.0)	75.0 (3.3)
+ demonstrations	70.0 (3.6)	**72.0** (3.1)	77.5 (3.5)	68.5 (5.4)	71.1 (5.3)	**78.1** (3.4)	67.7 (5.8)	**76.4** (6.2)
Fine-tuning (full)†	*89.8*	*89.5*	*92.6*	*93.3*	*80.9*	*91.4*	*81.7*	*91.9*

Table 3: Our main results using RoBERTa-large. †: full training set is used (see dataset sizes in Table B.1); ‡: no training examples are used; otherwise we use $K = 16$ (per class) for few-shot experiments. We report mean (and standard deviation) performance over 5 different splits (§3). Majority: majority class; FT: fine-tuning; man: manual prompt (Table 1); auto: automatically searched templates (§5.2); "GPT-3" in-context learning: using the in-context learning proposed in Brown et al. (2020) with RoBERTa-large (no parameter updates).

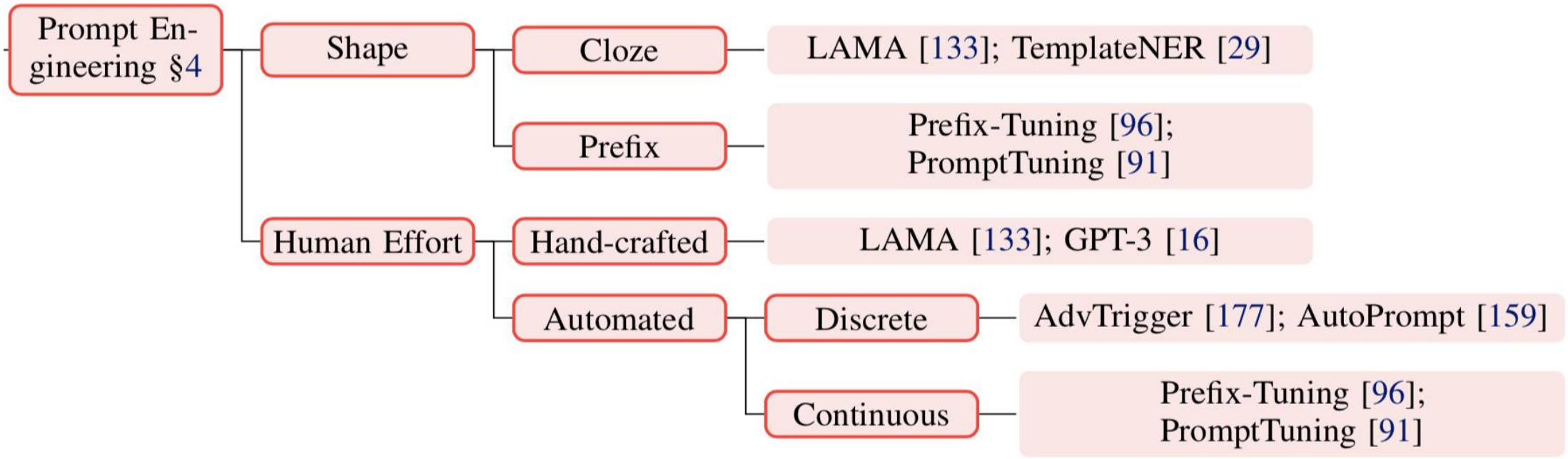
Gao et al. *Making Pre-trained Language Models Better Few-shot Learners*. Arxiv. 2021.6.2(陈丹琦, 普林斯顿大学).

Prompt Engineering

- **自动构造连续型Prompt (Soft Prompt)**
  - 由于prompt构造的目的是找到一种方法，使LM能够有效地执行任务，而不是让人类更好理解，因此没有必要将prompt限制为人类可解释的自然语言。同时，prompt可以有自己的参数，并可以根据下游任务的训练数据进行调整
  - **D1: Prefix Tuning**
    - 将Prompt以向量参数形式添加到输入中（如Prefix-tuning）
  - **D2: Tuning Initialized with Discrete Prompts**
    - 使用离散型模板中的token向量来初始化prompt向量，后续根据训练数据继续微调prompt参数
  - **D3: Hard-Soft Prompt Hybrid Tuning**
    - 向离散型模板中插入部分可学习的prompt参数（如P-tuning）

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

Prompt Engineering



Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

Answer Engineering

Prompt Engineering旨在构建合适的输入，Answer Engineering则是在搜索答案空间Z和原始标签Y的映射

- **Answer形式**
  - Token: 预训练LM词汇表中或词汇子集中的一个token，常见于分类任务
  - Span: 多个token组成的短语，常见于分类、抽取、序列标注任务等
  - Sentence: 句子，常见于文本生成任务，如翻译，摘要等
- **Answer空间**
  - 人工构造：答案和标签相同；特定任务中，构造多个输出与对应标签的映射，简单但可能不是最优
  - 自动搜索

Template	Label words	Accuracy
SST-2 (positive/negative)		mean (std)
$<S_1>$ It was [MASK] .	great/terrible	**92.7 (0.9)**
$<S_1>$ It was [MASK] .	good/bad	92.5 (1.0)
$<S_1>$ It was [MASK] .	cat/dog	91.5 (1.4)
$<S_1>$ It was [MASK] .	dog/cat	86.2 (5.4)
$<S_1>$ It was [MASK] .	terrible/great	83.2 (6.9)
Fine-tuning	-	81.4 (3.8)
SNLI (entailment/neutral/contradiction)		mean (std)
$<S_1>$? [MASK] , $<S_2>$	Yes/Maybe/No	**77.2 (3.7)**
$<S_1>$. [MASK] , $<S_2>$	Yes/Maybe/No	76.2 (3.3)
$<S_1>$? [MASK] $<S_2>$	Yes/Maybe/No	74.9 (3.0)
$<S_1>$ $<S_2>$ [MASK]	Yes/Maybe/No	65.8 (2.4)
$<S_2>$? [MASK] , $<S_1>$	Yes/Maybe/No	62.9 (4.1)
$<S_1>$? [MASK] , $<S_2>$	Maybe/No/Yes	60.6 (4.8)
Fine-tuning	-	48.4 (4.8)

Table 2: The impact of templates and label words on prompt-based fine-tuning ($K = 16$).

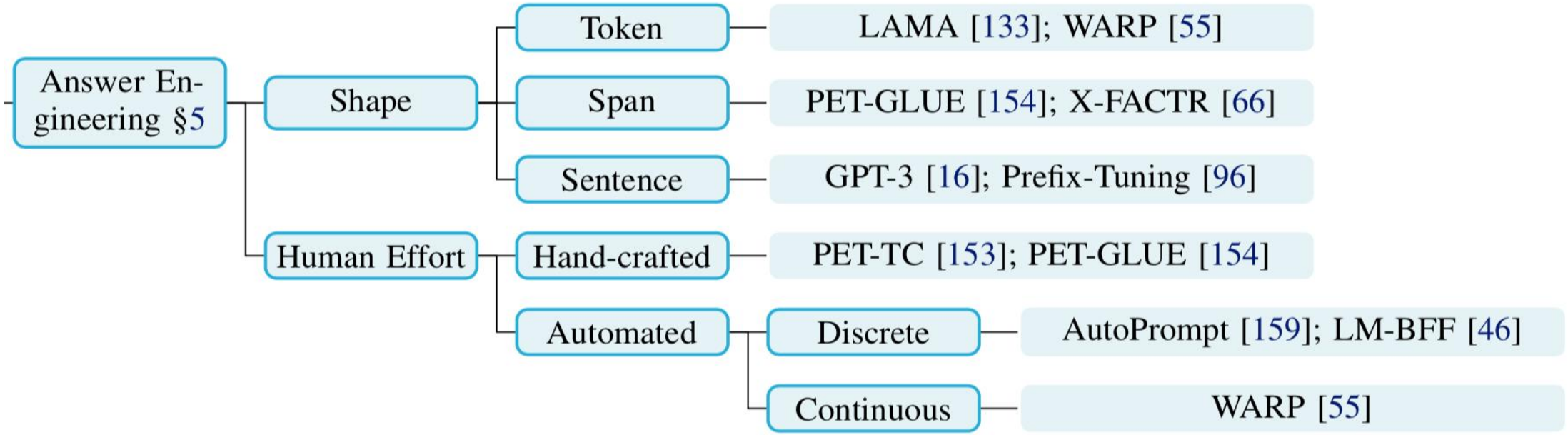
Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

Answer Engineering

- **自动离散型Answer搜索**
  - Answer Paraphrasing
    - 初始化种子answer集合，然后使用转义方法（back-translation、同义词表、转义模型）扩充种子词，最终label可以为扩充后词对应的概率和 $P(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\boldsymbol{z}\in \text{para}(\boldsymbol{z}')} P(\boldsymbol{z}|\boldsymbol{x})$
  - Prune-then-Search
    - 从整个词汇表选取一个答案子空间，然后在子空间进行搜索，例如在LM-BFF模型中首先利用语言模型直接预测[Z]位置的topK token，然后微调LM模型在验证集上选取topK中最优answer
  - Label Decomposition
    - 分解label，常见于关系抽取任务中，如关系*per:city of death*可分解为*{person, city, death}*
- **自动连续型Answer搜索**
  - 即answer不再对应自然语言的token，而仅是一个向量，不同向量对应不同的label

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

Answer Engineering



Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

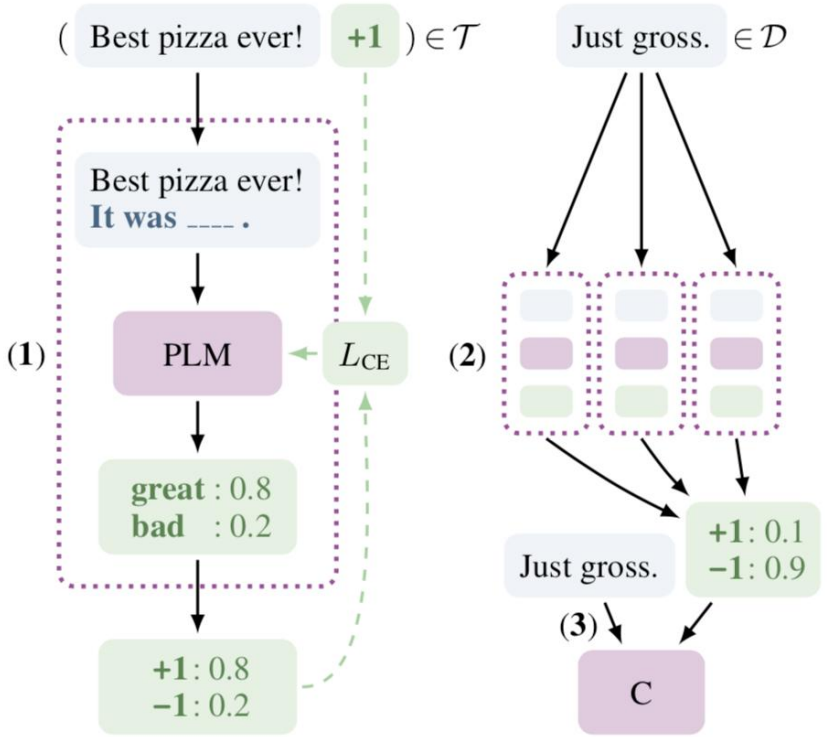
PET

■ Pattern-Exploiting Training

- 半监督方法，利用少量标签数据+大量无标签数据
- 构建多个不同的prompt，在few-shot数据集上微调PLM模型
- 使用微调后的模型标注无标签数据集D，其中每个样例的标注软标签如下所示（权重为每个不同prompt模型在训练集上的预测精度）：

$$s_{\mathcal{M}}(l \mid \mathbf{x}) = \frac{1}{Z} \sum_{\mathbf{p} \in \mathcal{P}} w(\mathbf{p}) \cdot s_{\mathbf{p}}(l \mid \mathbf{x})$$

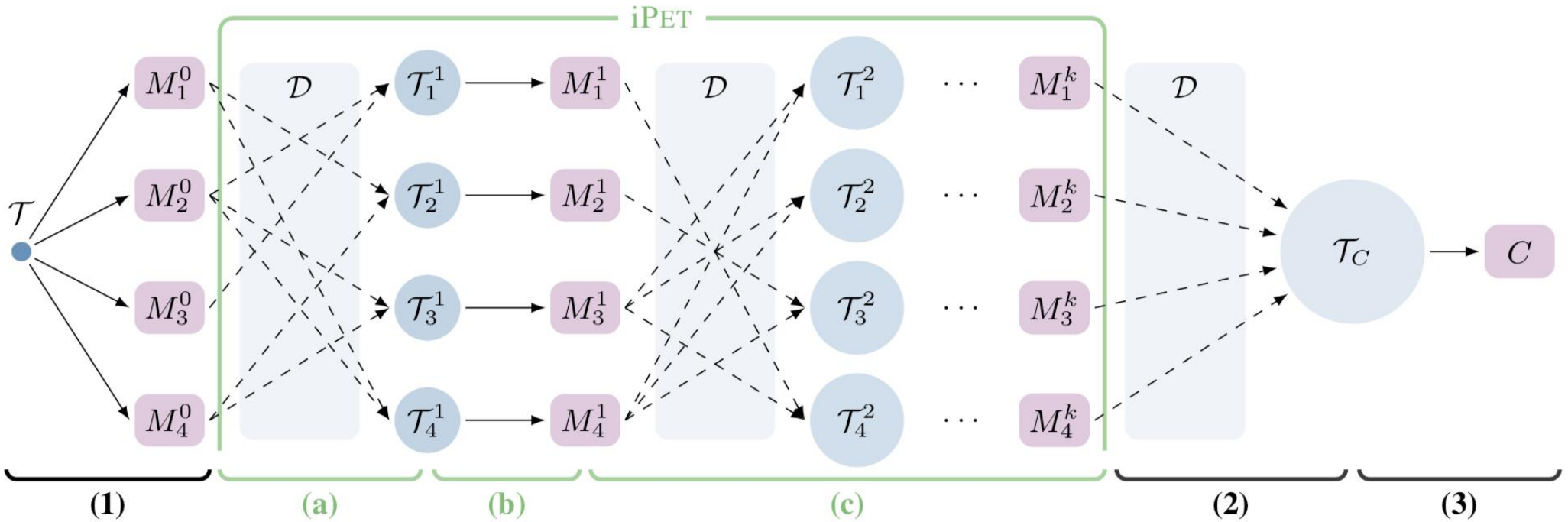
- 基于D中标注得到的软标签采用知识蒸馏的方式训练一个分类器



Schick et al. *Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference*. EACL2021. (德国慕尼黑大学)

PET

■ **PET & iPET**



- **iPET**: 对于每个使用不同prompt (pattern)的模型，采用迭代训练的方式训练K次，每次迭代使用的训练集为其他模型在无标签集合D上打标得到的数据集

Schick et al. *Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference*. EACL2021. (德国慕尼黑大学)

PET

■ 实验效果

Line	Examples	Method	Yelp	AG's	Yahoo	MNLI (m/mm)
1	$\|\mathcal{T}\| = 0$	unsupervised (avg)	33.8 ±9.6	69.5 ±7.2	44.0 ±9.1	39.1 ±4.3 / 39.8 ±5.1
2		unsupervised (max)	40.8 ±0.0	79.4 ±0.0	56.4 ±0.0	43.8 ±0.0 / 45.0 ±0.0
3		iPET	**56.7** ±0.2	**87.5** ±0.1	**70.7** ±0.1	**53.6** ±0.1 / **54.2** ±0.1
4	$\|\mathcal{T}\| = 10$	supervised	21.1 ±1.6	25.0 ±0.1	10.1 ±0.1	34.2 ±2.1 / 34.1 ±2.0
5		PET	52.9 ±0.1	87.5 ±0.0	63.8 ±0.2	41.8 ±0.1 / 41.5 ±0.2
6		iPET	**57.6** ±0.0	**89.3** ±0.1	**70.7** ±0.1	**43.2** ±0.0 / **45.7** ±0.1
7	$\|\mathcal{T}\| = 50$	supervised	44.8 ±2.7	82.1 ±2.5	52.5 ±3.1	45.6 ±1.8 / 47.6 ±2.4
8		PET	60.0 ±0.1	86.3 ±0.0	66.2 ±0.1	63.9 ±0.0 / 64.2 ±0.0
9		iPET	**60.7** ±0.1	**88.4** ±0.1	**69.7** ±0.0	**67.4** ±0.3 / **68.3** ±0.3
10	$\|\mathcal{T}\| = 100$	supervised	53.0 ±3.1	86.0 ±0.7	62.9 ±0.9	47.9 ±2.8 / 51.2 ±2.6
11		PET	61.9 ±0.0	88.3 ±0.1	69.2 ±0.0	74.7 ±0.3 / 75.9 ±0.4
12		iPET	**62.9** ±0.0	**89.6** ±0.1	**71.2** ±0.1	**78.4** ±0.7 / **78.6** ±0.5
13	$\|\mathcal{T}\| = 1000$	supervised	63.0 ±0.5	**86.9** ±0.4	70.5 ±0.3	73.1 ±0.2 / 74.8 ±0.3
14		PET	**64.8** ±0.1	**86.9** ±0.2	**72.7** ±0.0	**85.3** ±0.2 / **85.5** ±0.4

Table 1: Average accuracy and standard deviation for RoBERTa (large) on Yelp, AG's News, Yahoo and MNLI (m:matched/mm:mismatched) for five training set sizes $|\mathcal{T}|$.

Schick et al. *Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference*. EACL2021. (德国慕尼黑大学)

Multi-Prompt Learning

- **使用多重Prompt进一步提升Prompting方法的效果**

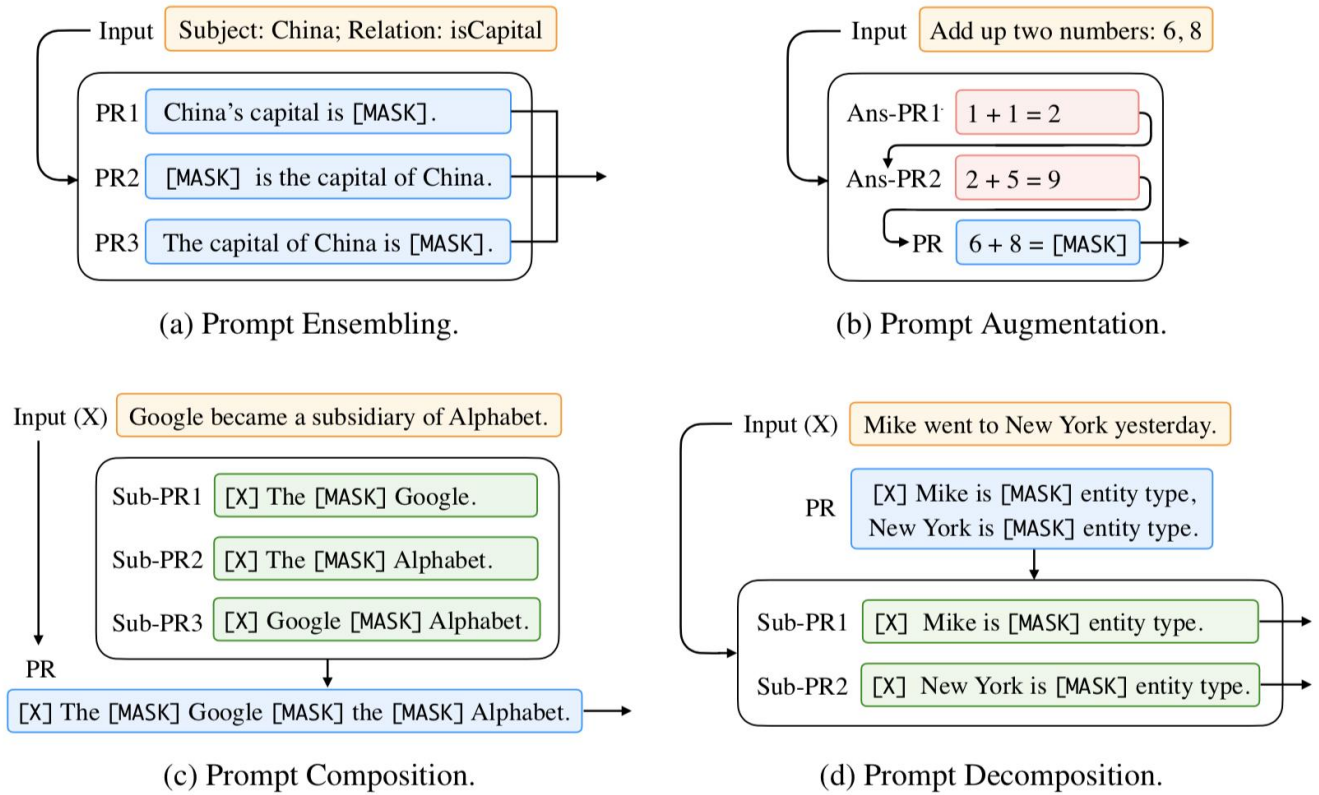
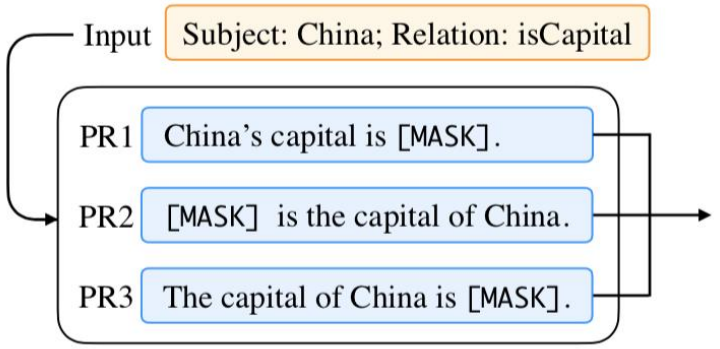


Figure 4: Different multi-prompt learning strategies. We use different colors to differentiate different components as follows. " " for input text, " " for prompt, " " for answered prompt. " " for sub-prompt. We use the following abbreviations. "PR" for prompt, "Ans-PR" for answered prompt, "Sub-PR" for sub-prompt.

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

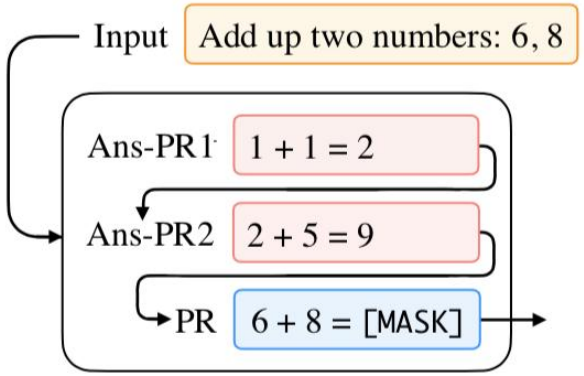
Multi-Prompt Learning

- **多重prompt策略**
  - **Prompt Ensembling**
    - 通过加权求和、多数投票等策略融合多个prompt的结果



(a) Prompt Ensembling.

  - **Prompt Augmentation**
    - 对于同一个prompt模板，加入更多样例
    - 样例的采样和排序是需要考虑的因素，如LM-BFF中仅对每个类别选择一个示例，同时基于语义相似度选择相似度最高作为演示示例

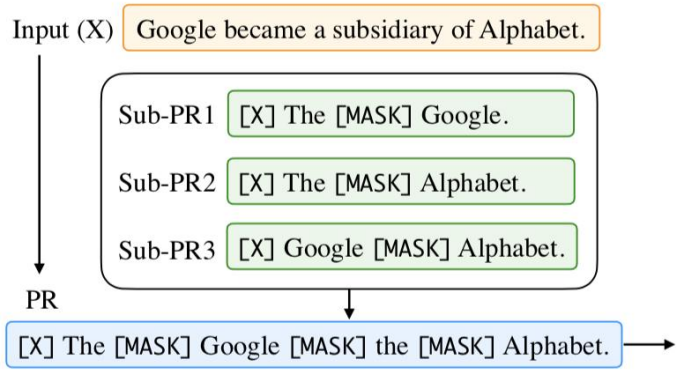


(b) Prompt Augmentation.

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

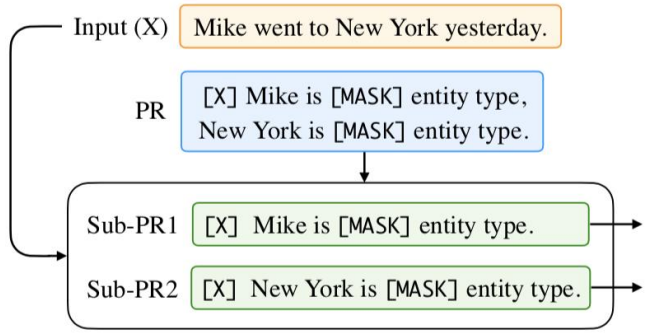
Multi-Prompt Learning

- **多重prompt策略**
  - **Prompt Composition**
    - 利用多个prompt构建prompt函数，每个prompt负责一个子任务，将多个prompt的内容进行融合，同时进行多个子任务的预测



(c) Prompt Composition.

  - **Prompt Decomposition**
    - 把具有多个预测值的任务拆分程多个单prompt任务去分别处理



(d) Prompt Decomposition.

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

Training Strategies

- **在基于prompt的下游任务学习中，包括预训练模型参数和Prompt参数**

Strategy	LM Params	Prompt Params		Example
		Additional	Tuned	
Promptless Fine-tuning	Tuned	-		ELMo [130], BERT [32], BART [94]
Tuning-free Prompting	Frozen	✗	✗	GPT-3 [16], AutoPrompt [159], LAMA [133]
Fixed-LM Prompt Tuning	Frozen	✓	Tuned	Prefix-Tuning [96], Prompt-Tuning [91]
Fixed-prompt LM Tuning	Tuned	✗	✗	PET-TC [153], PET-Gen [152], LM-BFF [46]
Prompt+LM Fine-tuning	Tuned	✓	Tuned	PADA [8], P-Tuning [103], PTR [56]

Table 6: Characteristics of different tuning strategies. "Additional" represents if there are additional parameters beyond LM parameters while "Tuned" denotes if parameters are updated.

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

Design Considerations

- **P1: 预训练模型选择**
  - PLM模型结构多样，同时预训练任务也各有特色，选择与特定任务匹配的PLM模型十分必要
- **P2: Prompt Engineering**
  - 针对不同任务，选择对应合适的Prompt输入至关重要
- **P3: Answer Engineering**
  - 主要涉及模型输出的粒度形式（token,span,sentence,etc.）及模型输出到实际标签的映射方式
- **P4: Multi-Prompt Learning**
  - 相对于Single Prompt, Multiple Prompt学习能在一定程度上提升最终效果
- **P5: Training Strategies**
  - 训练Prompt的方法多种多样，LM参数及Prompt参数可进行多种组合训练

Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. Arxiv. 2021.7.28.

Prompt Paradigm

Conclusion

- Prompt范式设计思想
  - 预训练语言模型的“知识”丰富（也可以理解为参数量大，见过的数据量多），为了更好的利用它，以重构任务为代价，设计更符合预训练任务的范式
- Prompt优点
  - 缩小了预训练目标与下游任务目标之间的Gap，降低了下游任务学习成本
  - 更充分地利用预训练模型已经学到的知识
- Prompt缺点
  - Prompt工程需要精细化设计
  - 依赖于先验，也就是取决于预训练模型的学习能力

2024年度暑期强化课程

欢迎加入DL4NLP！

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS