Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification

Method

Task

Source: Acl 2022
Advisor: JIA-LING KOH
Speaker: FAN-CHI-YU
Date:2023/12/12

Outline

- Introduction
- Method
- Experiment
- Conclusion

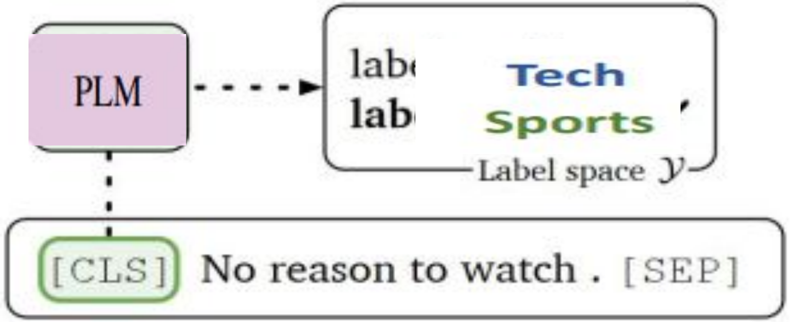
Introduction

Introduction(Fine-tuning)

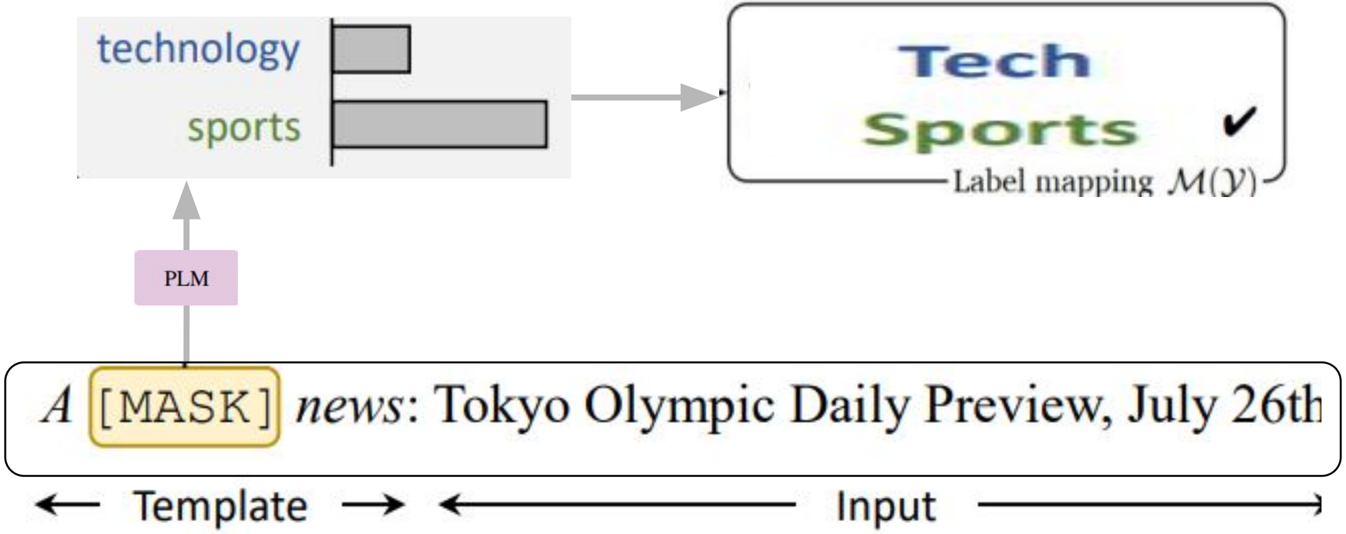
classifier

$$P(\cdot|x) = \mathrm{Softmax}(\boxed{F}(\mathbf{h}_{\texttt{[CLS]}})). \qquad (1)$$

The classifier and PLM are tuned by maximizing $\frac{1}{N}\sum_{i=1}^{N} \log P(y_i|x_i)$



Introduction(Prompt Tuning)

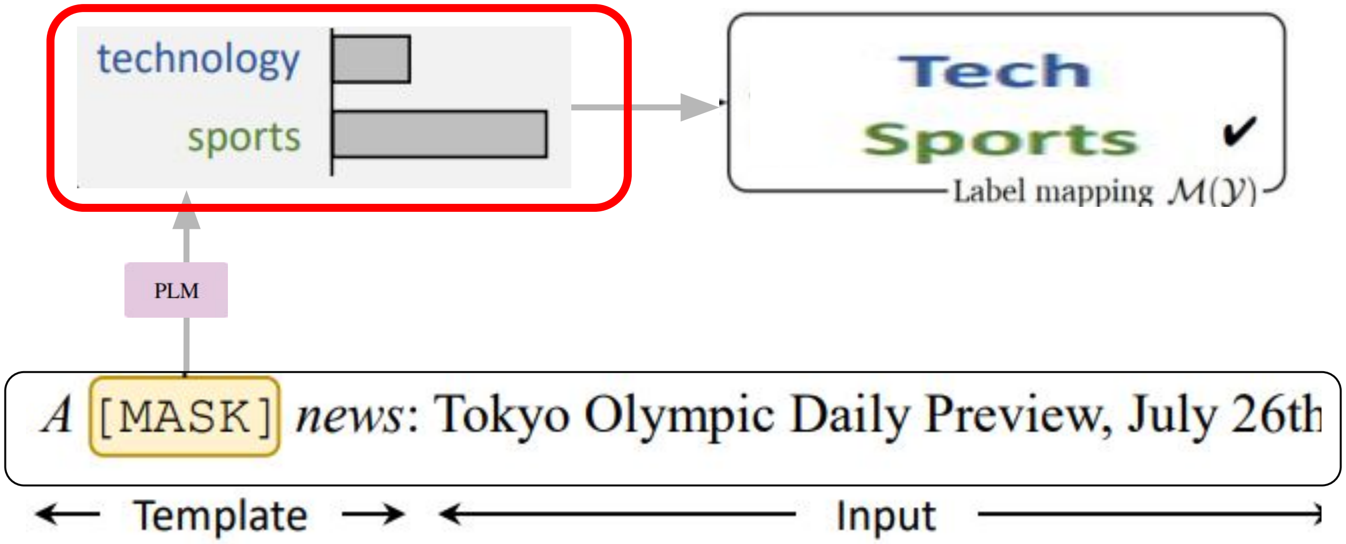


$$P(y|\mathbf{x}_\mathrm{p})=g\left(P_\mathcal{M}(\texttt{[MASK]}=v|\mathbf{x}_\mathrm{p})|v \in \mathcal{V}_y\right), \quad (1)$$

[Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference](#)

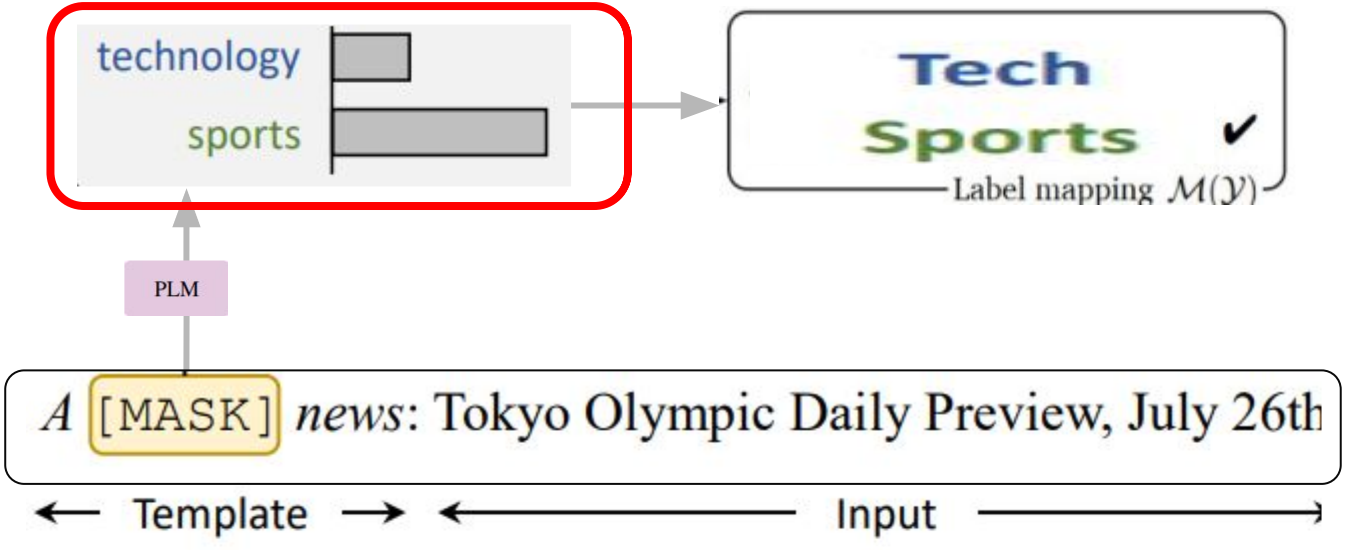
Introduction(Manual Verbalizer)

Defined by human with domain knowledge



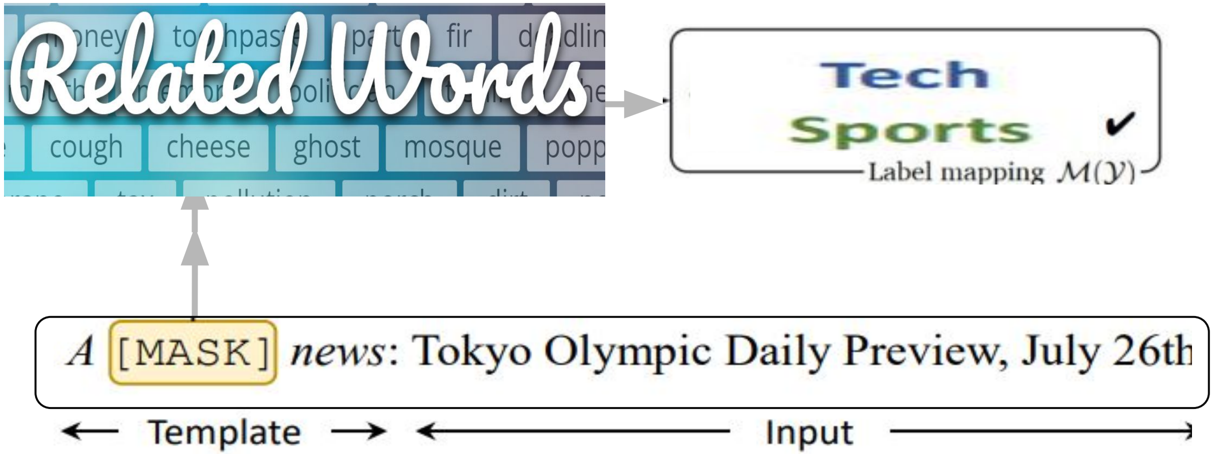
Introduction(Manual Verbalizer)

Because past handcraft have may lack coverage and hight variance to the result .



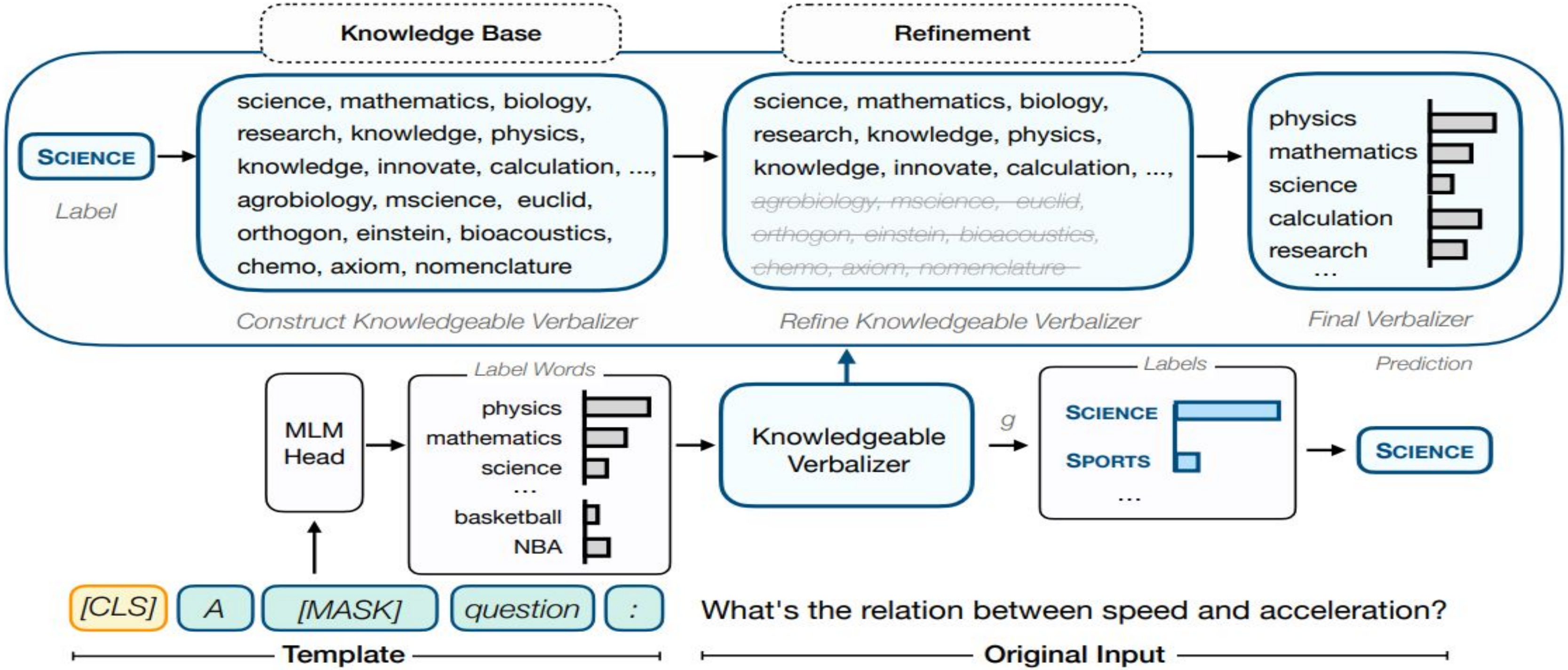
Introduction(Knowledgeable prompt-tuning)

We focus on **incorporating** external knowledge into the verbalizer



Method

Method



Knowledge Base
Refinement
SCIENCE
Label
science, mathematics, biology, research, knowledge, physics, knowledge, innovate, calculation, ..., agrobiology, mscience, euclid, orthogon, einstein, bioacoustics, chemo, axiom, nomenclature
science, mathematics, biology, research, knowledge, physics, knowledge, innovate, calculation, ..., ~~agrobiology, mscience, euclid, orthogon, einstein, bioacoustics, chemo, axiom, nomenclature~~
physics
mathematics
science
calculation
research
...
Construct Knowledgeable Verbalizer
Refine Knowledgeable Verbalizer
Final Verbalizer
MLM Head
Label Words
physics
mathematics
science
...
basketball
NBA
Knowledgeable Verbalizer
g
Labels
SCIENCE
SPORTS
...
Prediction
SCIENCE
[CLS]
A
[MASK]
question
:
What's the relation between speed and acceleration?
Template
Original Input

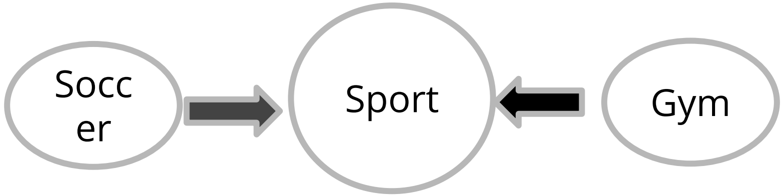
Verbalizer Construction

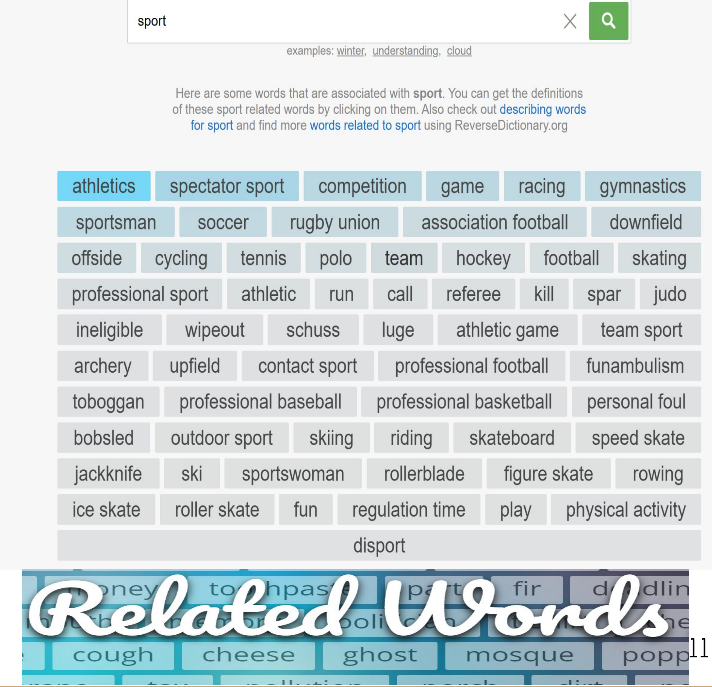
There is no standard correct answer, but **abundant** words may fit this context.

Neighborhood nodes

$$\mathcal{V}_y = \boxed{N_G}(v_0) \cup \{\boxed{v_0}\}$$

label word set





Verbalizer Refine

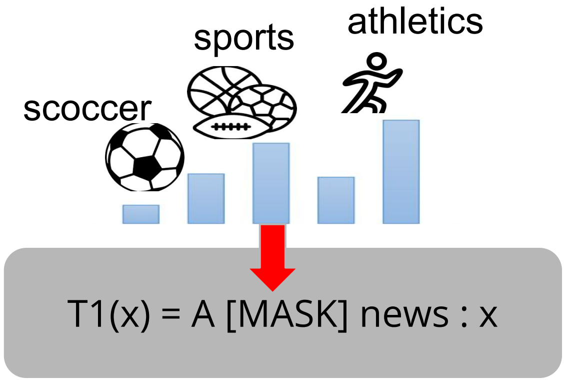
Refine verbalizer, keep high-quality words, reduce noise.

1. Frequency Refinement
2. Relevance Refinement
3. Contextualized Calibration
4. Learnable Refinement

Frequency Refinement

We propose to use **contextualized prior** of the label words to remove these words

$$P_{\mathcal{D}}(v) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} P_{\mathcal{M}}(\texttt{[MASK]} = v | \mathbf{x}_{\mathrm{p}}). \quad (2)$$



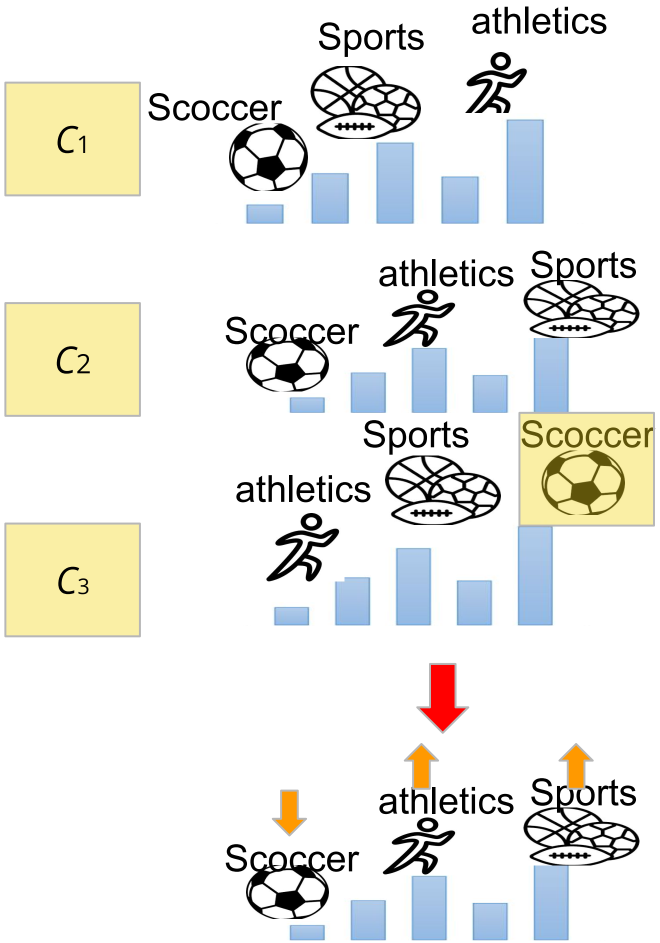
Frequency Refinement

We propose to use **contextualized prior** of the label words to remove these words

$$P_{\mathcal{D}}(v) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}} P_{\mathcal{M}}(\texttt{[MASK]}=v|\mathbf{x}_{\mathrm{p}}). \qquad (2)$$

$$P_{\mathcal{D}}(v) \approx \frac{1}{\boxed{|\tilde{\mathcal{C}}|}} \sum_{\mathbf{x}\in\tilde{\mathcal{C}}} P_{\mathcal{M}}(\texttt{[MASK]}=v|\mathbf{x}_{\mathrm{p}}). \qquad (3)$$

unlabeled support set *C*



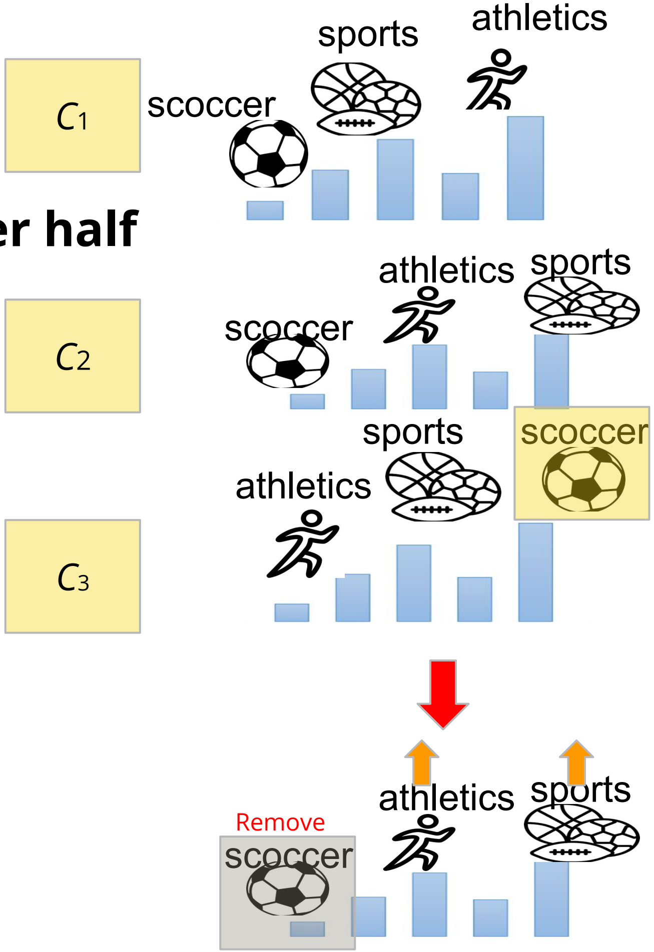
Frequency Refinement

We filter the label words that appear in the **lower half** of the **contextualized prior** probability.

$$P_{\mathcal{D}}(v)=\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}P_{\mathcal{M}}(\texttt{[MASK]}=v|\mathbf{x}_{\mathrm{p}}). \qquad (2)$$

$$P_{\mathcal{D}}(v)\approx\frac{1}{|\tilde{C}|}\sum_{\mathbf{x}\in\tilde{C}}P_{\mathcal{M}}(\texttt{[MASK]}=v|\mathbf{x}_{\mathrm{p}}). \qquad (3)$$

unlabeled support set *C*

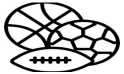


Relevance Refinement

Label words may be more relevant to their **belonging class** than the **others**

athletics

SPORTS



label word

Class:**SPORTS**

$$r(v, y) = \cos(\mathbf{q}^v, \mathbf{q}^y) = \cos(\mathbf{q}^v, \mathbf{q}^{v_0}). \quad (5)$$

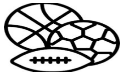
class's representation

Relevance Refinement

Label words may be more relevant to their **belonging class** than the **others**

athletics

SPORTS



label word

Class:**SPORTS**

$$r(v, y) = \cos(\mathbf{q}^{v}, \mathbf{q}^{y}) = \cos(\mathbf{q}^{v}, \mathbf{q}^{v_0}). \qquad (5)$$

class's representation

$$R(v) = r(v, f(v)) \frac{|\mathcal{Y}| - 1}{\sum_{y \in \mathcal{Y}, y \neq f(v)} (r(v, y))}, \qquad (6)$$

average relevance score for the **other classes**

Relevance Refinement

Label words may be more relevant to their **belonging class** than the **others**

belonging class

$$r(v, y) = \cos(\mathbf{q}^v, \mathbf{q}^y) = \cos(\mathbf{q}^v, \mathbf{q}^{v_0}). \quad (5)$$

$$R(v) = r(v, f(v)) \frac{|\mathcal{Y}| - 1}{\sum_{y \in \mathcal{Y}, y \neq f(v)} (r(v, y))}, \quad (6)$$

average relevance score for the **other classes**

Relevance Refinement

Label words may be more relevant to their **belonging class** than the **others**

If R(V)<1 we Remove it

belonging class

$$r(v, y) = \cos(\mathbf{q}^{v}, \mathbf{q}^{y}) = \cos(\mathbf{q}^{v}, \mathbf{q}^{v_0}). \qquad (5)$$

others classes

relevance score

$$R(v) = r(v, f(v)) \frac{\sum_{y \in \mathcal{Y}, y \neq f(v)} (r(v, y))}{|\mathcal{Y}| - 1}, \qquad (6)$$

class set

belonging class

average relevance score for the **other classes**

Contextualized Calibration

Before use the contextualized calibration (**CC**),KB tend to have more **diverse prior** probabilities(less likely to be predicted than the others).

$$\tilde{P}_{\mathcal{M}}([\text{MASK}]=v|\mathbf{x}_\mathrm{p}) \propto \frac{P_{\mathcal{M}}([\text{MASK}]=v|\mathbf{x}_\mathrm{p})}{\boxed{P_{\mathcal{D}}(v)}} \quad (7)$$

prior probability of the label word.

Learnable Refinement

In **few-shot learning**,the refinement can be strengthen by a learning process.

each label word

$$\alpha_v = \frac{\exp(w_v)}{\sum_{u \in \mathcal{V}_y} \exp(w_u)}. \qquad (8)$$

belonging class

Verbalizer Utilization

Mapping the predicted probability on each **refined label** word to the decision of the class label y

Average

$$\hat{y} = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \frac{\sum_{v \in \mathcal{V}_y} \tilde{P}_{\mathcal{M}}(\texttt{[MASK]} = v|\mathbf{x}_\mathrm{p})}{|\mathcal{V}_y|}. \quad (9)$$

Weighted Average

$$\hat{y} = \operatorname{argmax}_{y \in \mathcal{Y}} \frac{\exp\left(s(y|\mathbf{x}_\mathrm{p})\right)}{\sum_{y'} \exp\left(s(y'|\mathbf{x}_\mathrm{p})\right)}, \quad (10)$$

Weighted Average

$$s(y|\mathbf{x}_\mathrm{p}) = \sum_{v \in \mathcal{V}_y} \alpha_v \log P_{\mathcal{M}}(\texttt{[MASK]} = v|\mathbf{x}_\mathrm{p}). \quad (11)$$

Verbalizer Utilization(Average)

Each label word of a class **contributes equally** to predicting the label.

$$\hat{y} = \underset{y \in \mathcal{Y}}{\mathrm{argmax}} \frac{\sum_{v \in \mathcal{V}_y} \tilde{P}_{\mathcal{M}}(\texttt{[MASK]} = v \mid \mathbf{x}_{\mathrm{p}})}{|\mathcal{V}_y|}. \quad (9)$$

class

Verbalizer Utilization(Weighted Average)

Adopt a **weighted average** of label words' scores as the prediction score

Refinement Weights

$$s(y|\mathbf{x}_\mathrm{p}) = \sum_{v \in \mathcal{V}_y} \boxed{\alpha_v} \log P_{\mathcal{M}}(\texttt{[MASK]} = v|\mathbf{x}_\mathrm{p}). \quad (11)$$

$$\hat{y} = \mathrm{argmax}_{y \in \mathcal{Y}} \frac{\exp\left(s(y|\mathbf{x}_\mathrm{p})\right)}{\sum_{y'} \exp\left(s(y'|\mathbf{x}_\mathrm{p})\right)}, \quad (10)$$

Experiment

DataSet

Topic Classification

$\mathcal{T}_1(x) = \text{A } \texttt{[MASK]} \text{ news: } x$

$\mathcal{T}_2(x) = x \text{ This topic is about } \texttt{[MASK]}.$

$\mathcal{T}_3(x) = [\text{ Category : } \texttt{[MASK]} \text{] } x$

$\mathcal{T}_4(x) = [\text{ Topic : } \texttt{[MASK]} \text{] } x$

Name	Type	# Class	Test Size
AG's News	Topic	4	7600
DBPedia	Topic	14	70000
Yahoo	Topic	10	60000
Amazon	Sentiment	2	10000
IMDB	Sentiment	2	25000

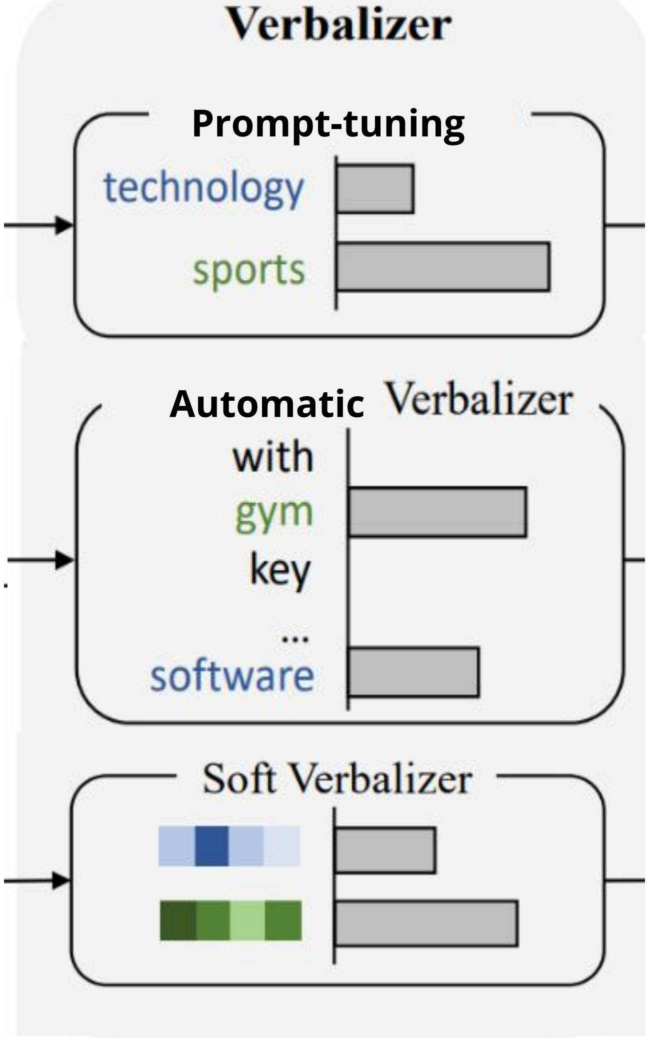
Baselines

1. Prompt-tuning(**PT**)
   a. Uses the class name

      as the only label word for each class

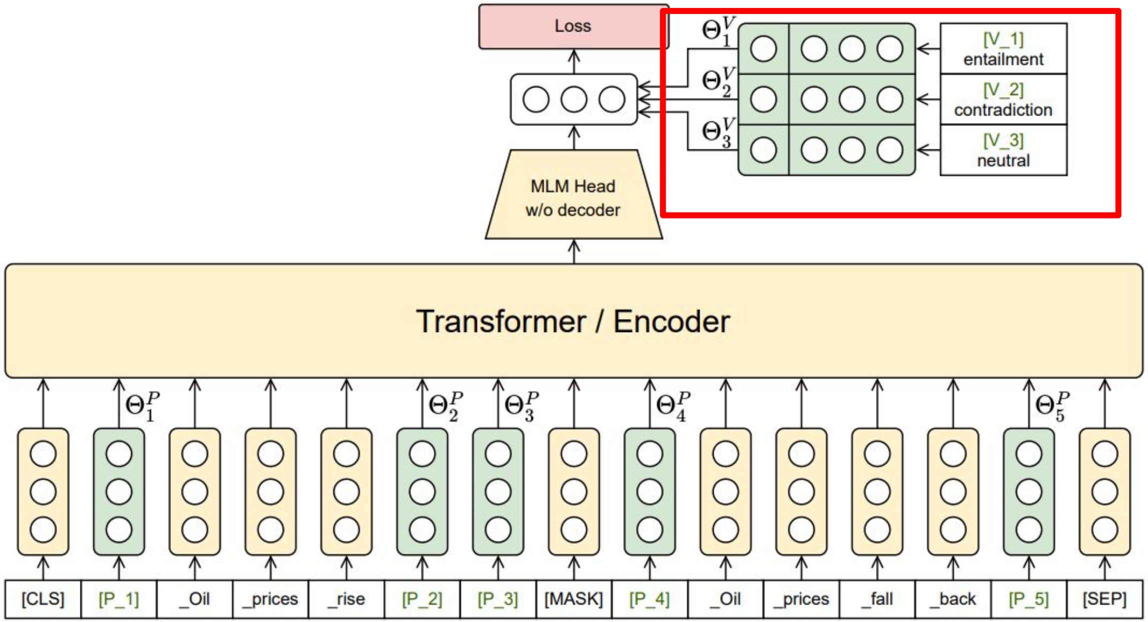
2. Automatic Verbalizer(**AUTO**)
   a. Select the most informative label word

3. Soft Verbalizer(**SOFT**)



Soft Verbalizer(WARP)

Learn the **Verbalizer** in Embedding method



Zero Shot baseline

Method	AG's News	DBPedia	Yahoo	Amazon	IMDB accuracy
PT	75.1 ± 6.2 (79.0)	66.6 ± 2.3 (68.4)	45.4 ± 7.0 (52.0)	80.2 ± 8.8 (87.8)	86.4 ± 4.0 (92.0)
PT+CC	79.9 ± 0.7 (81.0)	73.9 ± 4.9 (82.6)	58.0 ± 1.4 (58.8)	91.4 ± 1.6 (93.5)	91.6 ± 3.0 (93.7)
KPT	**84.8 ± 1.2 (86.7)**	**82.2 ± 5.4 (87.4)**	**61.6 ± 2.2 (63.8)**	**92.8 ± 1.2 (94.6)**	**91.6 ± 2.7** (94.0)
-FR	82.7 ± 1.5 (85.0)	81.8 ± 4.6 (86.2)	60.9 ± 1.5 (62.7)	**92.8 ± 1.2 (94.6)**	91.6 ± 2.8 **(94.1)**
-RR	81.4 ± 1.5 (83.7)	81.4 ± 4.5 (85.8)	60.1 ± 1.0 (61.4)	**92.8 ± 1.2 (94.6)**	91.6 ± 2.8 **(94.1)**
-CC	55.5 ± 2.8 (58.3)	64.5 ± 6.8 (73.0)	42.4 ± 5.0 (46.8)	86.2 ± 5.7 (92.5)	90.3 ± 2.8 **(94.1)**

Few Shot baseline

Shot	Method	AG's News	DBPedia	Yahoo	Amazon	IMDB **accuracy**
1	FT	19.8 ± 10.4	8.6 ± 4.5	11.1 ± 4.0	49.9 ± 0.2	50.0 ± 0.0
	PT	80.0 ± 6.0 (84.4)	92.2 ± 2.5 (94.3)	54.2 ± 3.1 (55.7)	91.9 ± 2.7 (93.2)	91.2 ± 3.7 (93.7)
	AUTO	52.8 ± 9.8 (57.6)	63.0 ± 8.9 (68.3)	23.3 ± 4.5 (25.0)	66.6 ± 12.5 (72.7)	75.5 ± 15.5 (83.1)
	SOFT	80.0 ± 5.6 (82.4)	92.3 ± 2.3 (93.3)	54.3 ± 2.7 (55.9)	90.9 ± 5.8 (93.6)	89.4 ± 8.9 (93.1)
	KPT	**83.7 ± 3.5 (84.6)**	**93.7 ± 1.8 (95.3)**	**63.2 ± 2.5 (64.1)**	93.2 ± 1.3 **(93.9)**	92.2 ± 3.0 **(93.6)**
5	FT	37.9 ± 10.0	95.8 ± 1.3	25.3 ± 14.2	52.1 ± 1.3	51.4 ± 1.4
	PT	82.7 ± 2.7 (84.0)	97.0 ± 0.6 **(97.3)**	62.4 ± 1.7 (63.9)	92.2 ± 3.3 (93.5)	91.9 ± 3.1 (92.7)
	AUTO	72.2 ± 10.1 (75.6)	88.8 ± 3.9 (91.5)	49.6 ± 4.3 (51.2)	87.5 ± 7.4 (90.8)	86.8 ± 10.1 (92.1)
	SOFT	82.8 ± 2.7 (84.3)	97.0 ± 0.6 (97.2)	61.8 ± 1.8 (63.1)	93.2 ± 1.6 **(94.2)**	91.6 ± 3.4 **(93.9)**
	KPT	85.0 ± 1.2 **(85.9)**	97.1 ± 0.4 (97.3)	**67.2 ± 0.8 (67.8)**	93.4 ± 1.9 (94.1)	92.7 ± 1.5 (92.9)
10	FT	75.9 ± 8.4	93.8 ± 2.2	43.8 ± 17.9	83.0 ± 7.0	76.2 ± 8.7
	PT	84.9 ± 2.4 (86.1)	97.6 ± 0.4 (97.8)	64.3 ± 2.2 (64.8)	93.9 ± 1.3 (94.6)	**93.0 ± 1.7 (94.0)**
	AUTO	81.4 ± 3.8 (84.1)	91.5 ± 3.4 (95.1)	58.7 ± 3.1 (60.9)	93.7 ± 1.2 (94.5)	91.1 ± 5.1 (93.3)
	SOFT	85.0 ± 2.8 (86.7)	97.6 ± 0.4 (97.8)	64.5 ± 2.2 (65.0)	93.9 ± 1.7 (93.9)	91.8 ± 2.6 (93.0)
	KPT	**86.3 ± 1.6** (87.0)	**98.0 ± 0.2 (98.1)**	**68.0 ± 0.6 (68.2)**	93.8 ± 1.2 (94.1)	92.9 ± 1.8 (93.3)
20	FT	85.4 ± 1.8	97.9 ± 0.2	54.2 ± 18.1	71.4 ± 4.3	78.5 ± 10.1
	PT	86.5 ± 1.6 (87.0)	97.9 ± 0.3 (98.1)	67.2 ± 1.1 (67.5)	93.5 ± 1.0 (94.4)	93.0 ± 1.1 (93.6)
	AUTO	85.7 ± 1.4 (86.1)	92.2 ± 2.7 (94.9)	65.0 ± 1.8 (66.9)	**93.9 ± 1.1** (94.1)	92.8 ± 2.0 **(94.0)**
	SOFT	86.4 ± 1.7 (87.1)	98.0 ± 0.3 (98.1)	67.4 ± 0.7 (67.5)	93.8 ± 1.6 (94.2)	**93.5 ± 0.9 (94.0)**
	KPT	87.2 ± 0.8 (87.5)	**98.1 ± 0.3 (98.2)**	68.9 ± 0.8 (69.3)	93.7 ± 1.6 (94.4)	93.1 ± 1.1 (93.5)

Few Shot baseline

accuracy

Shot	Method	AG's News	DBPedia	Yahoo	Amazon	IMDB
1	KPT	**83.7 ± 3.5** (84.6)	**93.7 ± 1.8** (95.3)	**63.2 ± 2.5** (64.1)	93.2 ± 1.3 (93.9)	92.2 ± 3.0 (93.6)
	- LR	83.5 ± 3.8 (84.3)	93.0 ± 1.8 (94.5)	62.2 ± 2.9 (63.6)	**93.3 ± 1.3** (93.9)	92.2 ± 2.8 (93.6)
	- RR	82.2 ± 3.2 (82.6)	92.9 ± 1.8 (94.1)	61.3 ± 4.2 (62.5)	93.1 ± 1.5 (93.7)	**92.6 ± 1.7** (93.6)
	- RR - LR	81.8 ± 3.3 (82.5)	91.3 ± 1.7 (92.6)	60.7 ± 4.2 (61.4)	93.2 ± 1.5 (93.9)	92.6 ± 1.5 (93.5)
5	KPT	85.0 ± 1.2 (85.9)	97.1 ± 0.4 (97.3)	**67.2 ± 0.8** (67.8)	93.4 ± 1.9 (94.1)	92.7 ± 1.5 (92.9)
	- LR	**85.1 ± 1.0** (85.8)	97.1 ± 0.4 (97.2)	67.0 ± 1.1 (67.5)	93.4 ± 1.9 (94.1)	92.8 ± 1.5 (93.0)
	- RR	84.3 ± 1.8 (84.9)	**97.2 ± 0.4** (97.3)	**67.2 ± 0.8** (67.7)	**93.6 ± 1.4** (94.1)	**93.0 ± 2.0** (93.8)
	- RR - LR	84.2 ± 1.7 (84.5)	97.1 ± 0.4 (97.3)	66.6 ± 1.4 (67.5)	93.4 ± 2.0 (94.1)	93.0 ± 2.1 (93.8)
10	KPT	**86.3 ± 1.6** (87.0)	**98.0 ± 0.2** (98.1)	**68.0 ± 0.6** (68.2)	93.8 ± 1.2 (94.1)	92.9 ± 1.8 (93.3)
	- LR	85.9 ± 1.9 (87.1)	98.0 ± 0.2 (98.1)	67.9 ± 0.7 (68.2)	93.9 ± 1.1 (94.1)	**93.0 ± 1.7** (93.2)
	- RR	85.6 ± 1.4 (86.2)	97.9 ± 0.2 (98.0)	67.5 ± 1.1 (68.1)	94.0 ± 1.0 (94.7)	92.7 ± 2.1 (93.0)
	- RR - LR	85.1 ± 1.4 (86.0)	97.8 ± 0.2 (97.8)	66.8 ± 1.1 (67.6)	**94.1 ± 0.9** (94.8)	93.0 ± 2.0 (93.4)
20	KPT	87.2 ± 0.8 (87.5)	**98.1 ± 0.3** (98.2)	68.9 ± 0.8 (69.3)	93.7 ± 1.6 (94.4)	93.1 ± 1.1 (93.5)
	- LR	**87.7 ± 0.6** (87.8)	**98.1 ± 0.3** (98.2)	68.8 ± 0.9 (69.8)	93.4 ± 2.3 (94.3)	93.4 ± 0.9 (93.6)
	- RR	87.3 ± 0.8 (87.5)	**98.1 ± 0.3** (98.2)	68.8 ± 0.9 (68.9)	93.6 ± 1.3 (94.2)	93.1 ± 0.8 (93.6)
	- RR - LR	87.1 ± 0.9 (87.4)	**98.1 ± 0.3** (98.2)	**69.0 ± 0.7** (69.3)	93.7 ± 0.9 (94.5)	93.1 ± 0.8 (93.7)

Handle the OOV Label Words(out-of-vocabulary)

The knowledgeable verbalizer is expanded using external resources

- Average of each token in the single token [**Mask**]

Shot	Method	AG's News	DBPedia	Yahoo	Amazon	IMDB accuracy
0	KPT + ST	84.9 ↑ ± 1.0 (86.3)	81.0 ↓ ± 4.3 (85.2)	62.7 ↑ ± 1.1 (64.4)	92.8 ± 1.2 (94.7)	91.5 ↓ ± 2.8 (94.1)
1	KPT + ST	83.4 ↓ ± 3.9 (84.2)	94.0 ↑ ± 1.8 (95.8)	62.5 ↓ ± 2.3 (63.5)	93.3 ↑ ± 1.4 (94.1)	92.1 ↓ ± 3.5 (93.6)
5	KPT + ST	84.7 ↓ ± 1.8 (85.4)	97.1 ± 0.5 (97.2)	66.8 ↓ ± 1.0 (67.3)	93.3 ↓ ± 2.1 (93.8)	93.1 ↑ ± 1.4 (93.3)
10	KPT + ST	86.3 ± 1.5 (86.8)	98.0 ± 0.2 (98.1)	67.6 ↓ ± 0.9 (67.9)	94.0 ↑ ± 1.0 (94.1)	92.7 ↓ ± 1.8 (93.6)
20	KPT + ST	87.2 ↓ ± 1.1 (87.6)	97.9 ↓ ± 0.4 (98.1)	68.6 ↓ ± 0.7 (69.1)	93.5 ↑ ± 1.8 (94.0)	92.9 ↓ ± 1.2 (93.4)

-

Conclusion

Conclusion

1. Propose KPT , which expands the **verbalizer** in prompt-tuning using the external KB.

2. Better utilize the KB, we propose refinement methods for the knowledgeable.

Open questions

1. Better approaches for combining KB and prompt-tuning in terms of template construction and verbalizer design.

2. Incorporating external knowledge into prompt-tuning for other tasks such as text generation.