Prototypical Verbalizer for Prompt-based Few-shot Tuning

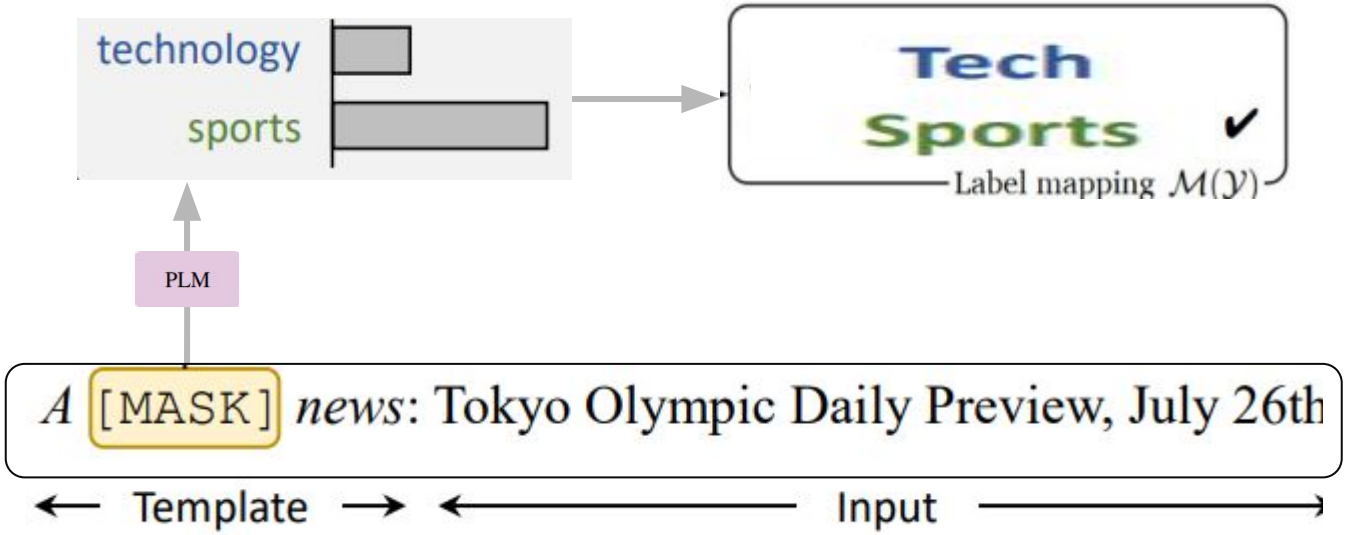
Source: Acl 2022
Advisor: JIA-LING KOH
Speaker: FAN-CHI-YU
Date:2023/10/03

Outline

- Introduction
- Method
- Experiment
- Conclusion

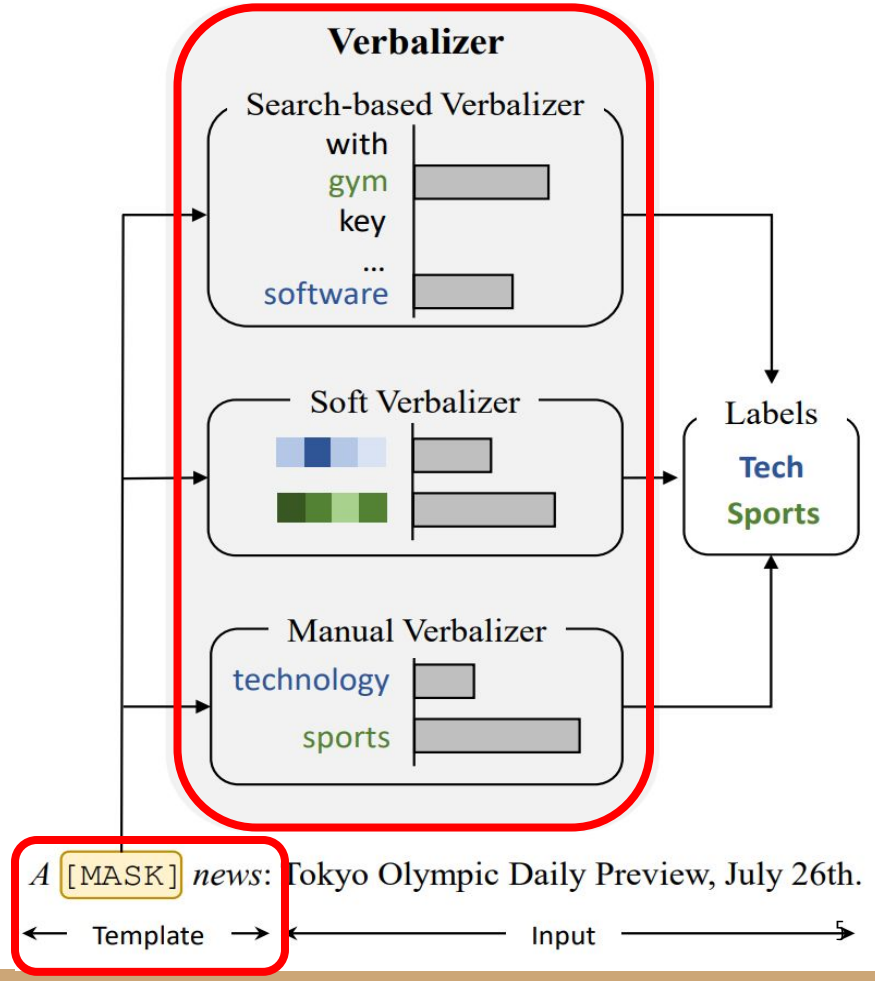
Introduction

Introduction(Prompt Tuning)



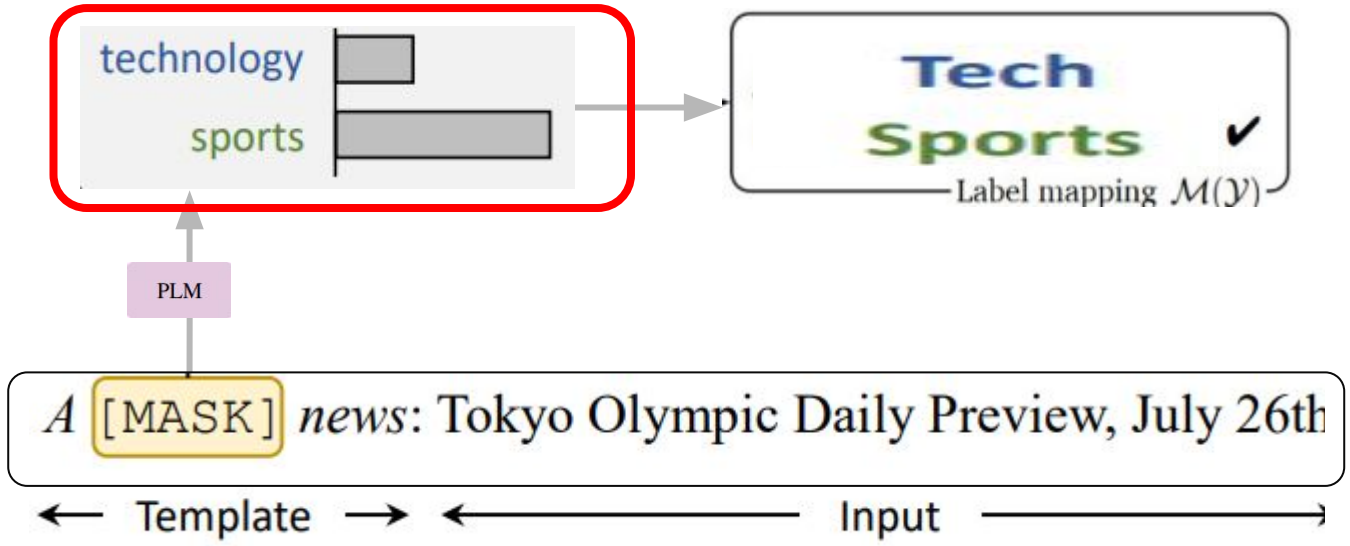
Introduction

- Prompt-based tuning includes two key points:
  - Template design
  - **Verbalizer design**



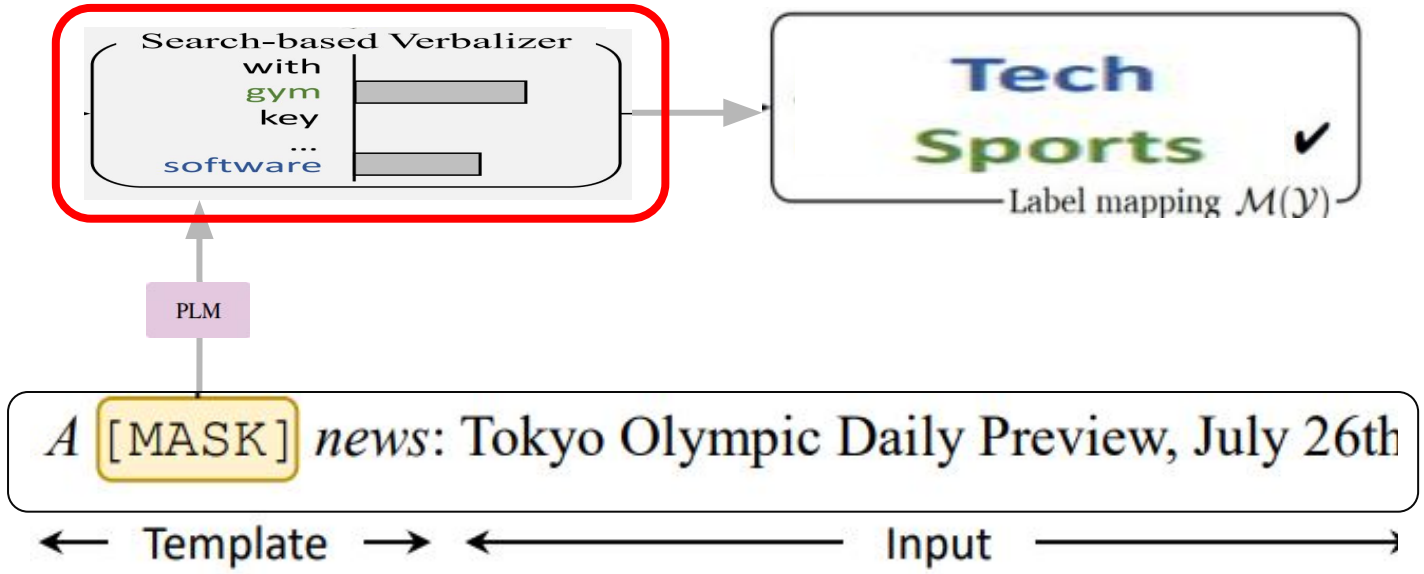
Introduction(Manual Verbalizer)

Defined by human with domain knowledge



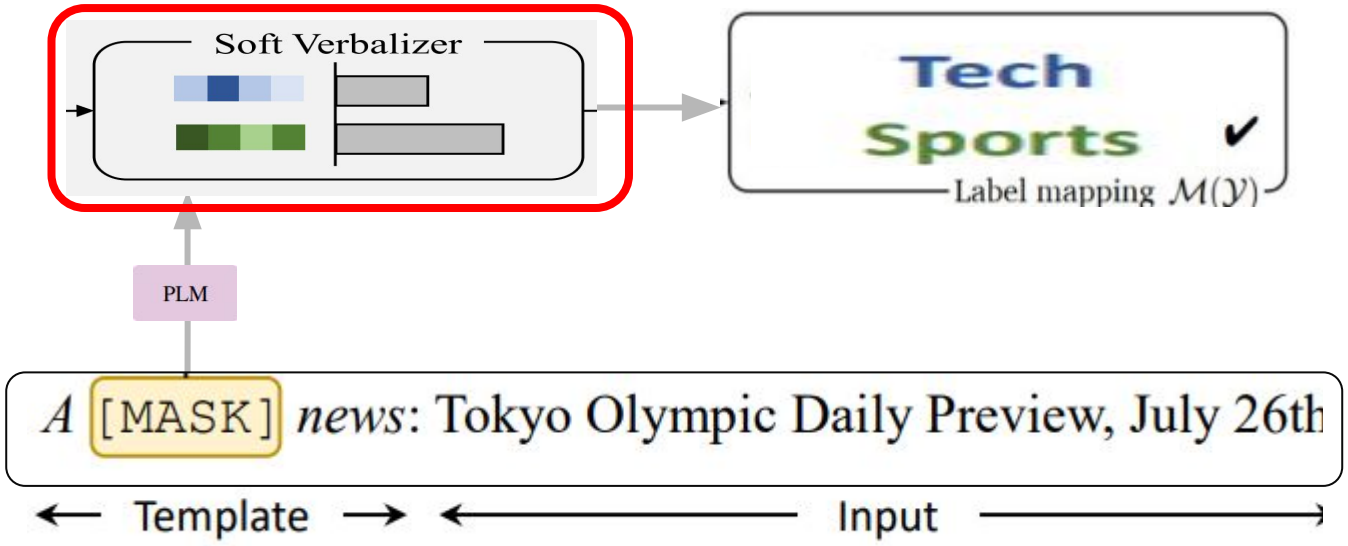
Introduction(Search-based Verbalizer)

Search for suitable words from vocabulary automatically



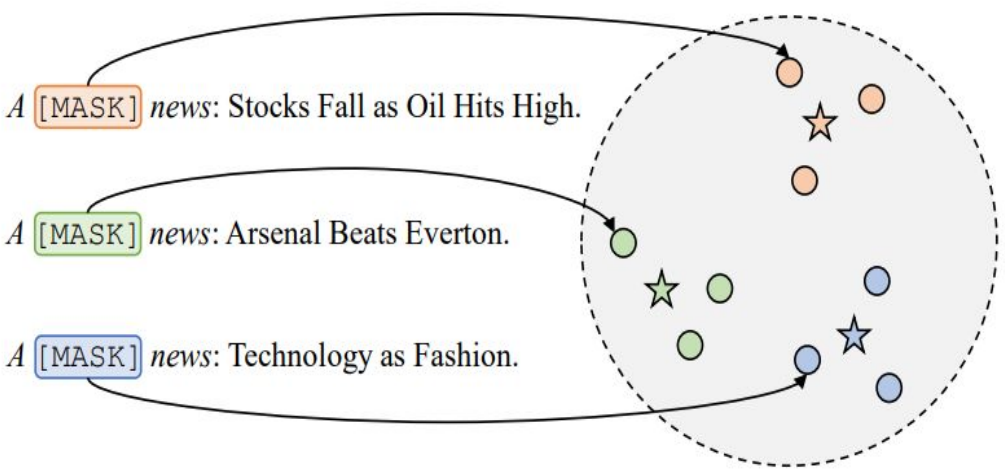
Introduction(Soft Verbalizer)

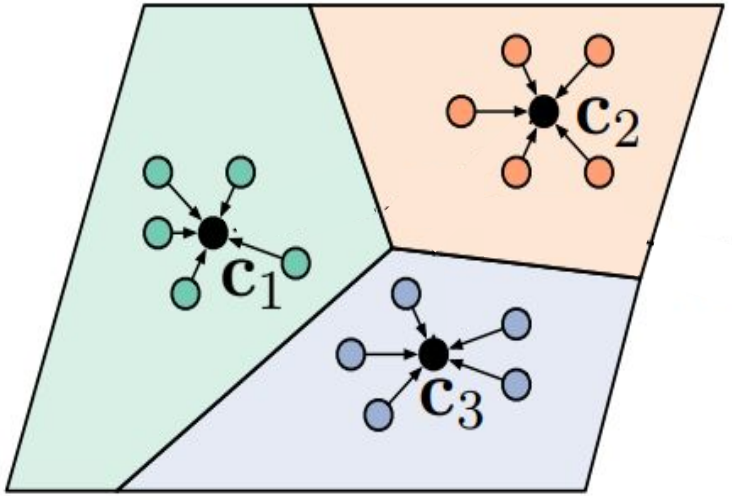
Trainable tokens as verbalizers



Prototypical Networks

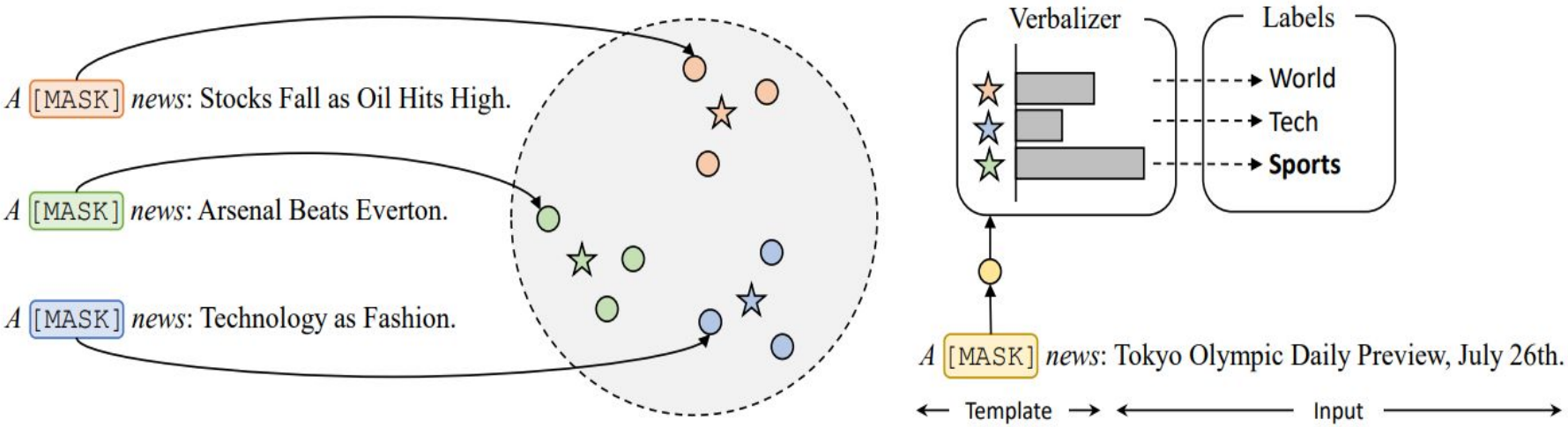
Few-shot prototypes **class** are **computed** as the **mean** of embedded support examples for each class





Method

Method

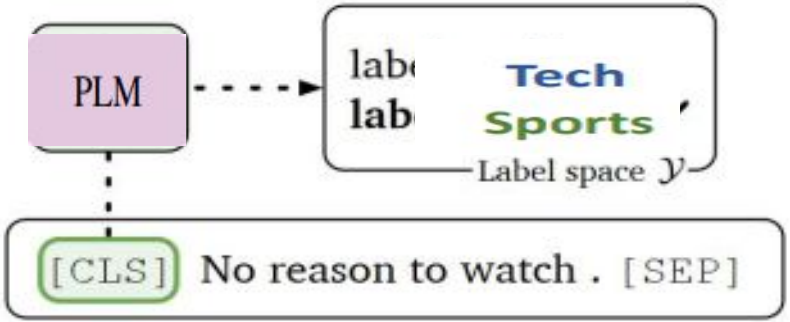

A [MASK] news: Stocks Fall as Oil Hits High.
A [MASK] news: Arsenal Beats Everton.
A [MASK] news: Technology as Fashion.
Verbalizer
Labels
World
Tech
Sports
A [MASK] news: Tokyo Olympic Daily Preview, July 26th.
Template
Input

Background(Fine-tuning)

classifier

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

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

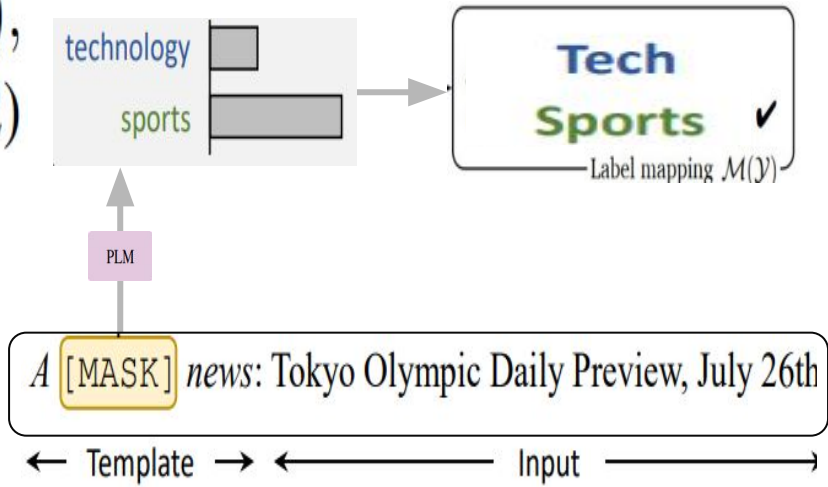


Background(Prompt Tuning)

aggregate multiple scores.

$$P_{\mathcal{M}}(y|x) = g(P_{\mathcal{M}}([\texttt{MASK}] = v|\mathcal{T}(x))|v \in \mathcal{V}_y), \quad (2)$$

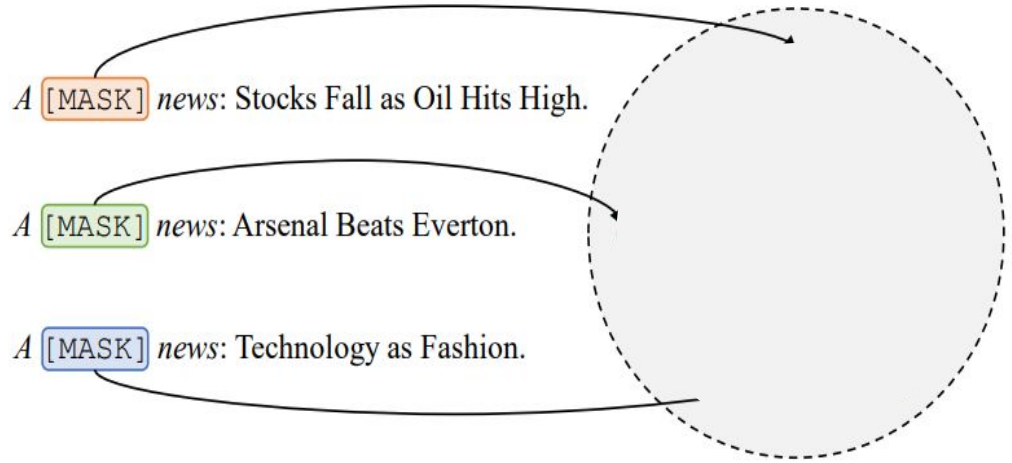
hidden vector



Projection

$$\mathbf{v} = E_\phi(x) = \mathbf{W}\mathbf{h}_{\texttt{[MASK]}}. \quad (3)$$

$$S(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i}{||\mathbf{v}_i||} \cdot \frac{\mathbf{v}_j}{||\mathbf{v}_j||}. \quad (4)$$



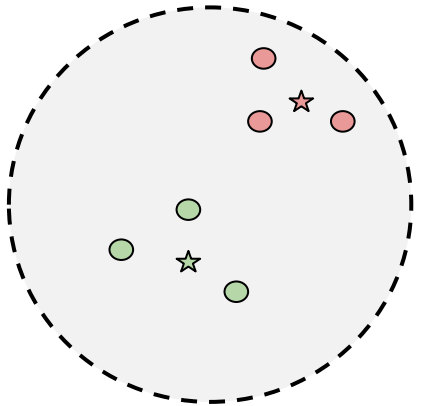
Prototypical networks

ProtoNet calculates prototype vectors by taking the average of instance vectors

prototype vectors

$$\mathcal{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_N\}$$

$$\mathbf{c}_n = \frac{1}{|S_n|} \sum_{(\mathbf{x}_i, y_i) \in S_n} E_\phi(\mathbf{x}_i)$$

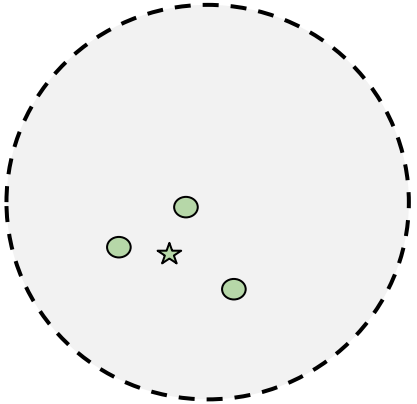


Loss(instance & instance)

Maximizes **intra**-instance similarity between instance & instance

$$\mathcal{L}_{\text{ins}} = \frac{-1}{N^2K^2} \sum_{n} \sum_{i,j} \log \frac{\exp S(\mathbf{v}_i^n, \mathbf{v}_j^n)}{\sum_{n\prime,j\prime} \exp S(\mathbf{v}_i^n, \mathbf{v}_{j\prime}^{n\prime})}, \tag{5}$$

instance pairs of the **same** class

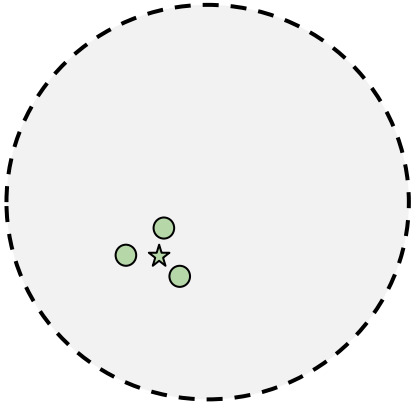


Loss(instance & instance)

Maximizes **intra**-instance similarity between instance & instance

$$\mathcal{L}_{\text{ins}} = \frac{-1}{N^2K^2} \sum_{n} \sum_{i,j} \log \frac{\exp S(\mathbf{v}_i^n, \mathbf{v}_j^n)}{\sum_{n\prime,j\prime} \exp S(\mathbf{v}_i^n, \mathbf{v}_{j\prime}^{n\prime})}, \tag{5}$$

instance pairs of the **same** class

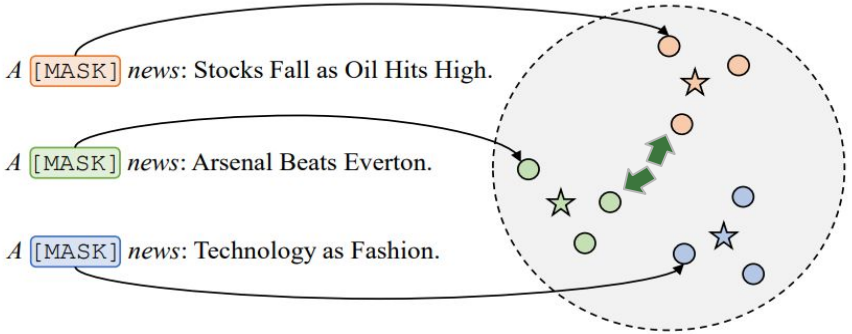


Loss(instance & instance)

Minimizes **inter**-class similarity between instance & instance

$$\mathcal{L}_{\text{ins}} = \frac{-1}{N^2K^2}\sum_{n}\sum_{i,j}\log\frac{\exp S(\mathbf{v}_i^n, \mathbf{v}_j^n)}{\sum_{n',j'}\exp S(\mathbf{v}_i^n, \mathbf{v}_{j'}^{n'})} \quad (5)$$

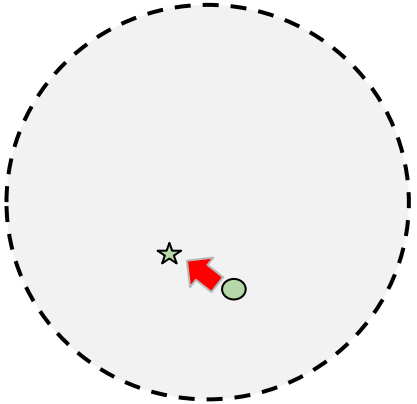
instance pairs of the **inter**-class



Loss(instance & class)

Maximizes **intra**-class similarity between instance & class

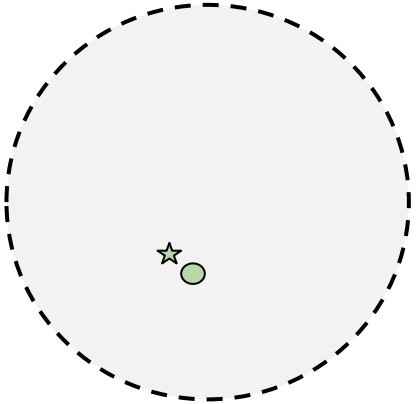
$$\mathcal{L}_{\text{proto}} = \frac{-1}{N^2 K} \sum_{i,n} \log \frac{\exp S(\mathbf{v}_i^n, \mathbf{c}_n)}{\sum_{n'} \exp S(\mathbf{v}_i^n, \mathbf{c}_{n'})}, \quad (6)$$



Loss(instance & class)

Maximizes **intra**-class similarity between instance & class

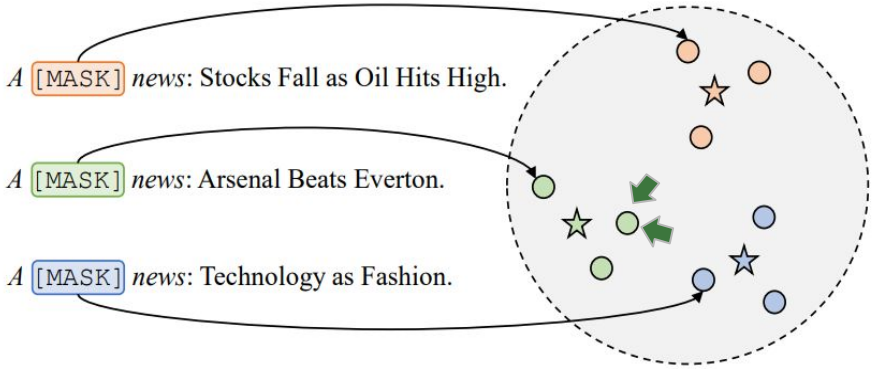
$$\mathcal{L}_{\text{proto}} = \frac{-1}{N^2 K} \sum_{i,n} \log \frac{\exp S(\mathbf{v}_i^n, \mathbf{c}_n)}{\sum_{n\prime} \exp S(\mathbf{v}_i^n, \mathbf{c}_{n\prime})}, \quad (6)$$



Loss(instance & class)

Minimizes **inter**-class similarity between instance & class

$$\mathcal{L}_{\text{proto}} = \frac{-1}{N^2 K} \sum_{i,n} \log \frac{\exp S(\mathbf{v}_i^n, \mathbf{c}_n)}{\sum_{n'} \exp S(\mathbf{v}_i^n, \mathbf{c}_{n'})}, \quad (6)$$



$$\mathcal{L} = \mathcal{L}_{\text{ins}} + \mathcal{L}_{\text{proto}}. \quad (7)$$

Inference

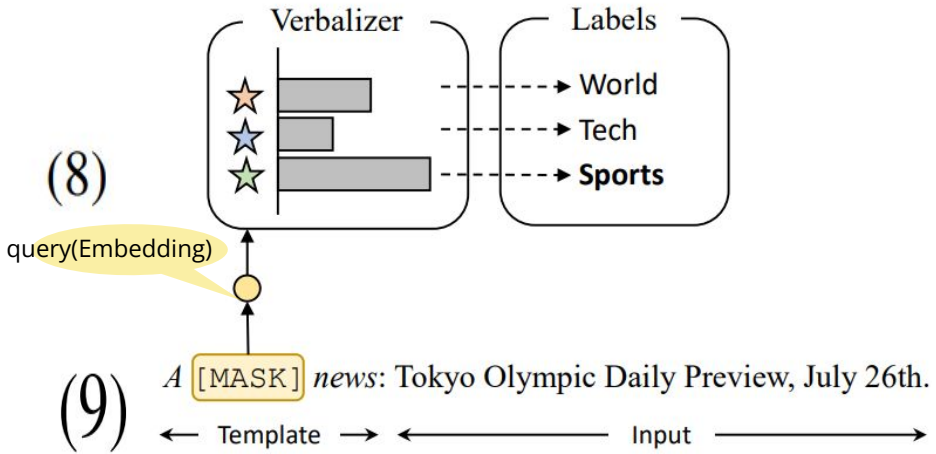
Calculate the similarity scores of query and prototypes

$$P_{\mathcal{M}}(y_k|x) = \frac{\exp S(\mathbf{v}, \mathbf{c}_k)}{\sum_{k'} \exp S(\mathbf{v}, \mathbf{c}_{k'})}. \quad (8)$$

query

prototypes

$$\tilde{y} = \arg\max_k P_{\mathcal{M}}(y_k|x). \quad (9)$$



Experiment

Experiment

Topic Classification

$\mathcal{T}_1(x) =$ A [MASK] news: x
$\mathcal{T}_2(x) = x$ This topic is about [MASK].
$\mathcal{T}_3(x) =$ [Category : [MASK]] x
$\mathcal{T}_4(x) =$ [Topic : [MASK]] x

Dataset	Task	#Class	#Test
AG's News	TC	4	7,600
DBPedia	TC	14	70,000
Yahoo	TC	10	60,000
FewNERD	ET	66	96,901

Entity Typing

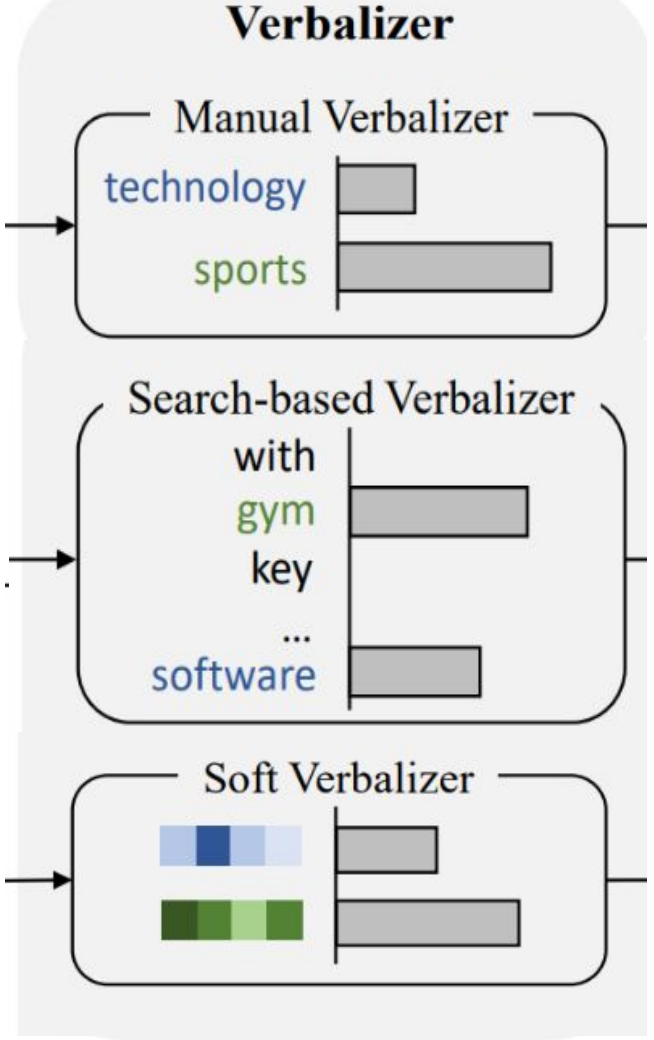
Example:The University of Washington[education] is a public research university in Seattle, Washington.[location]

Baselines

1. Manual Verbalizer

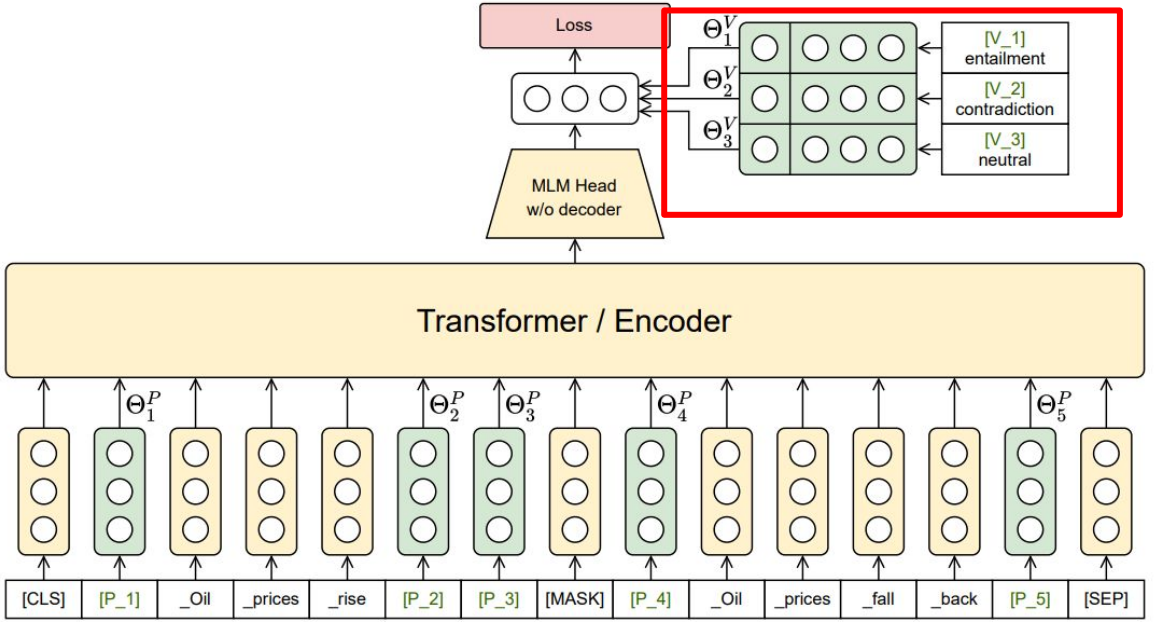
2. Search-based Verbalizer

3. Soft Verbalizer



Soft Verbalizer(WARP)

Learn the **Verbalizer** in Embedding method



Baselines

1. **ProtoVerb gets better results on topic classification(TC) than entity typing(ET).**

2. ProtoVerb catch up with ManualVerb with enough samples.

K	Method	AG	DB	Yahoo	Few
0	ManualVerb	*75.13*	*67.06*	*43.11*	*20.00*
1	ManualVerb	*76.67*	*85.47*	*50.22*	*41.68*
	SearchVerb	41.50	60.06	27.39	20.88
	SoftVerb	49.79	65.35	22.72	18.78
	ProtoVerb	**64.19**	**72.85**	**36.12**	**25.00**
2	ManualVerb	*81.06*	*93.61*	*58.65*	*46.44*
	SearchVerb	65.82	78.21	40.71	31.28
	SoftVerb	56.37	80.69	30.72	32.80
	ProtoVerb	**77.34**	**85.49**	**46.30**	**35.72**
16	ManualVerb	*84.74*	*96.05*	*58.77*	*61.17*
	SearchVerb	83.40	92.00	59.66	55.49
	SoftVerb	80.57	86.90	58.20	58.87
	ProtoVerb	**84.48**	**96.30**	**64.35**	**61.29**

Baselines

1. ProtoVerb gets better results on topic classification than entity typing.

2. **ProtoVerb catch up with ManualVerb with enough samples.**

3. ProtoVerb will surpass the Manual in **shot-2**

K	Method	AG	DB	Yahoo	Few
0	ManualVerb	*75.13*	*67.06*	*43.11*	*20.00*
1	ManualVerb	*76.67*	*85.47*	*50.22*	*41.68*
	SearchVerb	41.50	60.06	27.39	20.88
	SoftVerb	49.79	65.35	22.72	18.78
	ProtoVerb	**64.19**	**72.85**	**36.12**	**25.00**
2	ManualVerb	*81.06*	*93.61*	*58.65*	*46.44*
	SearchVerb	65.82	78.21	40.71	31.28
	SoftVerb	56.37	80.69	30.72	32.80
	ProtoVerb	**77.34**	**85.49**	**46.30**	**35.72**
16	ManualVerb	*84.74*	*96.05*	*58.77*	*61.17*
	SearchVerb	83.40	92.00	59.66	55.49
	SoftVerb	80.57	86.90	58.20	58.87
	ProtoVerb	**84.48**	**96.30**	**64.35**	**61.29**

Baselines

1. ProtoVerb+ provides a better way to utilize training data

2. ProtoVerb+ boosts them considerably on **all tasks.**

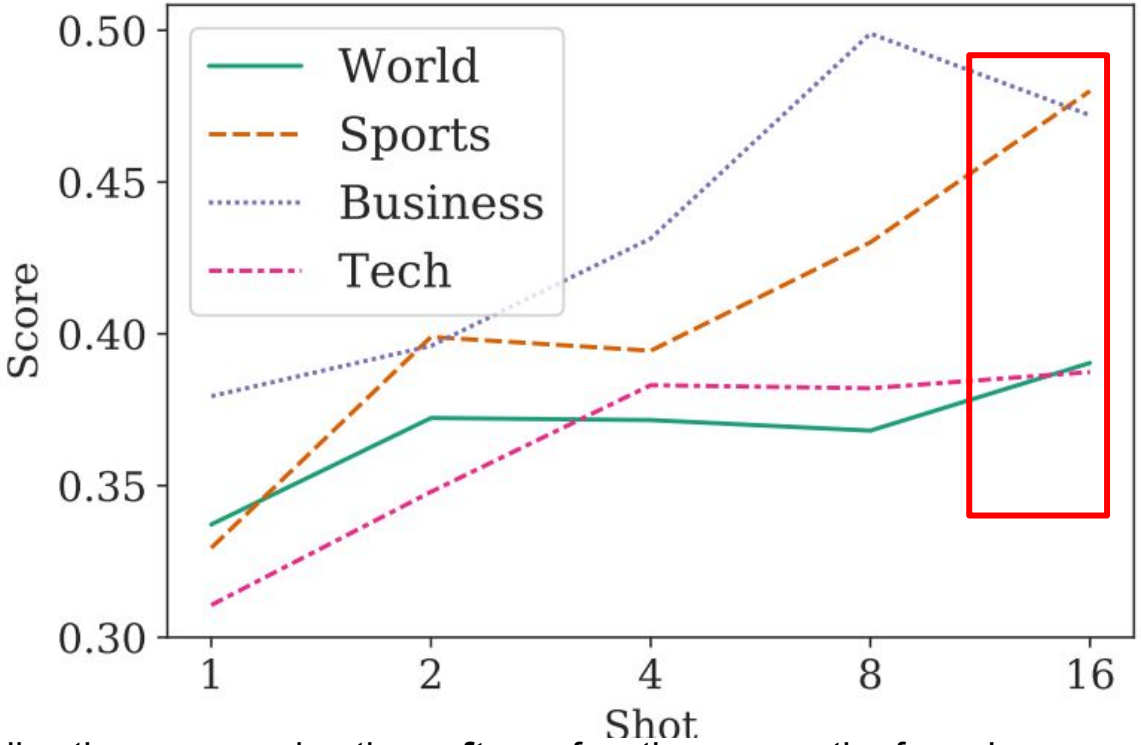
K	Method	AG	DB	Yahoo	Few
1	Fine-tuning	25.45	10.80	10.59	7.48
	ManualVerb	76.67	85.47	**50.22**	41.68
	ProtoVerb+	**77.71**	**88.16**	50.08	**43.20**
	w/o tuning	76.28	78.32	45.01	29.51
2	Fine-tuning	25.78	49.01	11.26	19.03
	ManualVerb	81.06	93.61	58.65	46.44
	ProtoVerb+	**84.09**	**94.77**	**59.33**	**48.69**
	w/o tuning	82.13	86.11	50.34	34.44
16	Fine-tuning	84.14	**97.25**	64.27	52.66
	ManualVerb	84.74	96.05	58.77	61.17
	ProtoVerb+	**87.98**	97.22	**65.65**	**62.55**
	w/o tuning	84.78	93.46	60.89	33.96

Is ProtoVerb Similar with ManualVerb ?

World and **Tech** news includes

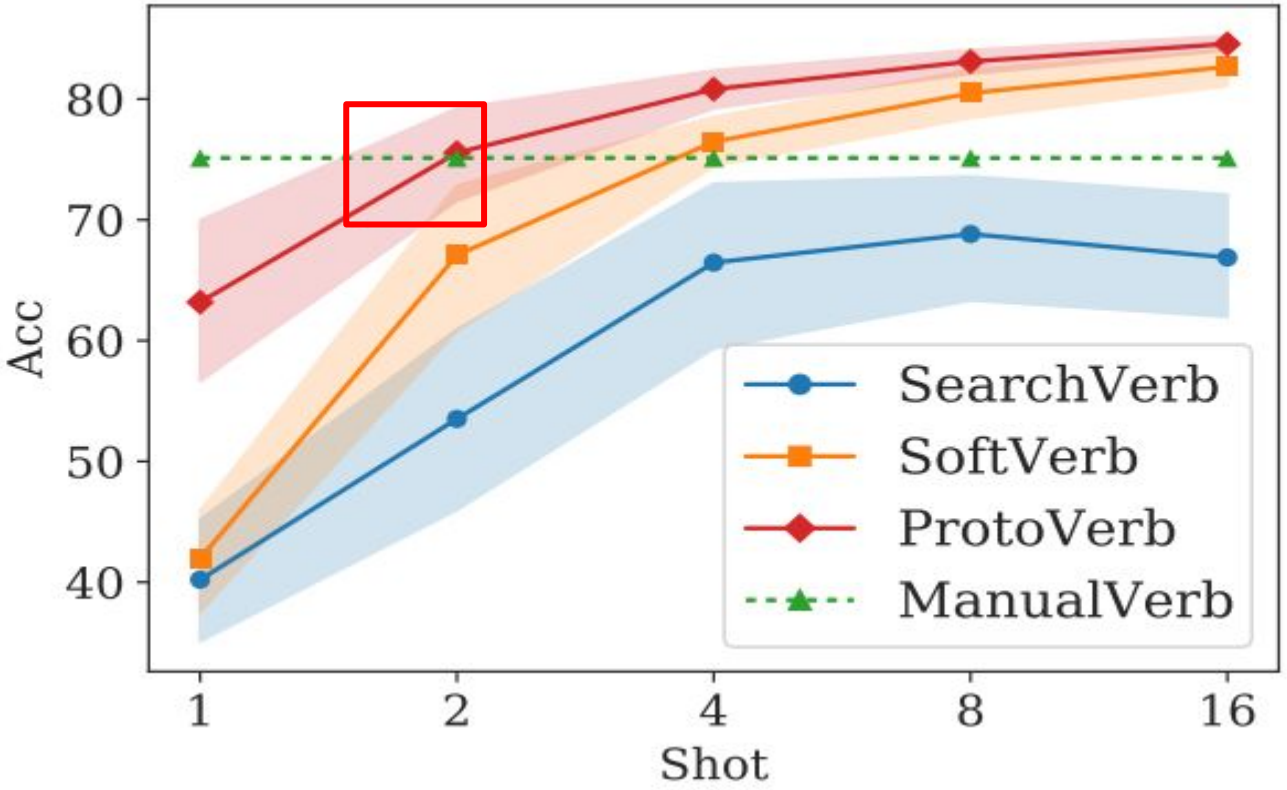
diverse sub-topics

that are hard to summarize.



Normalize the scores using the **softmax** function across the four classes

Fixed Model Experiments



Ablation

1. If the Sentence more than the accuracy will effect by **instance and instance loss.**
2. If the Sentence few will more effect by the **instance and class loss.**

Method	$K = 2$	$K = 4$	$K = 8$
$\mathcal{L}_{\text{ins}} + \mathcal{L}_{\text{proto}}$	**77.34**	**81.65**	**84.03**
$\mathcal{L}_{\text{proto}}$	76.37	81.06	82.91
Instance Mean	73.36	77.76	82.57

Noisy Samples

ProtoVerb is more robust than baseline methods when facing noisy samples.

K	Method	# Noisy Samples		
		1	2	3
8	SearchVerb	4.86	5.96	5.19
	SoftVerb	4.84	7.80	11.71
	ProtoVerb	**2.34**	**3.11**	**4.37**
16	SearchVerb	0.80	2.93	5.18
	SoftVerb	2.01	4.17	4.58
	ProtoVerb	**0.04**	**2.13**	**3.16**

Conclusion

Conclusion

1. A novel approach Automatic verbalizer construction in prompt-based tuning

2. ProtoVerb consistently **improve** promptbased tuning with **minor effort**.

3. ProtoVerb outperforms **state-of-the-art automatic verbalizers** considerably.