Making Pre-trained Language Models Better Few-shot Learners

Advisor : Jia-Ling, Koh

Speaker : Hsiao-Ting Huang

Source : ACL'2021

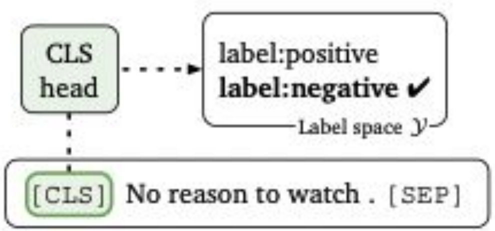
Date : 2023/05/02

Outline

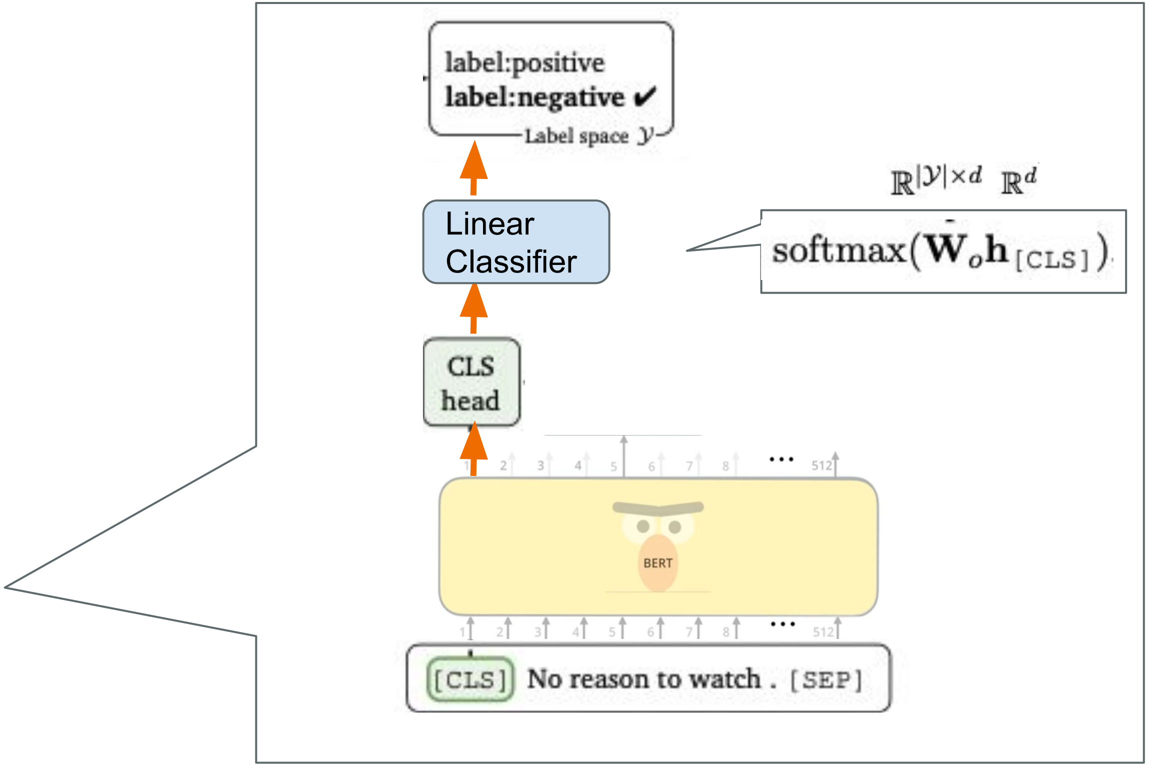
- Introduction
  - ❖ prompt-based fine-tuning
  - ❖ PET
  - ❖ in-context learning from GPT-3
- Method
  - ❖ Automatic Prompt Generation
  - ❖ Fine-tuning with Demonstrations
- Experiment
- Conclusion

Introduction :

- **standard fine-tuning**



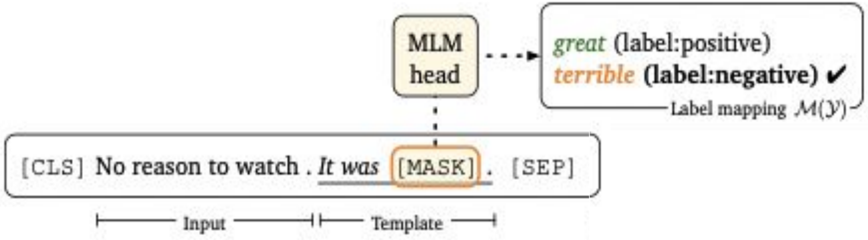
(b) Fine-tuning



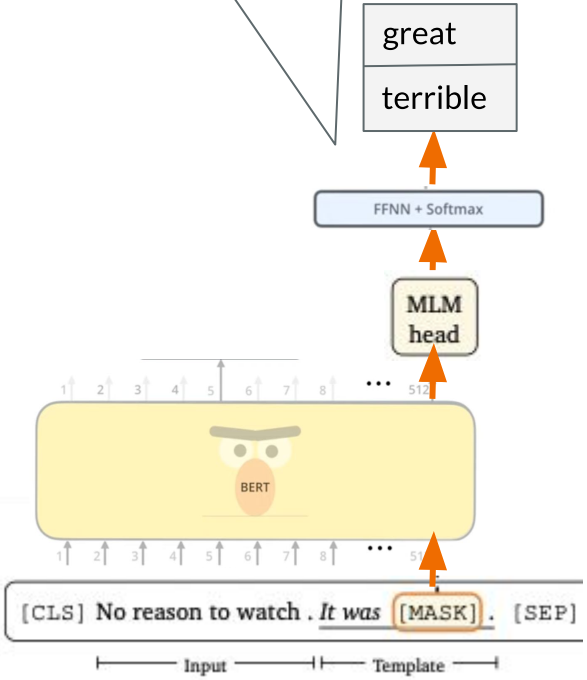
Introduction :

- **prompt-based fine-tuning**

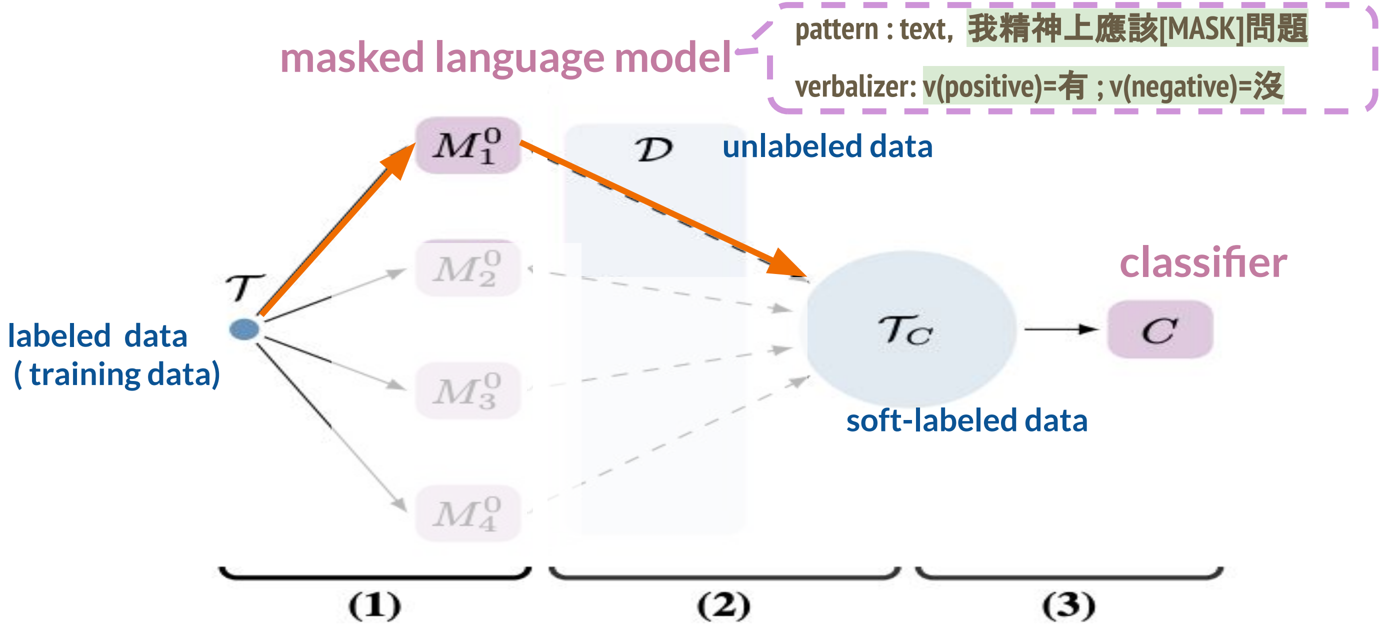
$$p(y \mid x_{\text{in}}) = p\left(\texttt{[MASK]} = \mathcal{M}(y) \mid x_{\text{prompt}}\right)$$
$$= \frac{\exp\left(\mathbf{w}_{\mathcal{M}(y)} \cdot \mathbf{h}_{\texttt{[MASK]}}\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\mathbf{w}_{\mathcal{M}(y')} \cdot \mathbf{h}_{\texttt{[MASK]}}\right)},$$



(c) Prompt-based fine-tuning



Introduction : PET



- problem:Finding the right prompts, however, is an art

Introduction : Manual prompts

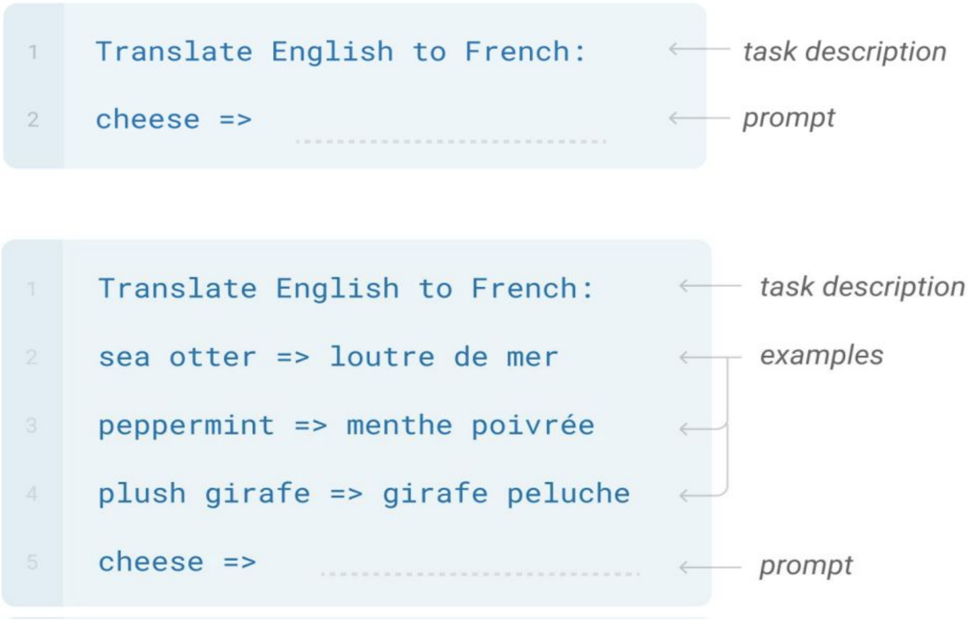
Template	Label words	Accuracy
SST-2 (positive/negative)		mean (std)
$<S_1>$ It was `[MASK]` .	great/terrible	**92.7 (0.9)**
$<S_1>$ It was `[MASK]` .	good/bad	92.5 (1.0)
$<S_1>$ It was `[MASK]` .	cat/dog	91.5 (1.4)
$<S_1>$ It was `[MASK]` .	dog/cat	86.2 (5.4)
$<S_1>$ It was `[MASK]` .	terrible/great	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)

sentiment-classification

Natural Language Inference

Introduction : in-context learning from GPT-3

- Using in-context learning of GPT-3 for machine translation.



- problem:GPT-3 consists of 175B parameters

Outline

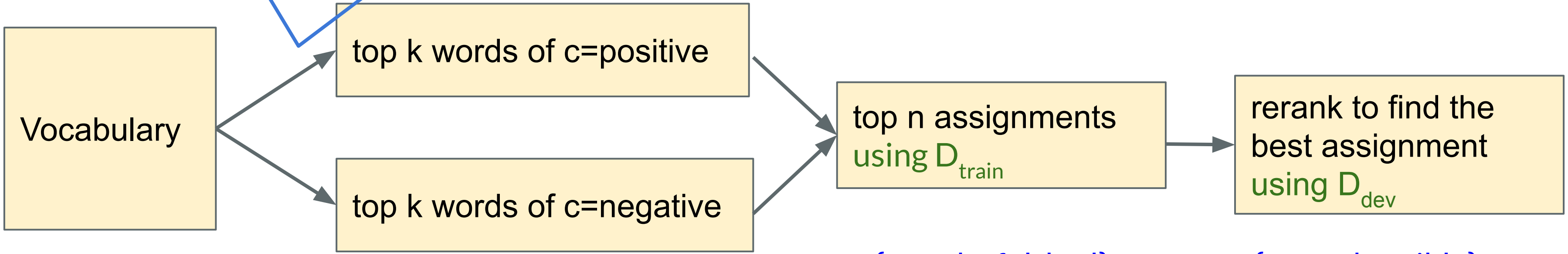
- Introduction
  - ❖ prompt-based fine-tuning
  - ❖ in-context learning from GPT-3
  - ❖ PET
- Method
  - ❖ Automatic Prompt Generation
  - ❖ Fine-tuning with Demonstrations
- Experiment
- Conclusion

Automatic Prompt Generation :

- **Automatic selection of label words**

given a fixed template T

$$\underset{v \in \mathcal{V}}{\text{Top-}k} \left\{ \sum_{x_{\text{in}} \in \mathcal{D}^{c}_{\text{train}}} \log P_{\mathcal{L}}\Big(\texttt{[MASK]} = v \mid \mathcal{T}(x_{\text{in}}) \Big) \right\}, \quad (3)$$



{wonderful, bad}
{good, bad}
{great, horrible}

{great, horrible}
{wonderful, bad}
{good, bad}

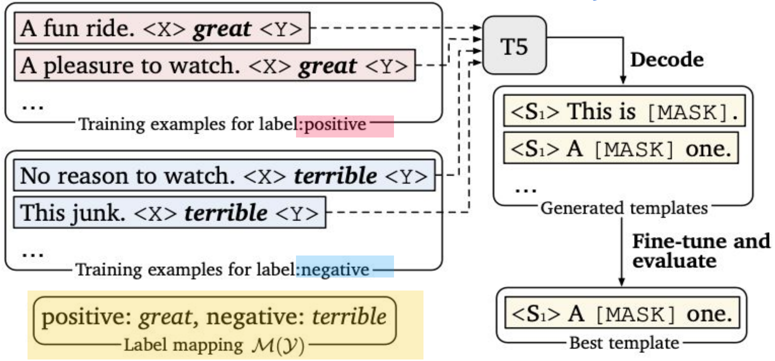
Automatic Prompt Generation :

- **Automatic generation of templates**

T5 input:

$$<S_1> \longrightarrow \ <\text{X}>\ \mathcal{M}(y)\ <\text{Y}>\ <S_1>,$$

$$<S_1> \longrightarrow \ <S_1>\ <\text{X}>\ \mathcal{M}(y)\ <\text{Y}>,$$



fix label word *M(y)*

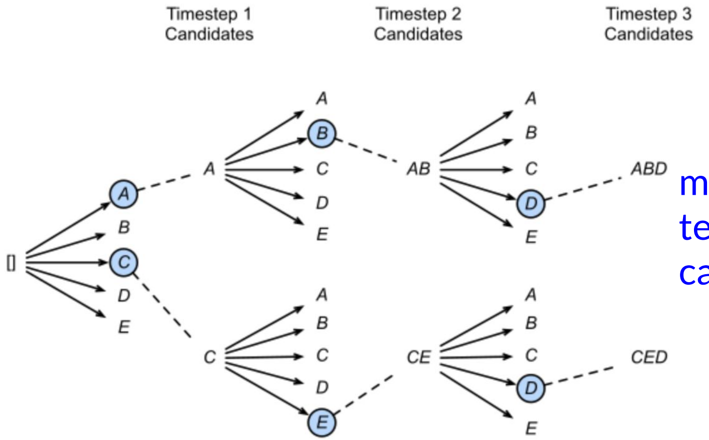
goal of T5 output:

$$\sum_{(x_{\text{in}},y)\in\mathcal{D}_{\text{train}}} \log P_{\text{T5}}(\mathcal{T} \mid \mathcal{T}_{\text{g}}(x_{\text{in}},y))$$

$$\sum_{j=1}^{|\mathcal{T}|} \sum_{(x_{\text{in}},y)\in\mathcal{D}_{\text{train}}} \log P_{\text{T5}}(t_j \mid t_1, ..., t_{j-1}, \mathcal{T}_{\text{g}}(x_{\text{in}},y)), \quad (4)$$

$(t_1, \dots, t_{|\mathcal{T}|})$ are the template tokens.

beam search(beam width = 2)



multiple template candidates

Fine-tuning with Demonstrations :

- **Sampling similar demonstrations**

$x_{\text{in}} \rightarrow$ pre-trained SBERT $\rightarrow \mathbf{e}(x_{\text{in}})$

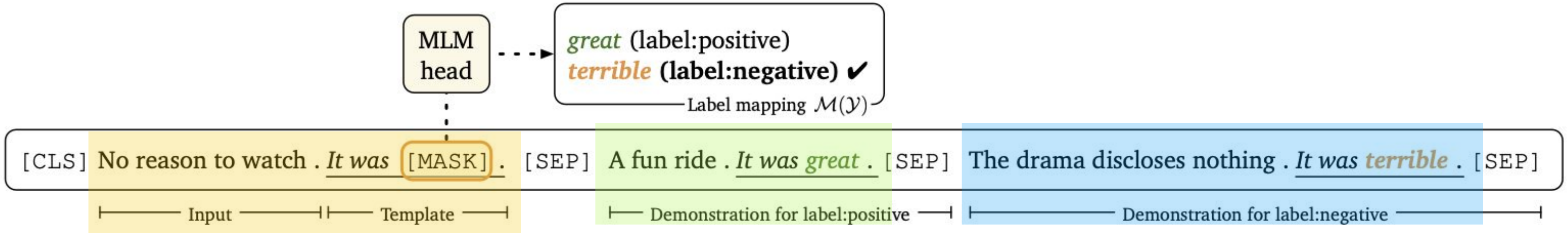
$x \in \mathcal{D}^{c}_{\text{train}} \rightarrow$ pre-trained SBERT $\rightarrow \mathbf{e}(x)$

$\rightarrow \cos(\mathbf{e}(x_{\text{in}}), \mathbf{e}(x))$

- **Training examples as demonstrations**

$$\mathcal{T}(x_{\text{in}}) \oplus \tilde{\mathcal{T}}(x_{\text{in}}^{(1)}, y^{(1)}) \oplus \cdots \oplus \tilde{\mathcal{T}}(x_{\text{in}}^{(|\mathcal{Y}|)}, y^{(|\mathcal{Y}|)}).$$

sample from the top r = 50% instances for each class



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

Outline

- Introduction
  - prompt-based fine-tuning
  - in-context learning from GPT-3
  - PET
- Method
  - Automatic Prompt Generation
  - Fine-tuning with Demonstrations
- Experiment
- Conclusion

Datasets-SST-2

Category	Dataset	$\|\mathcal{Y}\|$	L	#Train	#Test	Type	Labels (classification tasks)
	SST-2	2	19	6,920	872	sentiment	positive, negative

sentence (string)	label (class label)
"hide new secretions from the parental units "	0 (negative)
"contains no wit , only labored gags "	0 (negative)
"that loves its characters and communicates something rather beautiful about human nature "	1 (positive)

- manual prompt:

$<S_1>$ It was [MASK] .	positive: great, negative: terrible

- auto prompt:

$<S_1>$ A [MASK] one .	irresistible/pathetic
$<S_1>$ A [MASK] piece .	wonderful/bad
$<S_1>$ All in all [MASK] .	delicious/bad

Datasets-TREC

Category	Dataset	$\lvert\mathcal{Y}\rvert$	L	#Train	#Test	Type	Labels (classification tasks)
	TREC	6	10	5,452	500	question cls.	abbr., entity, description, human, loc., num.

text (string)	coarse_label label)
"How did serfdom develop in and then leave Russia ?"	2 (DESC)
"What films featured the character Popeye Doyle ?"	1 (ENTY)
"How many points make up a perfect fivepin bowling score ?"	5 (NUM)
"Who was the inventor of silly putty ?"	3 (HUM)
"What is the highest waterfall in the United States ?"	4 (LOC)
"What does the abbreviation AIDS stand for ?"	0 (ABBR)

- manual prompt:

[MASK] : $<S_1>$

abbreviation: Expression, entity: Entity, description: Description
human: Human, location: Location, numeric: Number

- auto prompt:

Q: [MASK] : $<S_1>$
$<S_1>$ Why [MASK]?
$<S_1>$ Answer: [MASK] .

Application/Advisor/Discussion/Culture/Assignment/Minute
Production/AE/Context/Artist/Assignment/Minute
Personality/Advisor/Conclusion/Hum/Assignment/Minute

Datasets-MNLI

Category	Dataset	$\lvert\mathcal{Y}\rvert$	L	#Train	#Test	Type	Labels (classification tasks)
	TREC	6	10	5,452	500	question cls.	abbr., entity, description, human, loc., num.

sentence1	sentence2	label
"Fun for adults and children."	"Fun for only children."	2 (contradiction)
"Issues in Data Synthesis."	"Problems in data synthesis."	0 (entailment)

- manual prompt:

$<S_1>$? [MASK] , $<S_2>$	entailment: Yes, netural: Maybe, contradiction: No

- auto prompt:

$<S_1>$. [MASK] , you are right , $<S_2>$	Fine/Plus/Otherwise
$<S_1>$. [MASK] you're right $<S_2>$	There/Plus/Otherwise
$<S_1>$. [MASK] ! $<S_2>$	Meaning/Plus/Otherwise

Experiment

	SST-2 (acc)	TREC (acc)	MNLI (acc)	SNLI (acc)	RTE (acc)	MRPC (F1)
Majority†	*50.9*	*18.8*	*32.7*	*33.8*	*52.7*	*81.2*
Prompt-based zero-shot‡	83.6	32.0	50.8	49.5	51.3	61.9
"GPT-3" in-context learning	84.8 (1.3)	26.2 (2.4)	52.0 (0.7)	47.1 (0.6)	60.4 (1.4)	45.7 (6.0)
Fine-tuning	81.4 (3.8)	88.8 (2.1)	45.8 (6.4)	48.4 (4.8)	54.4 (3.9)	76.6 (2.5)
Prompt-based FT (man)	92.7 (0.9)	84.8 (5.1)	68.3 (2.3)	77.2 (3.7)	69.1 (3.6)	74.5 (5.3)
+ demonstrations	92.6 (0.5)	87.5 (3.2)	**70.7** (1.3)	**79.7** (1.5)	68.7 (2.3)	77.8 (2.0)
Prompt-based FT (auto)	92.3 (1.0)	88.2 (2.0)	68.3 (2.5)	77.1 (2.1)	**73.9** (2.2)	76.2 (2.3)
+ demonstrations	**93.0** (0.6)	**89.4** (1.7)	70.0 (3.6)	77.5 (3.5)	71.1 (5.3)	**78.1** (3.4)
Fine-tuning (full)†	*95.0*	*97.4*	*89.8*	*92.6*	*80.9*	*91.4*

Experiment - ensemble model

Prompt-based Fine-tuning	MNLI	RTE
Our single manual $\mathcal{P}$	68.3 (2.3)	69.1 (3.6)
$\mathcal{P}_{\text{PET}}$	71.9 (1.5)	69.2 (4.0)
$\mathcal{P}_{\text{ours}}, \lvert\mathcal{P}_{\text{ours}}\rvert = \lvert\mathcal{P}_{\text{PET}}\rvert$	70.4 (3.1)	73.0 (3.2)
+ demonstrations	74.0 (1.9)	71.9 (4.6)
$\mathcal{P}_{\text{ours}}, \lvert\mathcal{P}_{\text{ours}}\rvert = 20$	72.7 (2.5)	**73.1** (3.3)
+ demonstrations	**75.4** (1.6)	72.3 (4.5)

manual prompt

auto prompt

Table 4: Ensemble models using manual prompts from PET (Schick and Schütze, 2021a,b) and our automatic templates. PET uses 4 prompts for MNLI and 5 for RTE. We also use an equal number of templates in $|\mathcal{P}_{\text{ours}}| = |\mathcal{P}_{\text{PET}}|$ for a fair comparison.

Experiment - manual prompts vs.auto prompt

		SST-2	SNLI	TREC	MRPC
manual prompt	Manual	**92.7**	**77.2**	84.8	74.5
auto prompt	Auto T	92.3	77.1	88.2	76.2
	Auto L	91.5	75.6	87.0	**77.2**
	Auto T + L	92.1	77.0	**89.2**	74.0

SST-2	(positive/negative)
Auto T	$\mathcal{M}(\mathcal{Y})$ = {great, terrible} #1. <S_1> A `[MASK]` one . #2. <S_1> A `[MASK]` piece . #3. <S_1> All in all `[MASK]` .
Auto L	$\mathcal{T}(x_{\text{in}})$ = <S_1> It was `[MASK]`. #1. irresistible/pathetic #2. wonderful/bad #3. delicious/bad
SNLI	(entailment/neutral/contradiction)
Auto T	$\mathcal{M}(\mathcal{Y})$ = {Yes, Maybe, No} #1. <S_1> . `[MASK]` , no , <S_2> #2. <S_1> . `[MASK]` , in this case <S_2> #3. <S_1> . `[MASK]` this time <S_2>
Auto L	$\mathcal{T}(x_{\text{in}})$ = <S_1> ? `[MASK]` , <S_2> #1. Alright/Watch/Except #2. Hi/Watch/Worse #3. Regardless/Fortunately/Unless

Experiment - Impact of demonstration sampling strategies

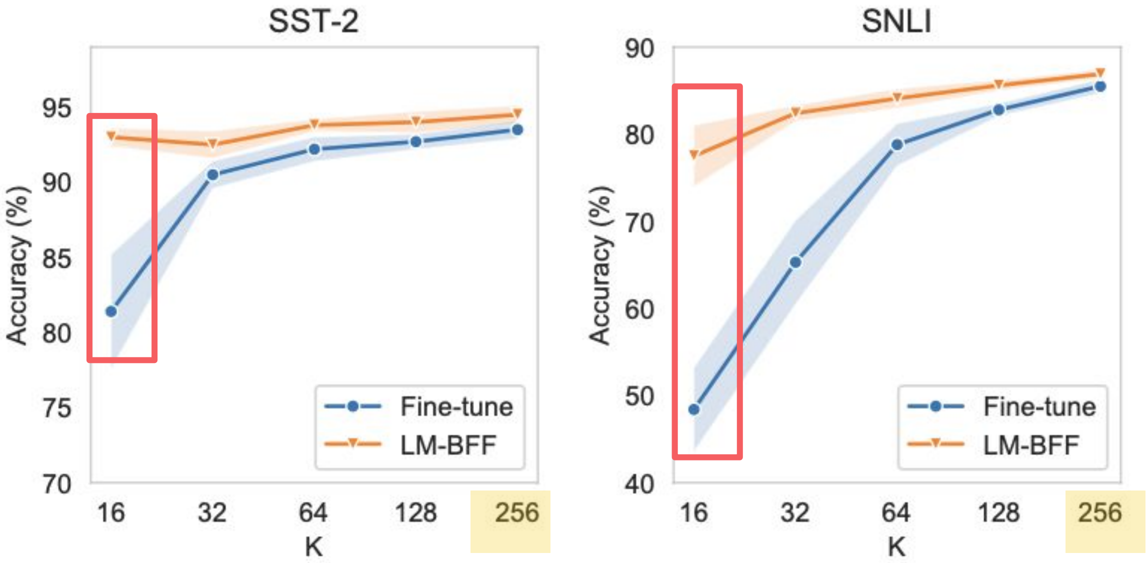
Prompt-based FT (man)

random sample for each class

sample from the top r = 50% for each class

	SST-2	SNLI	TREC	MRPC
Prompt-based FT	**92.7**	77.2	84.8	74.5
Uniform sampling	92.3	78.8	85.6	70.9
+ RoBERTa sel.	**92.7**	79.5	83.4	76.6
+ SBERT sel.	92.6	**79.7**	**87.5**	**77.8**

Experiment - fine-tuning vs our LM-BFF



Conclusion

- presented LM-BFF, a set of simple but effective techniques for fine-tuning language models using only a few examples

- (1) use prompt-based fine- tuning with automatically searched prompts
- (2) include selected task demonstrations (training examples) as part of the input context.