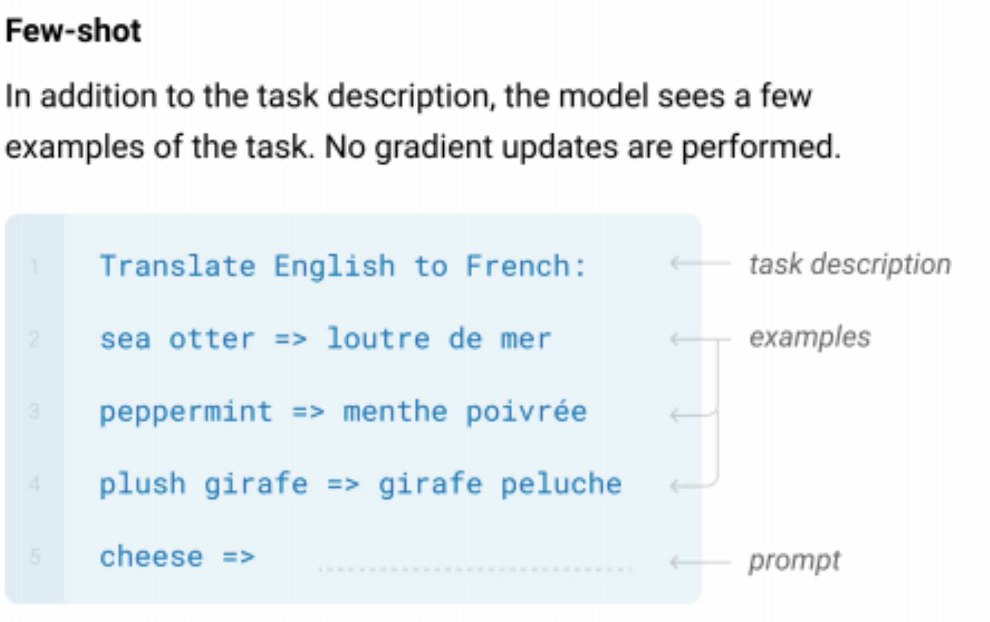
Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference

백인진 22.08.30

Backgrounds

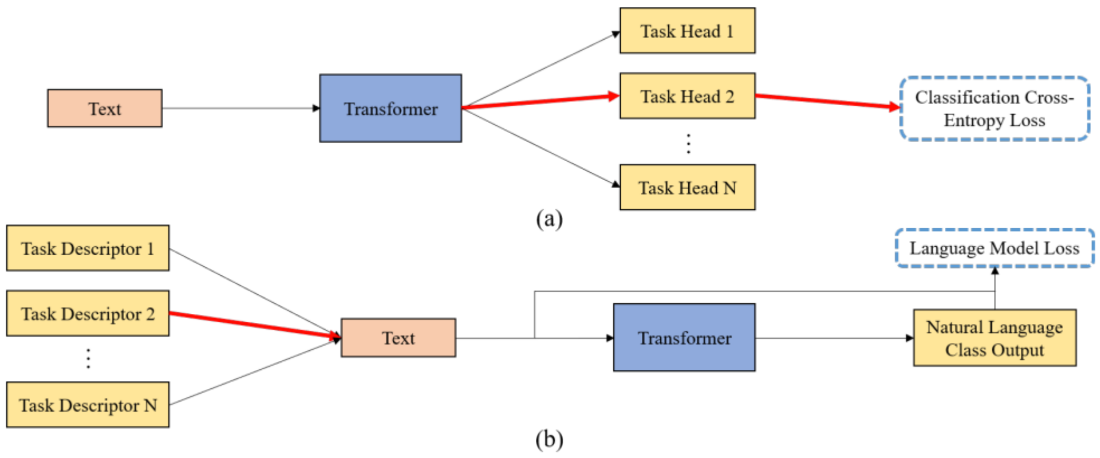
Few-Shot Learning



The vast number of languages, domains and tasks and the cost of annotating data

-> making few-shot learning a highly important research area

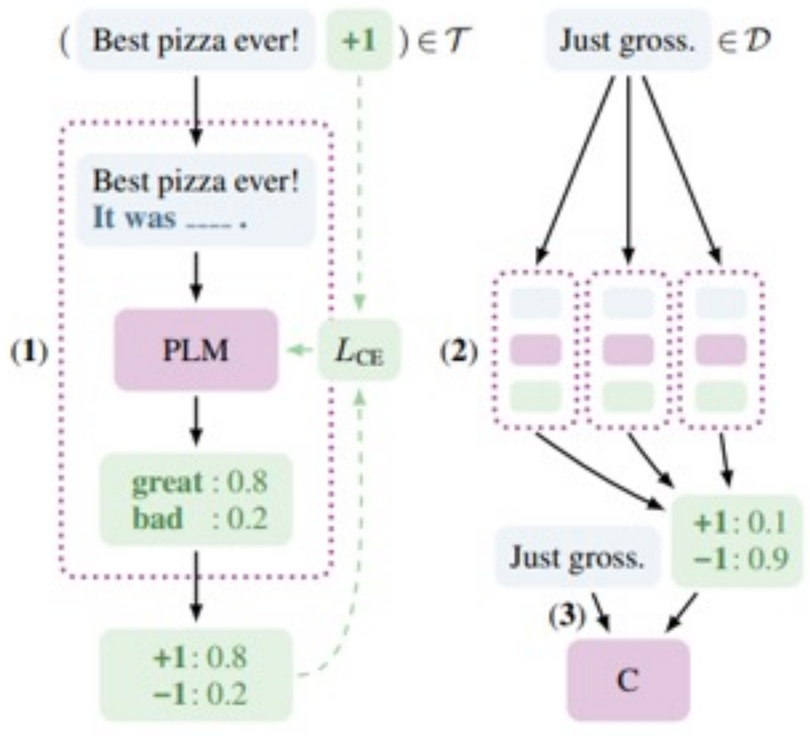
Providing Task Description



Simply appending task descriptions in natural language to an input

-> zero-shot scenarios where no training data is available at all

- PET



1. Train dataset(T) is transformed into a cloze question form to train PLM. (PLM is fine-tuned in each cloze question pattern)
2. Each PLM is ensembled to annotate the unlabeled data (D) as soft-label.
3. Text Classifier is trained with soft-labeled datasets.

PET notations

Notation

- M : Masked Language Model(MLM)
- V : Vocabulary
- __ : Mask Token ($\in V$)
- A : A specific Task
- $\mathcal{L}$: A set of Labels for classification task
- $x = (s_1, \dots, s_k)$: a sequence of phrases ($s_i \in V$)
- P : Pattern, where $P(x) \in V^*$
- v : Verbalizer, $\mathcal{L} \to V$ (mapping Label to a word belonging to V of M)
- (P, v) : Pattern-Verbalizer-Pair (PVP)

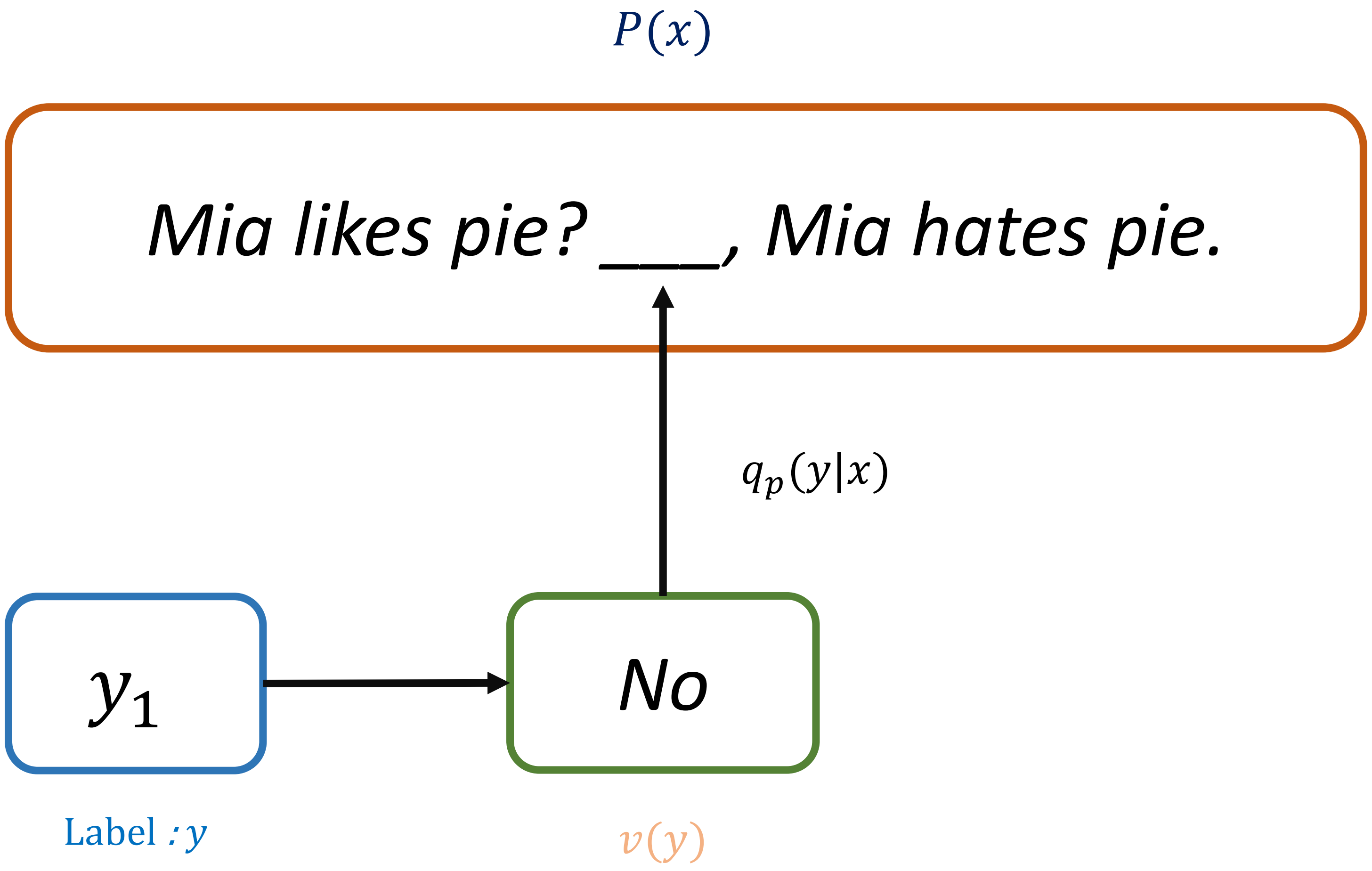
Examples

Task (A)	Identifying whether two sentences contradict each other or agree with each other
Sentence 1 (s_1)	Mia likes pie
Sentence 2 (s_2)	Mia hates pie
Label ($\mathcal{L}$)	y_1

$x : [s_1 . s_2] = [Mia\ likes\ pie, Mia\ hates\ pie]$
$v : y_0 \to Yes, \quad y_1 \to No$
$P : [s_1 ? \_\_, s_2]$

$P(x) : [Mia\ likes\ pie? \_\_, Mia\ hates\ pie]$

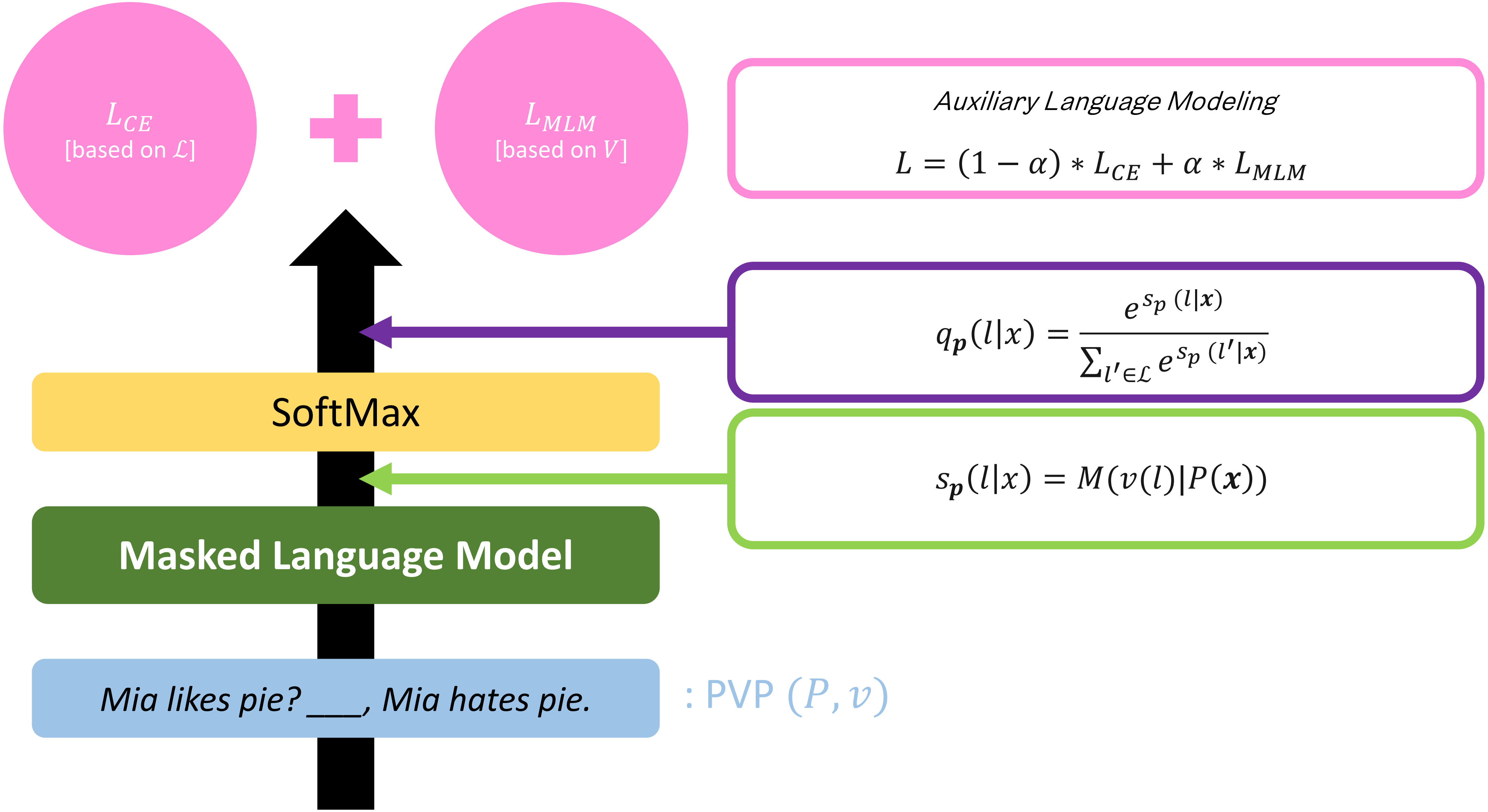
PVP method



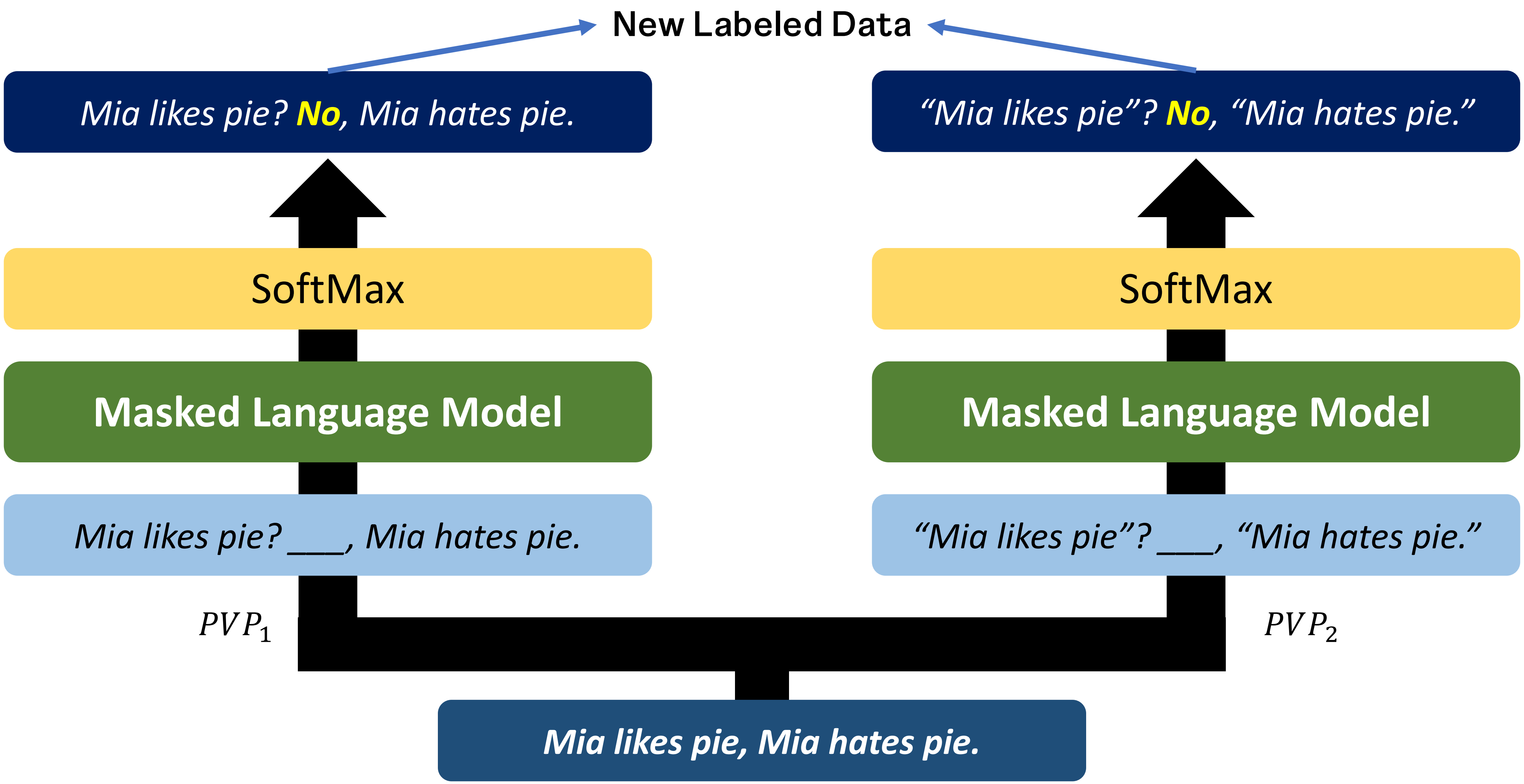
- **Notation**

- M : Masked Language Model(MLM)
- V : Vocabulary
- __ : Mask Token ($\in V$)
- A : A specific Task
- $\mathcal{L}$: A set of Labels for classification task
- $x = (s_1, \ldots, s_k)$: a sequence of phrases ($s_i \in V$)
- P : Pattern, where $P(x) \in V^*$
- v : Verbalizer, $\mathcal{L} \to V$
- (P, v) : Pattern-Verbalizer-Pair (PVP)

PVP training and inference

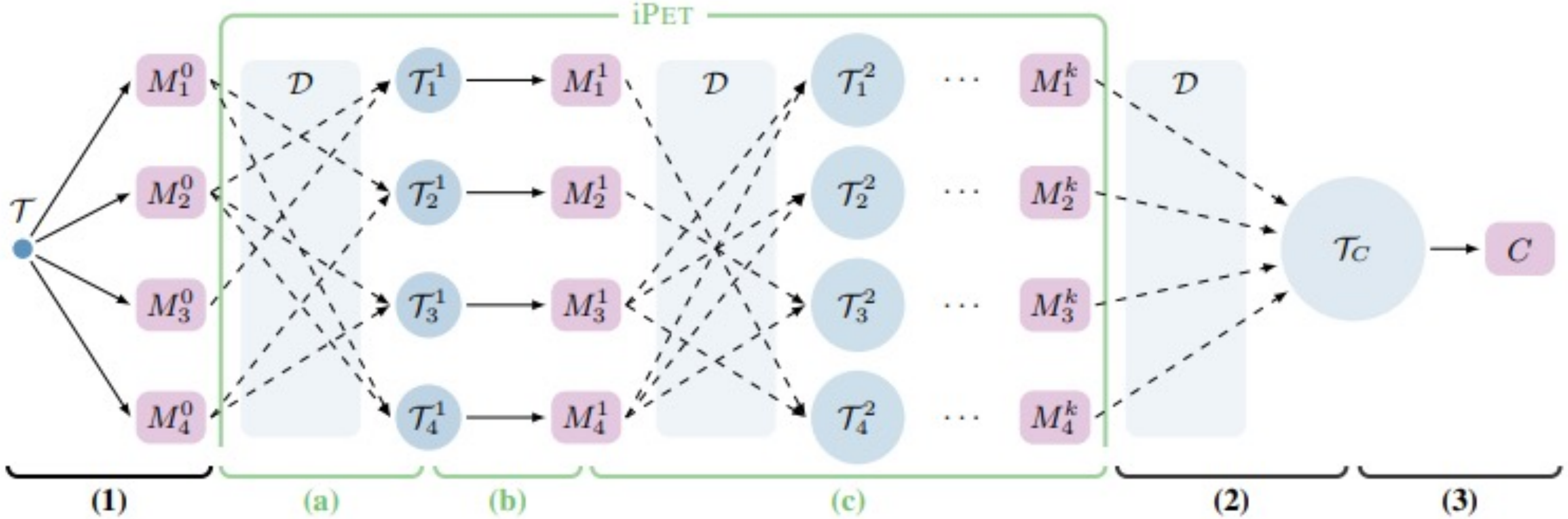


Combining PVPs



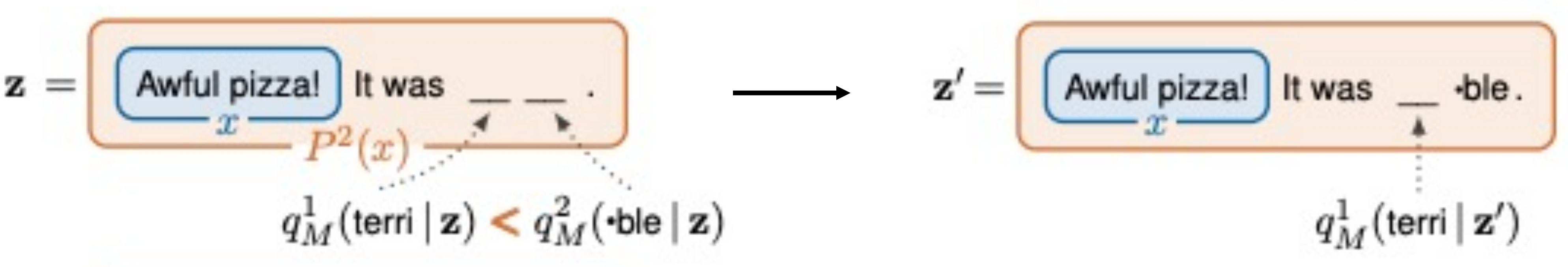
PET : using New Labeled Data

iPET



- PET : (1)&(2)&(3)
  - No interaction between patterns
- $iPET$: (1)&(a)&(b)&(c)&(2)&(3)
  - Iterative PET for interaction between pattern

PET with Multiple Masks



- *PET with Multiple Tasks*
  - Because output may not be made up of a single token, multiple MASK tokens are placed to populate in order of high probability

$$q(t_1, \ldots, t_k|\boldsymbol{z}) = \begin{cases} 1, & if\ k = 0 \\ q_M^j(t_j|\boldsymbol{z}) * (q(t'|\boldsymbol{z}'), & if\ k \geq 1 \end{cases}$$

- How to make patterns
  - So far, it's made in a **manual** way

for **WiC** task

"s_1" / "s_2". Similar sense of "w"? __.

s_1 s_2 Does w have the same meaning in both sentences? __

w. Sense (1) (a) "s_1" (__) "s_2"

- A problem of determining whether one word used in the two sentences has the same meaning.

for **MultiRC** task

p. Question: q? Is it a? __.

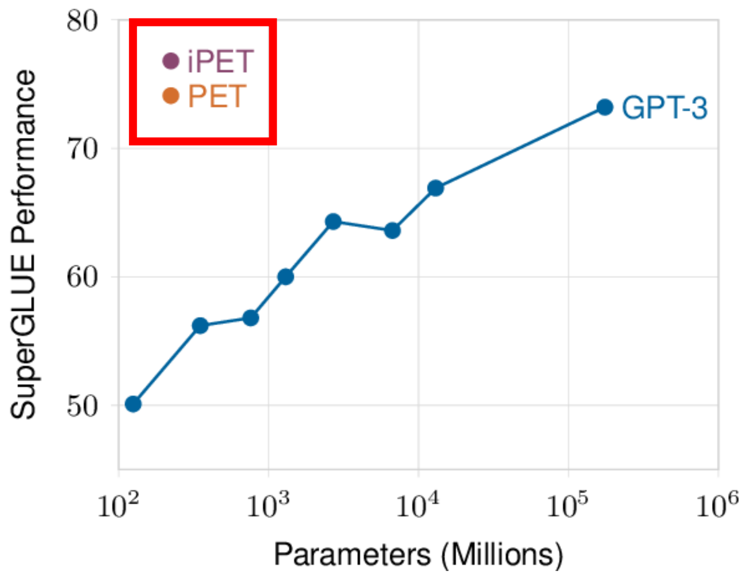
p. Question: q? Is the correct answer "a"? __.

p. Based on the previous passage, q? Is "a" a correct answer? __.

- As one of the QA tasks, it is a matter of determining whether an appropriate answer to the question is correct.

Result

	Model	Params (M)	BoolQ Acc.	CB Acc. / F1	COPA Acc.	RTE Acc.	WiC Acc.	WSC Acc.	MultiRC EM / F1a	ReCoRD Acc. / F1	Avg –
dev	GPT-3 Small	125	43.1	42.9 / 26.1	67.0	52.3	49.8	58.7	6.1 / 45.0	69.8 / 70.7	50.1
	GPT-3 Med	350	60.6	58.9 / 40.4	64.0	48.4	55.0	60.6	11.8 / 55.9	77.2 / 77.9	56.2
	GPT-3 Large	760	62.0	53.6 / 32.6	72.0	46.9	53.0	54.8	16.8 / 64.2	81.3 / 82.1	56.8
	GPT-3 XL	1,300	64.1	69.6 / 48.3	77.0	50.9	53.0	49.0	20.8 / 65.4	83.1 / 84.0	60.0
	GPT-3 2.7B	2,700	70.3	67.9 / 45.7	83.0	56.3	51.6	62.5	24.7 / 69.5	86.6 / 87.5	64.3
	GPT-3 6.7B	6,700	70.0	60.7 / 44.6	83.0	49.5	53.1	67.3	23.8 / 66.4	87.9 / 88.8	63.6
	GPT-3 13B	13,000	70.2	66.1 / 46.0	86.0	60.6	51.1	75.0	25.0 / 69.3	88.9 / 89.8	66.9
	GPT-3	175,000	77.5	82.1 / 57.2	92.0	72.9	**55.3**	75.0	32.5 / 74.8	**89.0 / 90.1**	73.2
	PET	223	79.4	85.1 / 59.4	**95.0**	69.8	52.4	**80.1**	**37.9 / 77.3**	86.0 / 86.5	74.1
	iPET	223	**80.6**	**92.9 / 92.4**	**95.0**	**74.0**	52.2	**80.1**	33.0 / 74.0	86.0 / 86.5	**76.8**
test	GPT-3	175,000	76.4	75.6 / 52.0	**92.0**	69.0	49.4	80.1	30.5 / 75.4	**90.2 / 91.1**	71.8
	PET	223	79.1	87.2 / 60.2	90.8	67.2	**50.7**	**88.4**	**36.4 / 76.6**	85.4 / 85.9	74.0
	iPET	223	**81.2**	**88.8 / 79.9**	90.8	**70.8**	49.3	**88.4**	31.7 / 74.1	85.4 / 85.9	**75.4**
	SotA	11,000	*91.2*	*93.9 / 96.8*	*94.8*	*92.5*	*76.9*	*93.8*	*88.1 / 63.3*	*94.1 / 93.4*	*89.3*



- It shows that the performance of the GPT is overtaken using PET and iPET, even though there is a huge difference in the number of parameters.

Conclusion

- Contributions
  - PET help leverage the knowledge contained within pretrained language models for downstream tasks.
  - When the initial amount of training data is limited, PET gives large improvements over standard supervised training and strong semi-supervised approaches
  - Achieve few-shot text classification performance similar to GPT-3 on SuperGLUE with LMs that have three orders of magnitude fewer parameters

References

- Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference

https://arxiv.org/abs/2001.07676

- It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners

https://arxiv.org/abs/2009.07118

- Zero-shot Text Classification With Generative Language Models

https://arxiv.org/abs/1912.10165