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

Timo Schick, Hinrich Schutze

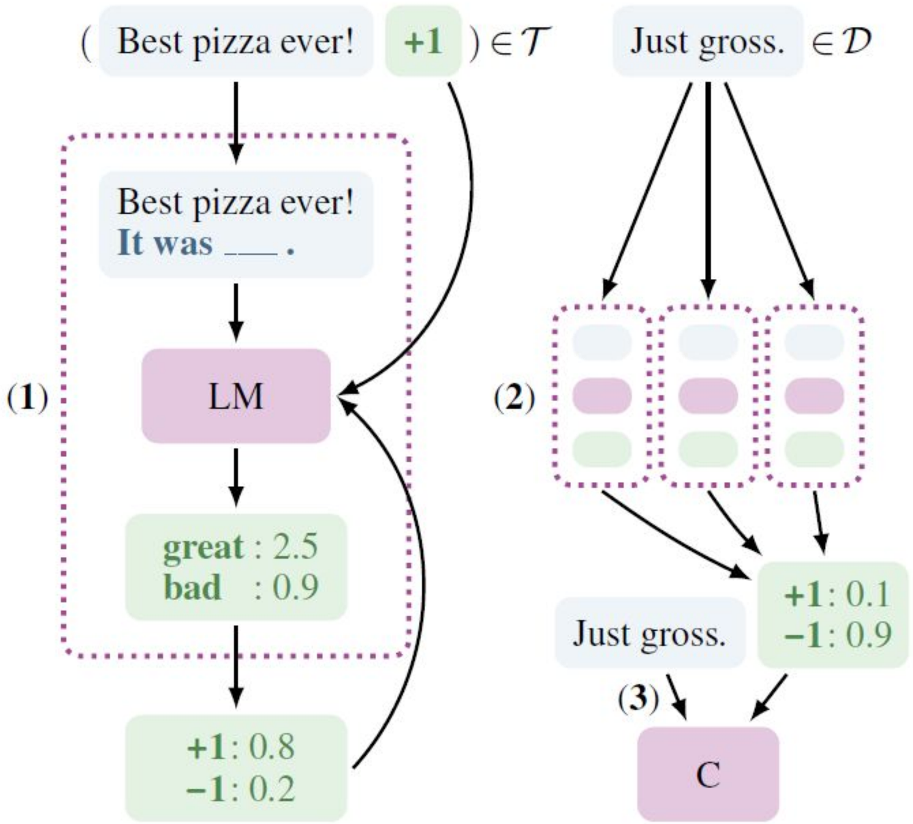
Motivation

Big LMs can perform some tasks with "task description". However, in some case, the performance is not comparable to supervised learning.

In this paper:

Employs patterns in small dataset to annotate soft-label

Train supervised model on that semi-supervised dataset.



Patterns exploiting training

Pattern Verbalize Pair (PVP): using masked LM to predict the labels of the pairs

$$P_1(a) = \text{It was } \_\_\_\_.\ a$$

$$P_2(a) = a.\ \text{All in all, it was } \_\_\_\_.$$

$$P_3(a) = \text{Just } \_\_\_\_!\ \|\ a$$

$$P_4(a) = a\ \|\ \text{In summary, the restaurant is } \_\_\_\_.$$

We define a single verbalizer v for all patterns as

$$v(1) = \text{terrible} \quad v(2) = \text{bad} \quad v(3) = \text{okay}$$
$$v(4) = \text{good} \quad v(5) = \text{great}$$

Yelp review

$$P_1(\mathbf{x}) = \text{“}\ a\ \text{”}\ ?\ \|\ \_\_\_\_\ ,\ \text{“}\ b\ \text{”}$$

$$P_2(\mathbf{x}) = a\ ?\ \|\ \_\_\_\_\ ,\ b$$

and consider two different verbalizers v_1 and v_2 that are defined as follows:

$$v_1(0) = \text{Wrong} \quad v_1(1) = \text{Right} \quad v_1(2) = \text{Maybe}$$
$$v_2(0) = \text{No} \quad v_2(1) = \text{Yes} \quad v_2(2) = \text{Maybe}$$

MNLI

Combining PVPs

PVPs varies that we don’t know which one is good. So

1. For each sample x in a labeled dataset, finetune a LM on x
2. On a unsupervised dataset, combine all finetuned LMs to get a soft label for all samples
3. Then use this scores as training signal to train a supervised model

This might propagate falsely labeled data into the model

Iterative PET (iPET)

Train multiple classifiers on increasing dataset sizes

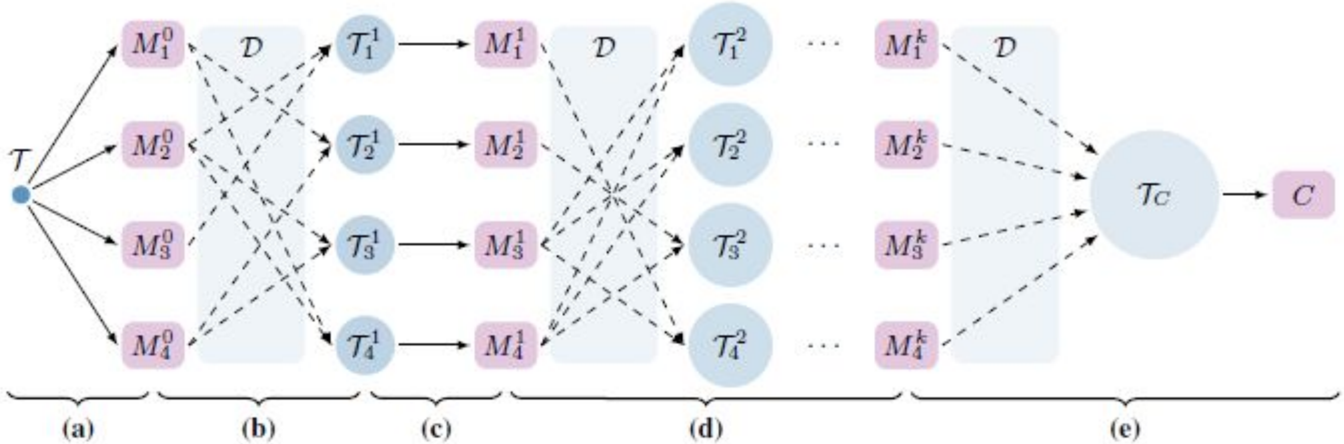


Figure 2: Schematic representation of iPET (a) The initial training set is used to train an ensemble of models as in regular PET. (b) For each model, a random subset of two other models ($\lambda = 2/3$) is used to generate a new training set by labeling examples from $\mathcal{D}$. (c) A new set of PET models is trained using the larger, model-specific datasets. (d) The previous two steps are repeated k times, each time increasing the size of the generated training sets by a factor of d. (e) The set of models at iteration k is used to create a soft-labeled dataset $\mathcal{T}_C$ as in regular PET; the final classifier C is trained on this dataset.

Result

Examples	Training Mode	Yelp	AG's News	Yahoo	MNLI (m)
$\lvert\mathcal{T}\rvert = 0$	unsupervised (avg)	33.8 ± 9.6	69.5 ± 7.2	44.0 ± 9.1	39.1 ± 4.3
	unsupervised (max)	40.8 ± 0.0	79.4 ± 0.0	56.4 ± 0.0	43.8 ± 0.0
	iPET	$\mathbf{56.7 \pm 0.2}$	$\mathbf{87.5 \pm 0.1}$	$\mathbf{70.7 \pm 0.1}$	$\mathbf{53.6 \pm 0.1}$
$\lvert\mathcal{T}\rvert = 10$	supervised	21.1 ± 1.6	25.0 ± 0.1	10.1 ± 0.1	34.2 ± 2.1
	PET	52.9 ± 0.1	87.5 ± 0.0	63.8 ± 0.2	41.8 ± 0.1
	iPET	$\mathbf{57.6 \pm 0.0}$	$\mathbf{89.3 \pm 0.1}$	$\mathbf{70.7 \pm 0.1}$	$\mathbf{43.2 \pm 0.0}$
$\lvert\mathcal{T}\rvert = 50$	supervised	44.8 ± 2.7	82.1 ± 2.5	52.5 ± 3.1	45.6 ± 1.8
	PET	60.0 ± 0.1	86.3 ± 0.0	66.2 ± 0.1	63.9 ± 0.0
	iPET	$\mathbf{60.7 \pm 0.1}$	$\mathbf{88.4 \pm 0.1}$	$\mathbf{69.7 \pm 0.0}$	$\mathbf{67.4 \pm 0.3}$
$\lvert\mathcal{T}\rvert = 100$	supervised	53.0 ± 3.1	86.0 ± 0.7	62.9 ± 0.9	47.9 ± 2.8
	PET	61.9 ± 0.0	88.3 ± 0.1	69.2 ± 0.0	74.7 ± 0.3
	iPET	$\mathbf{62.9 \pm 0.0}$	$\mathbf{89.6 \pm 0.1}$	$\mathbf{71.2 \pm 0.1}$	$\mathbf{78.4 \pm 0.7}$
$\lvert\mathcal{T}\rvert = 1000$	supervised	63.0 ± 0.5	$\mathbf{86.9 \pm 0.4}$	70.5 ± 0.3	73.1 ± 0.2
	PET	$\mathbf{64.8 \pm 0.1}$	$\mathbf{86.9 \pm 0.2}$	$\mathbf{72.7 \pm 0.0}$	$\mathbf{85.3 \pm 0.2}$

Table 1: Results for RoBERTa (large) on Yelp, AG's News, Yahoo and MNLI (matched) for various training set sizes. Scores for PET were obtained using the weighted variant with manually defined verbalizers.