Cold-Start Data Selection for Better Few-shot Language Model Fine-tuning: A Prompt-based Uncertainty Propagation Approach

Source: Acl 2023
Advisor: JIA-LING KOH
Speaker: FAN-CHI-YU
Date:2023/08/07

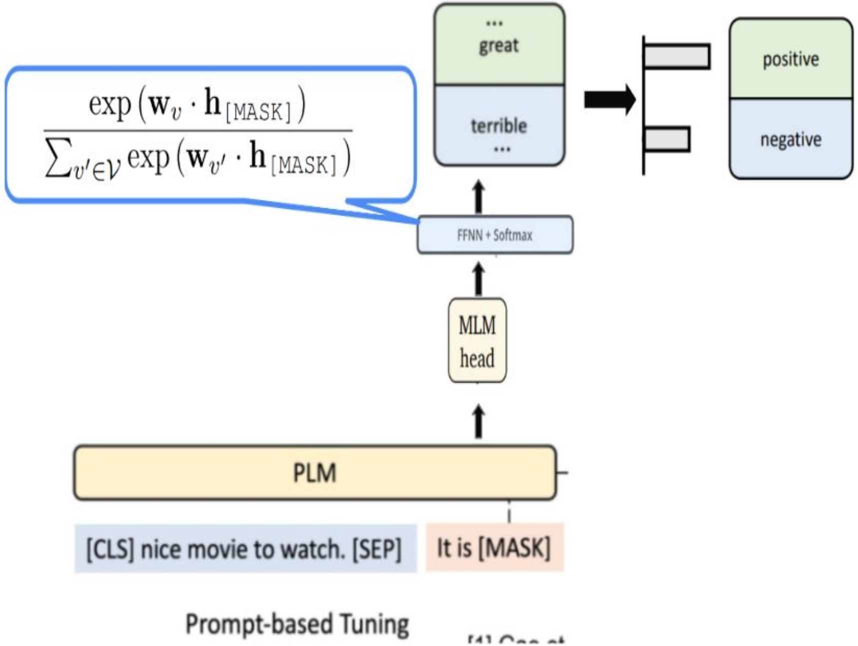
Outline

- Introduction
  - Prompt-base Tuning
  - Active learning
  - Cold Start Data Selection
- Method
  - Uncertainty Estimation with Prompt
  - Uncertainty Propagation for Data Utility Estimation
  - Partition-then-rewrite
- Experiment
  - Baseline
  - Ablation Study
  - label efficiency
- Conclusion

Introduction

Introduction(Prompt-base Tuning)

Choose a label word mapping , which maps task labels to individual words



Input: x = “I love this movie”

Template: [x] Overall, it was a [z] movie

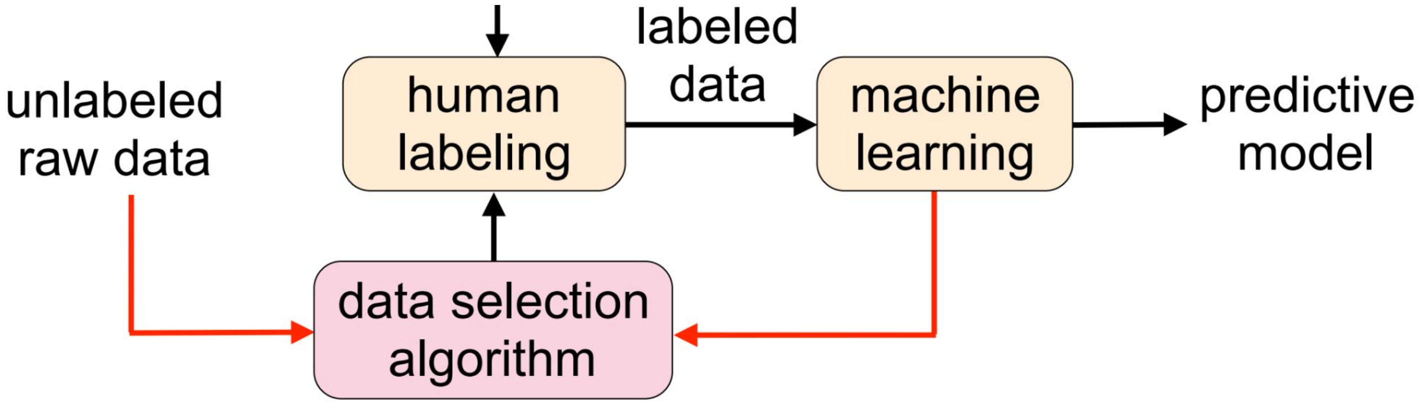
Prompting: x’ = “I love this movie. Overall it was a [z] movie.”

Predicting: x’ = “I love this movie. Overall it was a fantastic movie.”

Mapping: fantastic => Positive

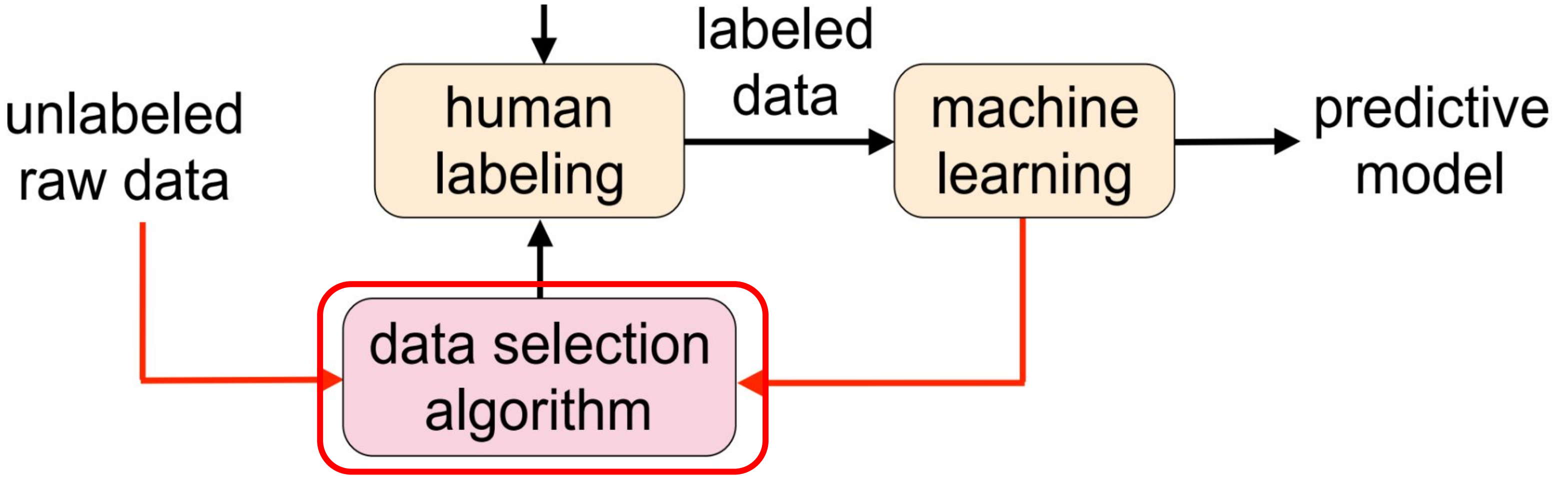
Introduction(Active learning)

- Active learning (AL) aims at reducing labeling effort by identifying the **most valuable unlabeled** data points from a large pool.



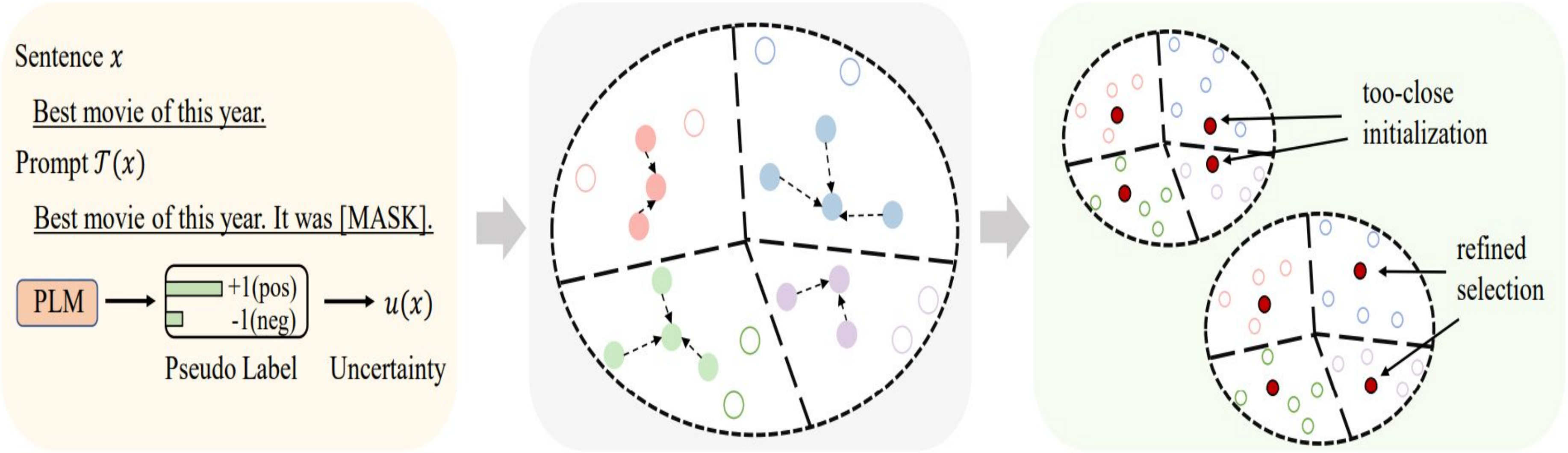
Introduction(Cold Start Data Selection)

We have **only unlabeled data** and **zero initial labels**, and need to design acquisition functions to effectively query samples for PLM fine-tuning



Method

Method



1. Uncertainty Estimation with Prompts
2. Uncertainty Propagation
3. Partition-then-rewrite (PTR)

Method

Algorithm 1: Process of PATRON Strategy.

Input: Unlabeled samples $\mathcal{X}_u$; Pre-trained LM $\mathcal{M} = f(\cdot;\theta)$, number of acquired samples B, the number of iterations T (T=2 in this work).

// **Step 1**: *Uncertainty Propagation for Utility Estimation.*

1a. Calculate uncertainty for samples $x \in \mathcal{X}_u$ with prompts based on Eq. (5).

1b. Estimate uncertainty $\widetilde{u}_{\text{prop}}$ with Eq. (6) and (7).

// **Step 2**: *Predict-then-propagate* (PTR) *for Diversity Promoting Selection.*

2a. Run K-Means on $\mathcal{X}_u$ with k=B until convergence.

2b. Select initial sample set $\mathcal{Q}^{(0)}$ based on Eq. (8).

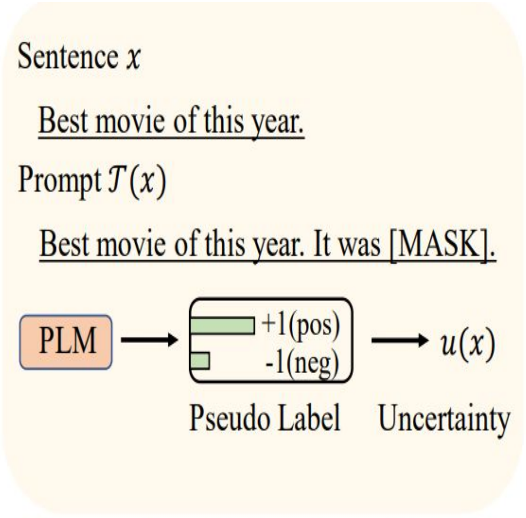
for $t = 1, 2, \cdots, T$ **do**

- **2c**. Building the additional KNN graph to obtain $\mathcal{X}_{\text{c-KNN}}$ with Eq. (9).
- **2d**. Update $\mathcal{Q}^{(t)}$ by optimizing the selected sample within each cluster $\widetilde{q}$ with Eq. (10).

Output: The final selected labeled data $\mathcal{Q}^{(T)}$.

Uncertainty Estimation with Prompt

- PLM probability problematic; tackle via contextualized label word priors calculation.



1. Uncertainty Estimation with Prompts

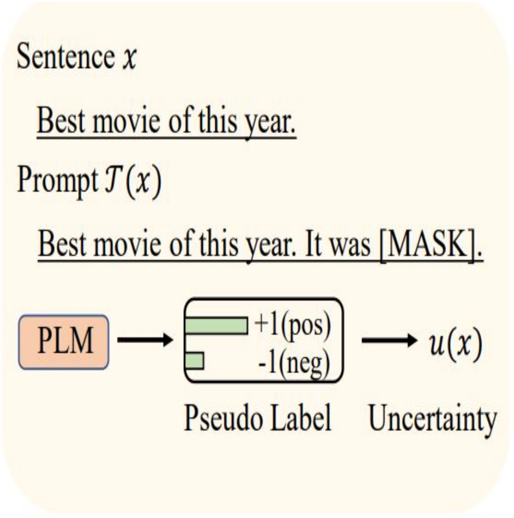
$$
\begin{aligned}
p(y \mid x) &= p\left(\texttt{[MASK]} = \mathcal{V}(y) \mid \mathcal{T}(x)\right) \\
&= \frac{\exp\left(\boldsymbol{w}_{\mathcal{V}(y)}^{T} \boldsymbol{h}_{\texttt{[MASK]}}\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\boldsymbol{w}_{\mathcal{V}(y')}^{T} \boldsymbol{h}_{\texttt{[MASK]}}\right)}
\end{aligned} \tag{1}
$$

$$
\mathcal{S} = \bigcup_{i \in \{1,2,\ldots,c\}} \underset{x \in \mathcal{D}_u}{\text{Top-k}}\; p(y_i \mid x). \tag{2}
$$

$$
P(v) \approx \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} P_{\mathcal{M}}\left(\texttt{[MASK]} = v \mid x\right), \tag{3}
$$

Uncertainty Estimation with Prompt

- PLM probability problematic; tackle via contextualized label word priors calculation.

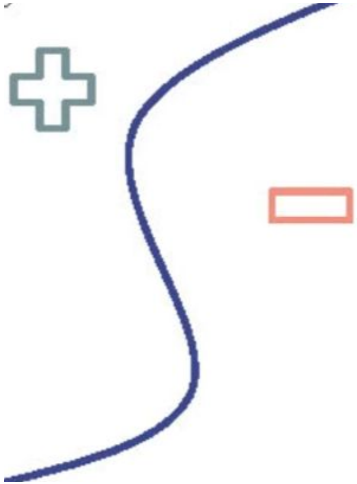


1. Uncertainty Estimation with Prompts

	Top-k $p(y_i|x)$		

Uncertainty Estimation with Prompt

When sample selection yields suboptimal results, calibration is used to improve the pseudo labels



$$\widehat{y_i} = \left(\frac{p(y_i|x)}{P(\mathcal{V}(y_i))} \right) / \left(\sum_{j=1}^{C} \frac{p(y_j|x)}{P(\mathcal{V}(y_j))} \right). \quad (4)$$

contextualized priors

$$u(x) = -\sum_{i=1}^{C} \widehat{y_i} \log \widehat{y_i}. \quad (5)$$

Method

Algorithm 1: Process of PATRON Strategy.

Input: Unlabeled samples $\mathcal{X}_u$; Pre-trained LM $\mathcal{M} = f(\cdot;\theta)$, number of acquired samples B, the number of iterations T (T=2 in this work).

// **Step 1**: *Uncertainty Propagation for Utility Estimation.*

1a. Calculate uncertainty for samples $x \in \mathcal{X}_u$ with prompts based on Eq. (5).

1b. Estimate uncertainty $\widehat{u}_{\text{prop}}$ with Eq. (6) and (7).

// **Step 2**: *Predict-then-propagate (PTR) for Diversity Promoting Selection.*

2a. Run K-Means on $\mathcal{X}_u$ with k=B until convergence.

2b. Select initial sample set $\mathcal{Q}^{(0)}$ based on Eq. (8).

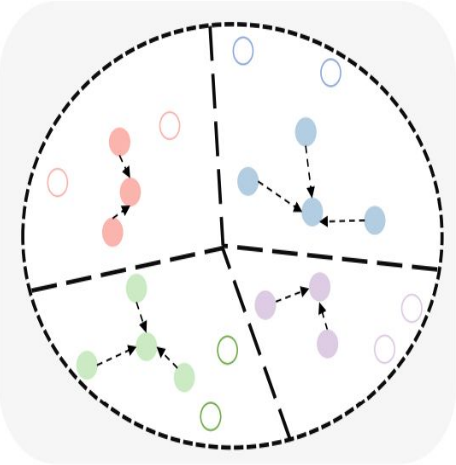
for $t = 1, 2, \cdots, T$ **do**

- **2c**. Building the additional KNN graph to obtain $\mathcal{X}_{\text{c-KNN}}$ with Eq. (9).
- **2d**. Update $\mathcal{Q}^{(t)}$ by optimizing the selected sample within each cluster $\widetilde{q}$ with Eq. (10).

Output: The final selected labeled data $\mathcal{Q}^{(T)}$.

Method:
Uncertainty Propagation for Data Utility Estimation

Result in **higher propagated uncertainty**, indicating the PLMs are uncertain about the **surrounding regions** around the sample.



2. Uncertainty Propagation

$$\kappa\left(x_i, x_j\right) = \exp\left(-\rho \left\|\mathbf{z}_i - \mathbf{z}_j\right\|_2^2\right), \qquad (6)$$

$$\widehat{u}_{\text{prop}}(x) = u(x) + \frac{\sum_{x_i \in \mathcal{X}_{\text{KNN}}(x)} \kappa(x, x_i) \cdot u(x_i)}{\left|\mathcal{X}_{\text{KNN}}(x)\right|}. \qquad (7)$$

Method

Algorithm 1: Process of PATRON Strategy.

Input: Unlabeled samples $\mathcal{X}_u$; Pre-trained LM $\mathcal{M} = f(\cdot;\theta)$, number of acquired samples B, the number of iterations T (T=2 in this work).

// **Step 1**: *Uncertainty Propagation for Utility Estimation.*

1a. Calculate uncertainty for samples $x \in \mathcal{X}_u$ with prompts based on Eq. (5).

1b. Estimate uncertainty $\widehat{u}_{\text{prop}}$ with Eq. (6) and (7).

// **Step 2**: *Predict-then-propagate (*PTR*) for Diversity Promoting Selection.*

2a. Run K-Means on $\mathcal{X}_u$ with k=B until convergence.

2b. Select initial sample set $\mathcal{Q}^{(0)}$ based on Eq. (8).

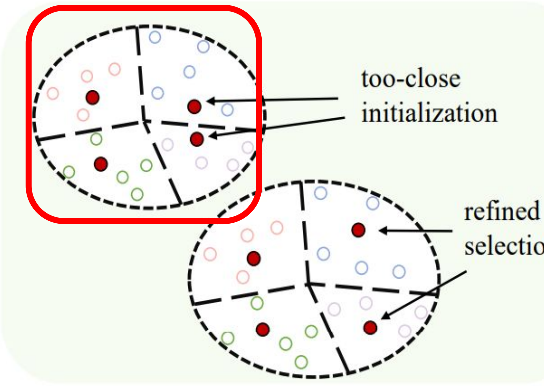
for $t = 1, 2, \cdots, T$ **do**

- **2c**. Building the additional KNN graph to obtain $\mathcal{X}_{\text{c-KNN}}$ with Eq. (9).
- **2d**. Update $\mathcal{Q}^{(t)}$ by optimizing the selected sample within each cluster $\widetilde{q}$ with Eq. (10).

Output: The final selected labeled data $\mathcal{Q}^{(T)}$.

Method: Partition-then-rewrite(PTR)

- Diversity-Promoting Data Selection
- **K-Means** clustering partitions pool Du into diverse clusters based on embeddings.



3. Partition-then-rewrite (PTR)

$$\bar{\mathbf{z}}_i = \frac{1}{|C_i|} \sum_{x_j \in C_i}$$

$$q_i = \operatorname*{argmax}_{x_j \in C_i} \left(\hat{u}_{\text{prop}}(x_j) - \beta \left\| \mathbf{z}_j - \bar{\mathbf{z}}_i \right\|_2^2 \right), \quad (8)$$

Method

Algorithm 1: Process of PATRON Strategy.

Input: Unlabeled samples $\mathcal{X}_u$; Pre-trained LM $\mathcal{M} = f(\cdot;\theta)$, number of acquired samples B, the number of iterations T (T=2 in this work).

// **Step 1**: *Uncertainty Propagation for Utility Estimation.*

1a. Calculate uncertainty for samples $x \in \mathcal{X}_u$ with prompts based on Eq. (5).

1b. Estimate uncertainty $\widehat{u}_{\text{prop}}$ with Eq. (6) and (7).

// **Step 2**: *Predict-then-propagate (*PTR*) for Diversity Promoting Selection.*

2a. Run K-Means on $\mathcal{X}_u$ with k=B until convergence.

2b. Select initial sample set $\mathcal{Q}^{(0)}$ based on Eq. (8).

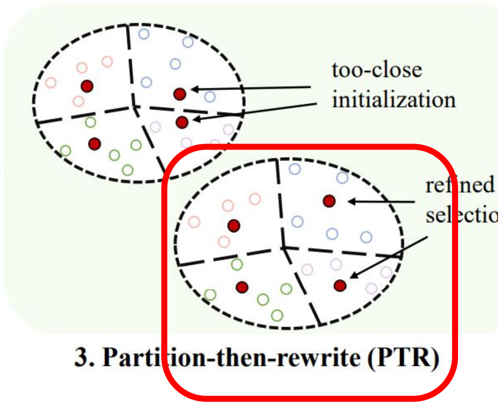
for $t = 1, 2, \cdots, T$ **do**

- **2c**. Building the additional KNN graph to obtain $\mathcal{X}_{\text{c-KNN}}$ with Eq. (9).
- **2d**. Update $\mathcal{Q}^{(t)}$ by optimizing the selected sample within each cluster $\widetilde{q}$ with Eq. (10).

Output: The final selected labeled data $\mathcal{Q}^{(T)}$.

Method: Partition-then-rewrite (PTR)

- Samples can still be very close to other selected samples in adjacent clusters, leading to limited overall diversity.
- Prevent samples in adjacency clusters from being overly close.



3. Partition-then-rewrite (PTR)

$$\mathcal{X}_{\text{c-KNN},i} = \text{KNN}(q_i, \mathcal{Q}). \tag{9}$$

$$\begin{aligned}\widetilde{q}_i = \underset{x_j \in \mathcal{C}_i}{\operatorname{argmax}} \Big(&\widehat{u}_{\text{prop}}(x_j) - \beta \|\mathbf{z}_j - \bar{\mathbf{z}}_i\|_2 \\ &- \gamma \sum_{q_k \in \mathcal{X}_{\text{c-knn},i}} [m - \|\mathbf{z}_j - \mathbf{z}_k\|_2]_+ \Big),\end{aligned} \tag{10}$$

Experiment

Experiment (Dataset)

Dataset	Domain	Classes c	#Unlabeled	#Test	Type	Template	Label words
IMDB	Movie Review	2	25k	25k	sentiment	$\langle S \rangle$. It was [MASK].	terrible, great
Yelp-full	Restaurant Review	2	560k	38k	sentiment	$\langle S \rangle$. It was [MASK].	terrible, bad, okay, good, great
AG News	News	4	120k	7.6k	News Topic	[MASK] News: $\langle S \rangle$	World, Sports, Business, Tech
Yahoo! Answers	Web QA	10	300k	60k	QA Topic	[Category: [MASK]] $\langle S \rangle$	Society, Science, Health, Education, Computer, Sports, Business, Entertainment, Relationship, Politics
DBPedia	Wikipedia Text	14	420k	70k	Wikipedia Topic	$\langle T \rangle \langle S \rangle . \langle T \rangle$ is a [MASK]]	Company, School, Artist, Athlete, Politics, Transportation, Building, Mountain, Village, Animal, Plant, Album, Film, Book
TREC	Web Text	6	5k	0.6k	Question Topic	$\langle S \rangle$. It was [MASK].	Expression, Entity, Description, Human, Location, Number

Experiment (Baseline : Uncertainty-based)

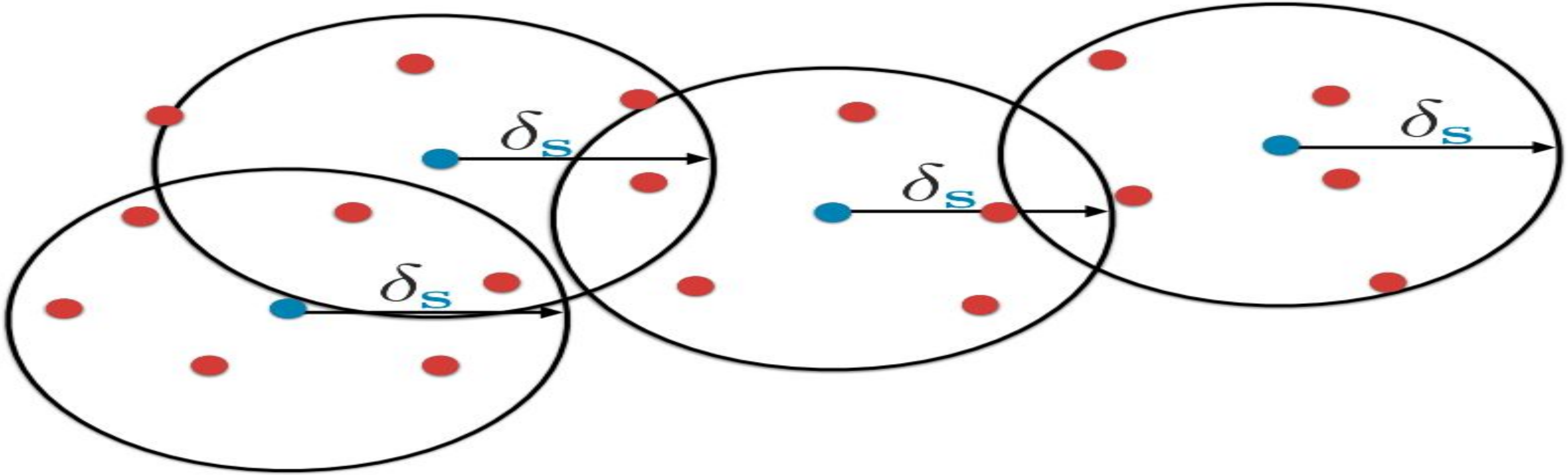
Methods focus on choosing **hard samples** without considering the sample diversity, **leading to imbalanced** label distribution

- Uncertainty : Use the highest uncertainty by entropy
- CAL : KL div Predict the prediction of itself and its neighbors

Experiment (Baseline : Diversity-based)

Tend to select **diverse yet** easy examples for the model

- **Coreset** : Samples largest distance between a data point and its nearest center is minimized.
-



Experiment (Baseline : Diversity-based)

Tend to select **diverse yet** easy examples for the model

- **Coreset** : Samples largest distance between a data point and its nearest center is minimized.
- **BERT-KM** : Each cluster that is closest to the center of the cluster
- **Margin-KM** : Minimum margin between the two most likely probabilities from each cluster
- **ALPS** : Uses the masked language model (MLM) loss of BERT to generate surprisal embeddings to query samples.
- **TPC** : Calculates the density for each data point, and then selects those with the highest density from each cluster

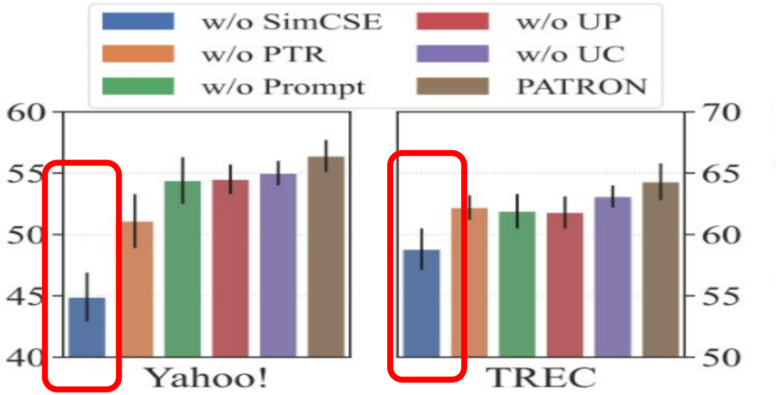
Experiment (Baseline)

Task	c	$\|B\|$	Random	Uncertainty-based: Uncertainty	CAL	Diversity-based: BERT-KM	Coreset	Margin-KM	ALPS	TPC	PATRON (Ours)
Yahoo! Ans.	10	32	58.5 ± 4.0	55.0 ± 3.0	54.0 ± 1.5	61.4 ± 1.8	55.3 ± 2.1	57.8 ± 2.6	61.9 ± 0.9	57.0 ± 1.6	$\mathbf{63.2 \pm 1.2}^{*}$
		64	62.2 ± 1.0	60.4 ± 0.7	58.6 ± 1.3	62.8 ± 0.7	59.5 ± 0.7	58.8 ± 1.2	$\underline{63.3 \pm 0.8}$	60.8 ± 0.7	$\mathbf{66.2 \pm 0.3}^{**}$
		128	64.7 ± 1.3	63.0 ± 1.2	60.1 ± 1.8	65.4 ± 1.2	62.7 ± 1.0	65.4 ± 0.7	65.9 ± 0.7	$\underline{66.2 \pm 0.6}$	$\mathbf{67.6 \pm 0.5}^{**}$
TREC	6	32	69.4 ± 2.8	66.4 ± 3.5	41.6 ± 2.5	68.1 ± 2.3	61.0 ± 4.6	64.8 ± 2.7	$\underline{72.1 \pm 2.3}$	59.5 ± 3.3	$\mathbf{76.1 \pm 1.1}^{**}$
		64	75.4 ± 1.4	68.0 ± 2.3	49.8 ± 1.5	78.8 ± 2.0	78.6 ± 1.3	74.2 ± 1.4	$\underline{80.6 \pm 0.9}$	77.8 ± 1.5	$\mathbf{81.9 \pm 1.3}^{*}$
		128	85.0 ± 2.1	78.8 ± 2.0	67.2 ± 2.7	85.6 ± 1.8	84.2 ± 2.4	78.0 ± 1.9	$\underline{86.5 \pm 2.0}$	80.6 ± 1.4	$\mathbf{88.9 \pm 1.0}^{**}$

Diversity-based methods generally achieve **better performance** over the uncertainty-based strategies

Experiment (Ablation Study)

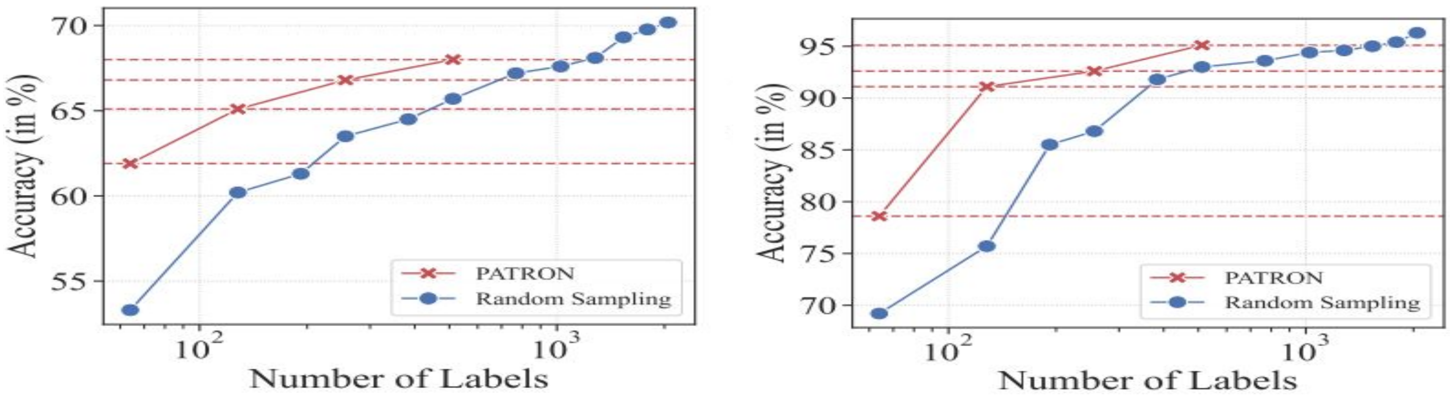
The **SimCSE** embeddings with the prompt-based pseudo labels and improve the performance significantly.



(a) Ablation Study

Experiment (label efficiency)

- PATRON improves the label efficiency over baselines by **3.4%–6.9%** on average.
- With 512 labels as the budget, PATRON achieves better performance with 2X~2.5 labels(**Yahoo1280 labels**,**TREC 1024 labels**)



(d) Yahoo! (f) TREC.

Conclusion

- By leveraging prompts, we can **distill the task-specific knowledge** from the frozen PLM to guide data acquisition.
- It ‘s possible to extend our method to (**PET,LMBFF**)tasks.
- This paper achieve sample **representativeness** and **diversity**.