MetricPrompt: Prompting Model as a Relevance Metric for Few-shot Text Classification

task

Advisor : Jia-Ling, Koh

Speaker : Yu-Zhi, Liu

Source : KDD '23

Date : 2023/11/21

Outline

- Introduction
- Method
- Experience
- Conclusion

Text Classification

text

Classification

sport

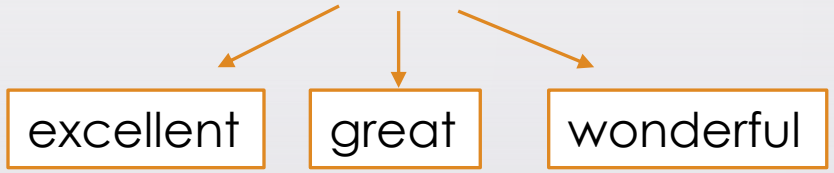
politics

business

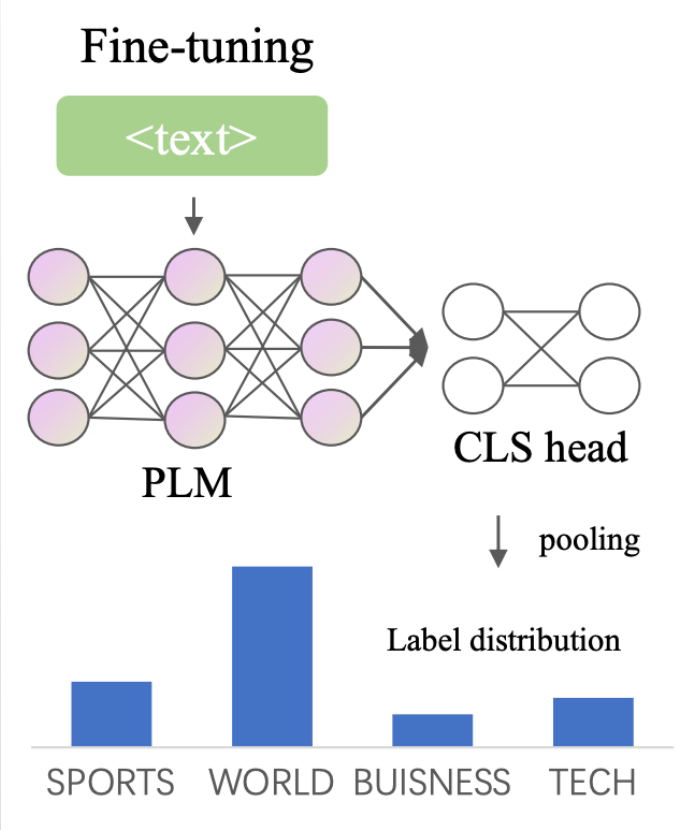
tech

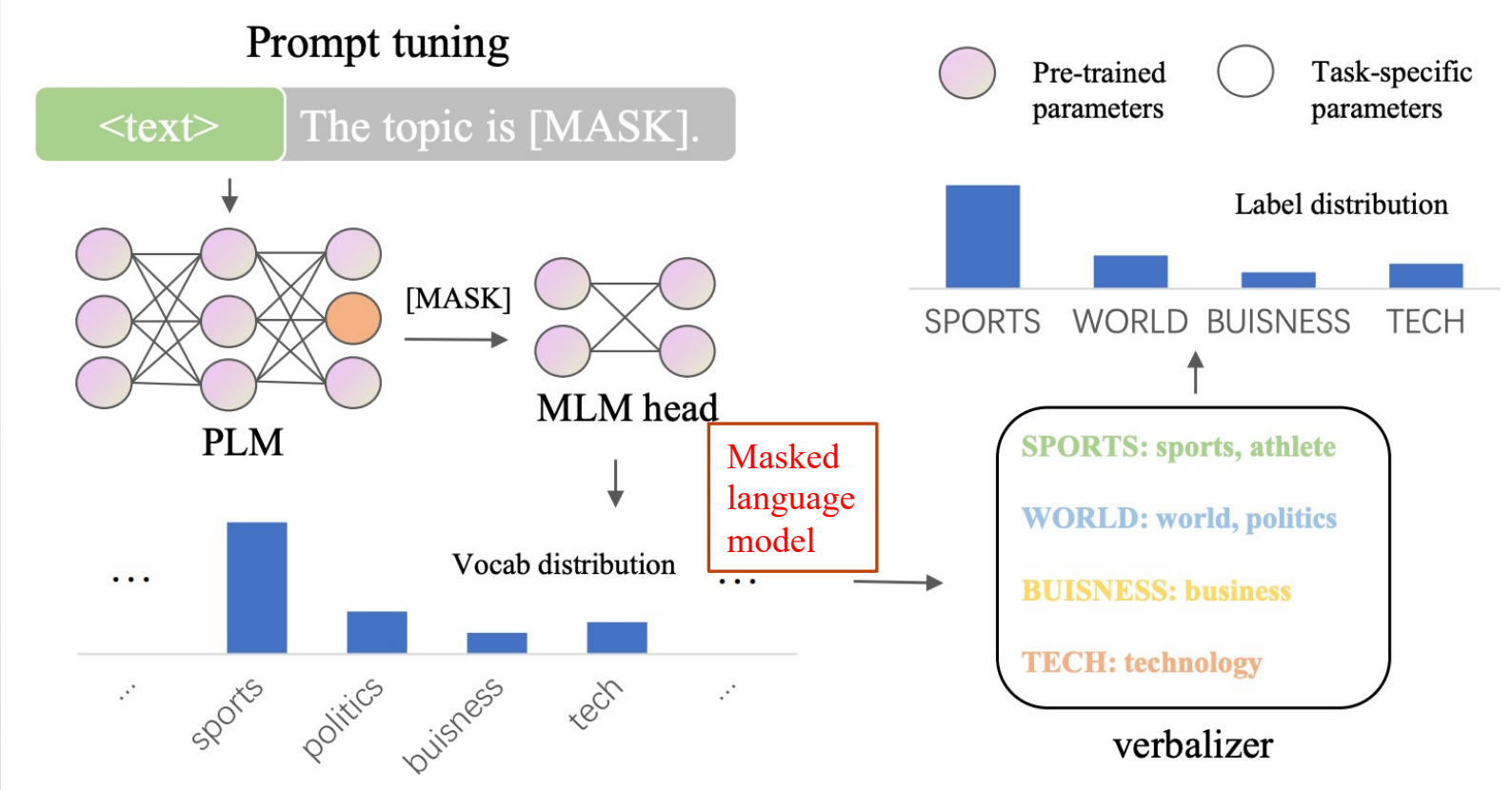
Prompting Model

- Input : “I love this movie.”
- Prompt: “I love this movie. Overall it was a [Z] movie.”

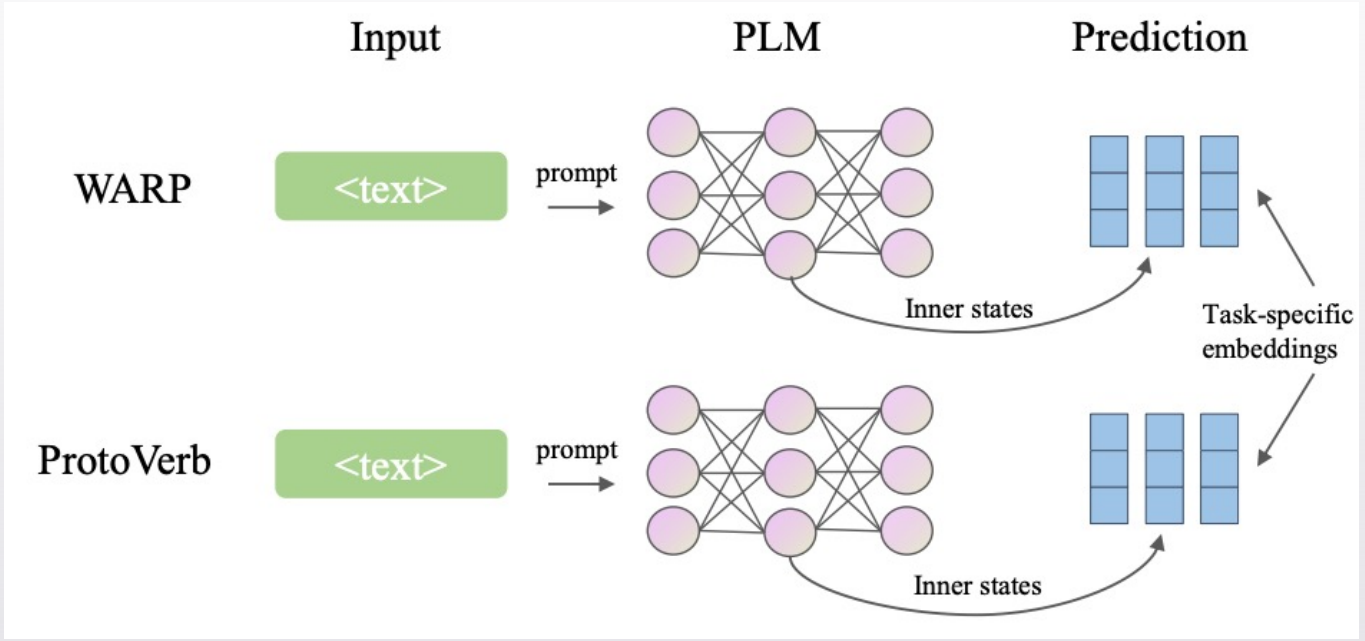


A comparison between fine-tuning and prompt-tuning





Problem



Overfitting - problem

MetricPrompt and other verbalizers

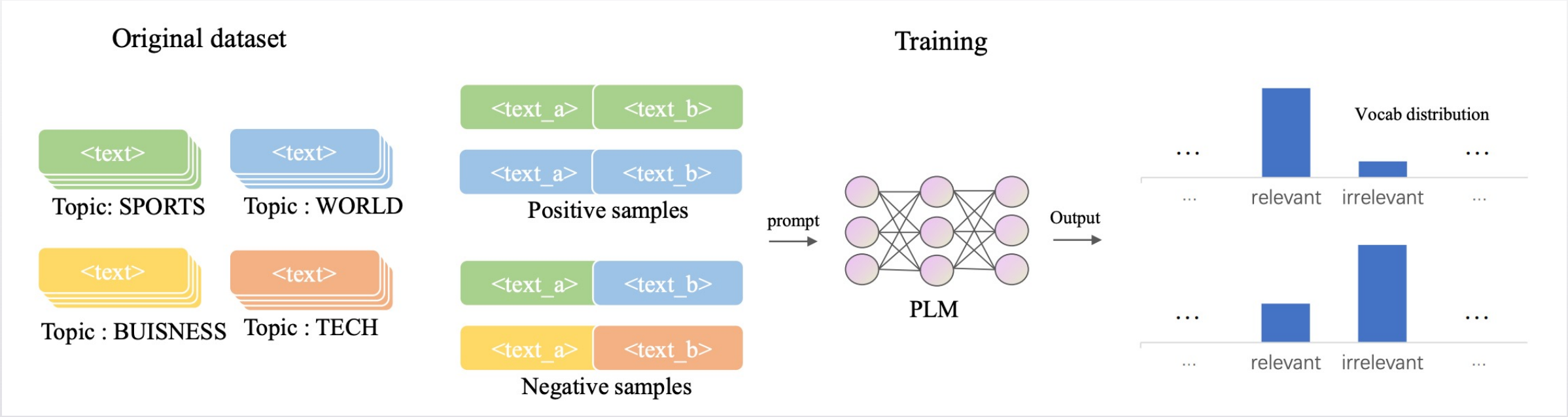
Method	Dataset	Prompt template	Task-specific verbalizer
ManualVerb	AG's News	A [MASK] news: <text>	sports, politics, business, technology
	DBPedia	<text> In this sentence, the topic is [MASK]	company, school, artist, athlete, politics, transportation, building, river, village, animal, plant, album, film, book
	Yahoo	A [MASK] question: <text>	society, science, health, education, computers, sports, business, entertainment, relationships, politics
AVS	AG's News	A [MASK] news: <text>	Automatically searched label words
	DBPedia	<text> In this sentence, the topic is [MASK]	Automatically searched label words
	Yahoo	A [MASK] question: <text>	Automatically searched label words
SoftVerb	AG's News DBPedia Yahoo	<text> In this sentence, the topic is [MASK]	Soft label embeddings
ProtoVerb	AG's News	A [MASK] news: <text>	Soft label embeddings
	DBPedia	<text> In this sentence, the topic is [MASK]	Soft label embeddings
	Yahoo	A [MASK] question: <text>	Soft label embeddings
MetricPrompt	AG's News DBPedia Yahoo	<text_a> A news of [MASK] topic: <text_b>	–

Means no task-specific verbalizer is required

Outline

- Introduction
- Method
- Experience
- Conclusion

Metricprompt's data construction and training procedure



Data construction

MetricPromt prompting function

$$\mathcal{D}_t^M = \bigcup_{(d_i,d_j)\in\mathcal{D}_t\times\mathcal{D}_t} \{(p(\mathbf{x}_{d_i}, \mathbf{x}_{d_j}), \mathbf{y}_{ij})\},$$

Training data

$$\mathcal{D}_q^M = \bigcup_{(d_i,d_j)\in\mathcal{D}_q\times\mathcal{D}_t} \{(p(\mathbf{x}_{d_i}, \mathbf{x}_{d_j}), \mathbf{y}_{ij})\}.$$

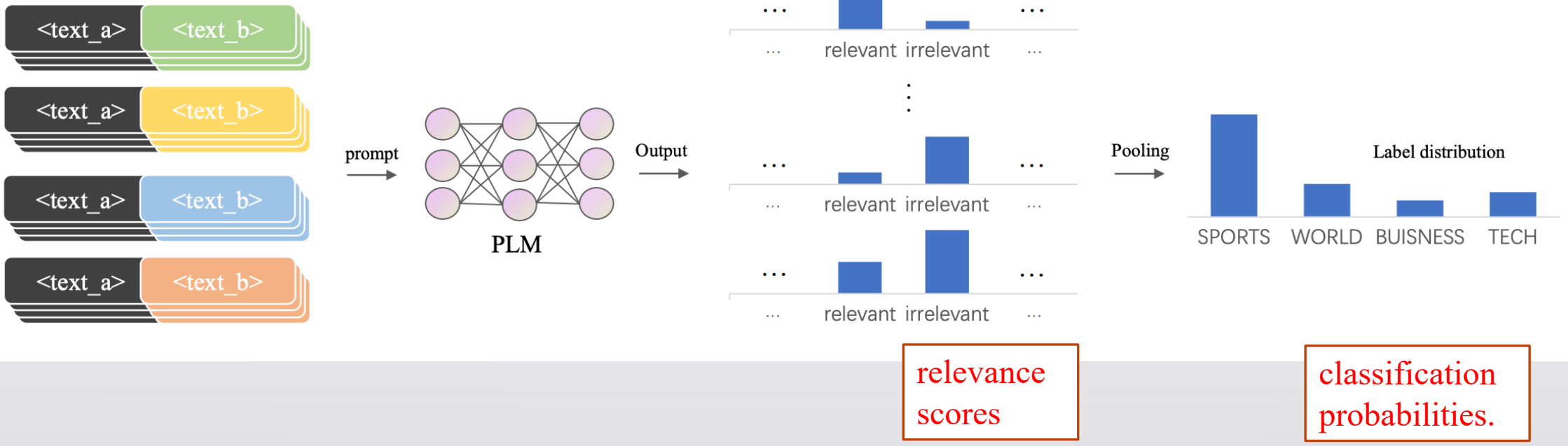
query sample

->$p(\ )$ is MetricPromt's prompting function.

->Xd is sample text.

->y represents its label

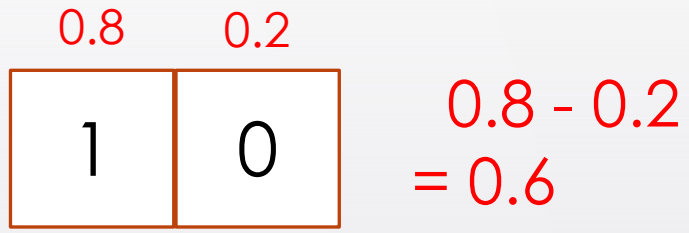
Inference


<text_a> <text_b>
<text_a> <text_b>
<text_a> <text_b>
<text_a> <text_b>
prompt
PLM
Output
Vocab distribution
relevant irrelevant
relevant irrelevant
relevant irrelevant
Pooling
Label distribution
SPORTS WORLD BUISNESS TECH
relevance scores
classification probabilities.

Inference

Probability at 1 and 0

$$s_{d_i} = \Delta(f_{cls}(p(\mathbf{x}_{d_q}, \mathbf{x}_{d_i}); \hat{\theta})),$$



$$s_l = \sum_{d_i \in \mathcal{D}_l} s_{d_i} / |\mathcal{D}_l|.$$

Relevance score

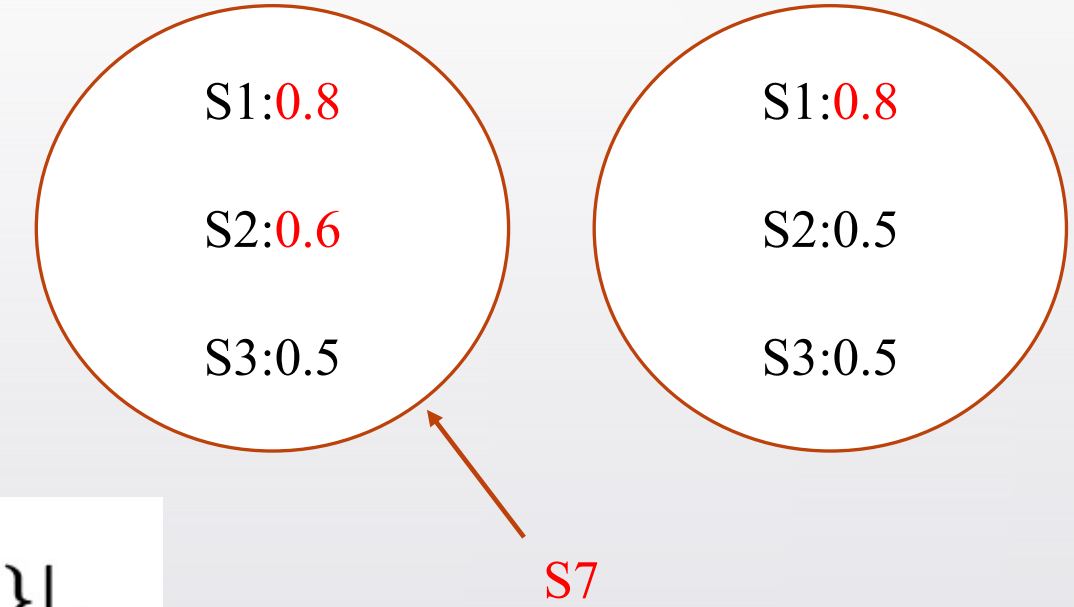
$$\hat{l} = \arg\max_{l} s_l.$$

Highest relevance score

Inference

KNN pooling

$$s_l = \max_{d_i \in \mathcal{D}_l} s_{d_i}.$$



$$s_l = |\{d_i | d_i \in \mathcal{D}_{topk}, \mathbf{y}_{d_i} = l_i\}|.$$

k training samples most relevant to dq in Dt

More Efficient Inference(pivot samples)

relevance score between samples dj and di

$$r_{d_i} = \frac{\sum_{d_j \in \mathcal{D}_l} s_{d_j}}{|\{d_j | d_j \in \mathcal{D}_l\}|} - \frac{\sum_{d_k \in \mathcal{D}_t - \mathcal{D}_l} s_{d_k}}{|\{d_k | d_k \in \mathcal{D}_t - \mathcal{D}_l\}|}.$$

Time complexity: $O(n * k)$ -> $O(n)$

Outline

- Introduction
- Method
- Experience
- Conclusion

Datasets

Dataset	# Class	# Test	Avg len
AG's News	4	7,600	52
DBPedia	14	70,000	68
Yahoo	10	60,000	130

Three text datasets

Dataset	2-shot	4-shot	8-shot	16-shot
AG's News	120	60	30	15
DBPedia	32	16	8	4
Yahoo	36	18	9	5

Training epochs

Experience

Method	AG's News		DBPedia		Yahoo		Average	
	2-shot	4-shot	2-shot	4-shot	2-shot	4-shot	2-shot	4-shot
ManualVerb	*45.87*	*76.22*	*69.81*	*84.15*	*34.82*	*55.56*	*50.17*	*71.98*
AVS [2021a]	44.93	57.49	32.22	53.55	21.88	28.44	33.01	46.49
SoftVerb [2021]	48.86	61.15	53.98	76.19	22.63	33.43	41.82	56.92
ProtoVerb [2022]	58.38	65.04	60.89	74.49	28.80	43.01	49.36	60.85
MetricPrompt$_{KNN}$	62.69	73.17	66.27	86.06	26.02	50.90	51.66	70.04
MetricPrompt$_{MAX}$	65.64	76.12	**71.28**	**88.44**	**28.85**	52.99	**55.26**	72.52
MetricPrompt$_{MEAN}$	**65.77**	**76.33**	71.20	**88.44**	28.76	**53.54**	55.24	**72.77**
MetricPrompt$_{PIVOT}$	65.76	74.53	71.20	86.12	28.76	51.32	55.24	70.66

accuracy

Experience

Method	AG's News		DBPedia		Yahoo		Average	
	8-shot	16-shot	8-shot	16-shot	8-shot	16-shot	8-shot	16-shot
ManualVerb	*78.94*	*83.66*	*94.24*	*97.27*	*58.30*	*62.42*	*77.16*	*81.12*
AVS [2021a]	71.37	77.81	75.91	85.36	46.53	57.68	64.60	73.62
SoftVerb [2021]	73.28	80.61	90.34	96.93	45.01	59.09	69.54	78.88
ProtoVerb [2022]	75.57	80.31	87.45	**97.16**	52.87	61.57	71.96	79.68
MetricPrompt$_{KNN}$	80.64	84.43	94.25	96.55	58.09	62.05	77.66	81.01
MetricPrompt$_{MAX}$	81.03	84.27	94.28	96.55	59.68	62.66	78.33	81.16
MetricPrompt$_{MEAN}$	**82.04**	**84.69**	**94.57**	96.59	**59.68**	**62.45**	**78.76**	**81.24**
MetricPrompt$_{PIVOT}$	81.19	84.15	94.13	96.22	58.63	61.78	77.98	80.72

Experience

Method	AG's News			Method	DBPedia			Method	Yahoo		
	1-shot	2-shot	4-shot		1-shot	2-shot	4-shot		1-shot	2-shot	4-shot
PROTOVERB	46.79	58.38	65.04	PROTOVERB	45.86	60.89	74.49	PROTOVERB	21.60	28.80	43.01
+DBPEDIA	57.43	65.72	71.27	+AG's NEWS	49.00	63.29	75.56	+AG's NEWS	28.93	39.15	49.96
+YAHOO	63.63	71.84	75.34	+YAHOO	**54.33**	66.78	77.56	+DBPEDIA	30.13	39.57	51.39
METRICPROMPT	39.16	65.77	76.33	METRICPROMPT	32.31	71.20	88.44	METRICPROMPT	18.80	28.76	53.54
+DBPEDIA	66.95	71.40	77.34	+AG's NEWS	53.23	74.21	88.34	+AG's NEWS	32.10	**44.27**	52.29
+YAHOO	**71.00**	**73.99**	**79.57**	+YAHOO	53.03	**76.41**	**89.47**	+DBPEDIA	**32.77**	43.63	**53.78**

+Out-Of-Domain (OOD) data

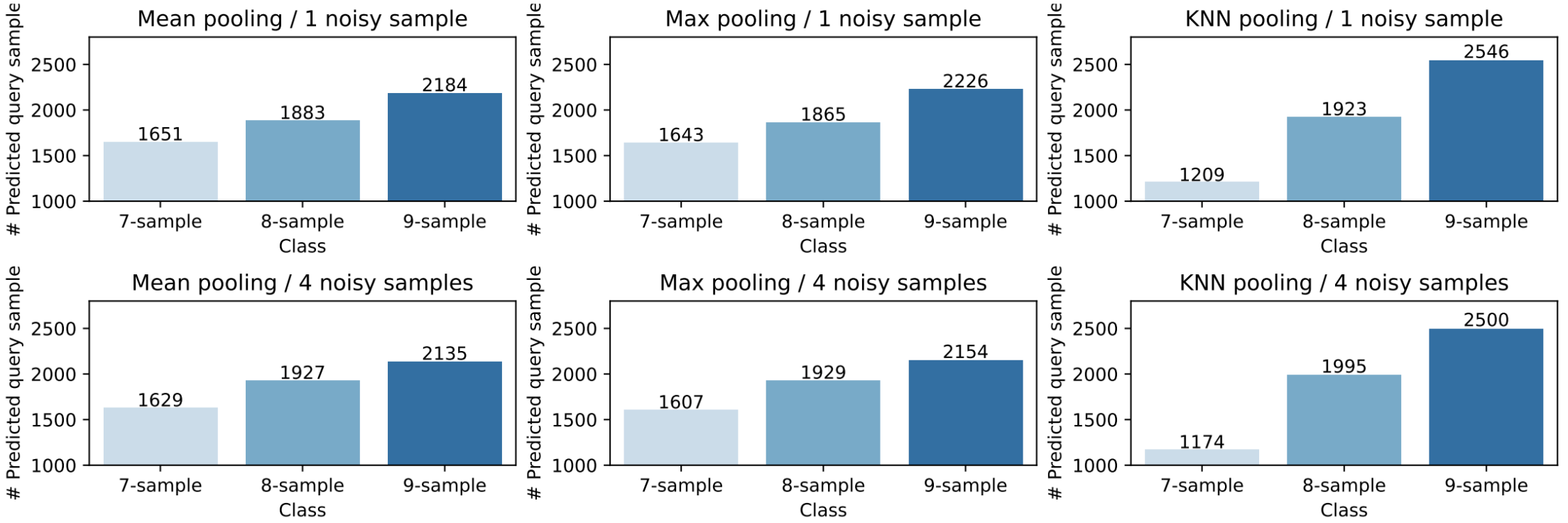
Experience

Noisy samples

Method	1 wrong		2 wrong		4 wrong		Average	
	8-shot	16-shot	8-shot	16-shot	8-shot	16-shot	8-shot	16-shot
ProtoVerb	2.79	0.83	4.95	1.85	11.31	3.71	6.35	2.13
MetricPrompt$_{KNN}$	5.74	2.61	5.66	3.08	12.38	3.38	7.93	3.02
MetricPrompt$_{MAX}$	**1.59**	0.59	3.12	**0.89**	**7.01**	**1.21**	3.91	**0.90**
MetricPrompt$_{MEAN}$	1.81	**0.55**	**2.72**	1.04	7.06	1.52	**3.86**	1.04

performance drop

Experience


Mean pooling / 1 noisy sample
Predicted query sample
2500
2000
1500
1000
1651
1883
2184
7-sample
8-sample
9-sample
Class
Max pooling / 1 noisy sample
Predicted query sample
2500
2000
1500
1000
1643
1865
2226
7-sample
8-sample
9-sample
Class
KNN pooling / 1 noisy sample
Predicted query sample
2500
2000
1500
1000
1209
1923
2546
7-sample
8-sample
9-sample
Class
Mean pooling / 4 noisy samples
Predicted query sample
2500
2000
1500
1000
1629
1927
2135
7-sample
8-sample
9-sample
Class
Max pooling / 4 noisy samples
Predicted query sample
2500
2000
1500
1000
1607
1929
2154
7-sample
8-sample
9-sample
Class
KNN pooling / 4 noisy samples
Predicted query sample
2500
2000
1500
1000
1174
1995
2500
7-sample
8-sample
9-sample
Class

Outline

- Introduction
- Method
- Experience
- Conclusion

Conclusion

- Propose MetricPrompt, which frees human labor from task-specific verbalizer design by reformulating few-shot text classification task into a text pair relevance estimation problem
- But they still suffer from prompting methods' susceptibility to the design of verbalizers