Metadata-Induced Contrastive Learning for Zero-Shot Multi-Label Text Classification

Yu Zhang[1], Zhihong Shen[2], Chieh-Han Wu[2], Boya Xie[2], Junheng Hao[3],
Ye-Yi Wang[2], Kuansan Wang[2], Jiawei Han[1]
[1]University of Illinois at Urbana-Champaign [2]Microsoft [3]University of California, Los Angeles
[1]{yuz9, hanj}@illinois.edu, [2]{zhihosh, chiewu, boxie, yeyiwang, kuansanw}@microsoft.com, [3]jhao@cs.ucla.edu

(WWW-2022)

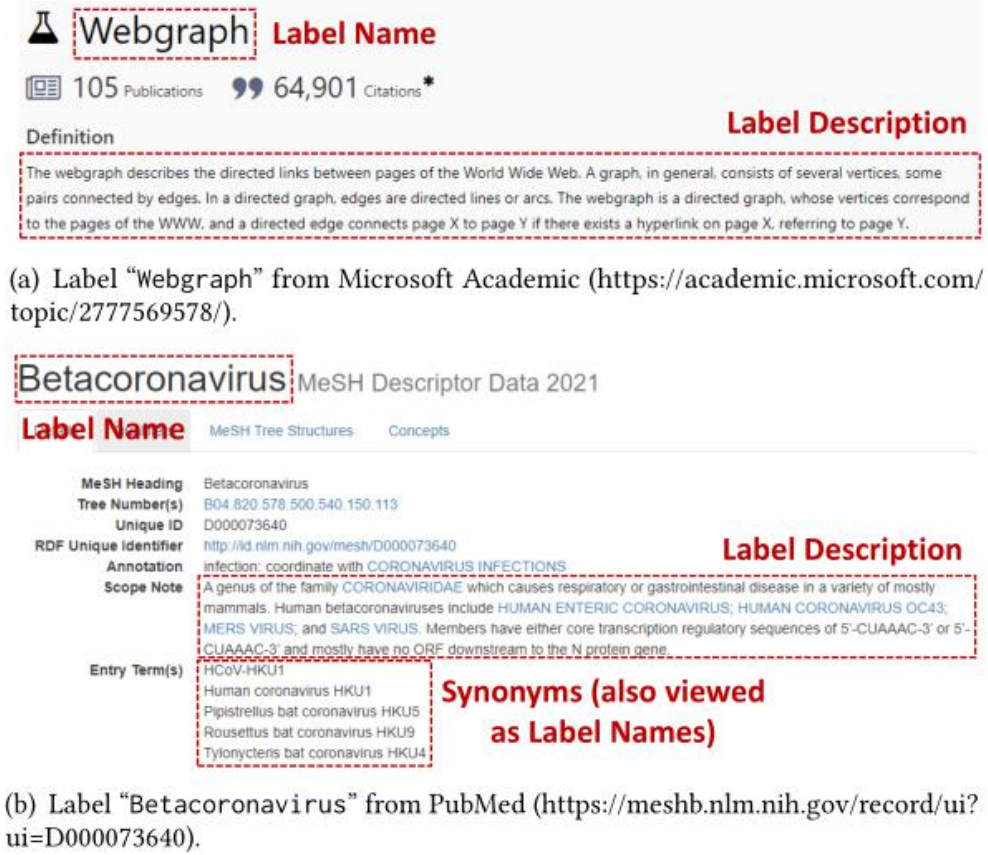
code :https://github.com/yuzhimanhua/MICoL.

1. Introduction

2. Approach

3. Experiments

Introduction



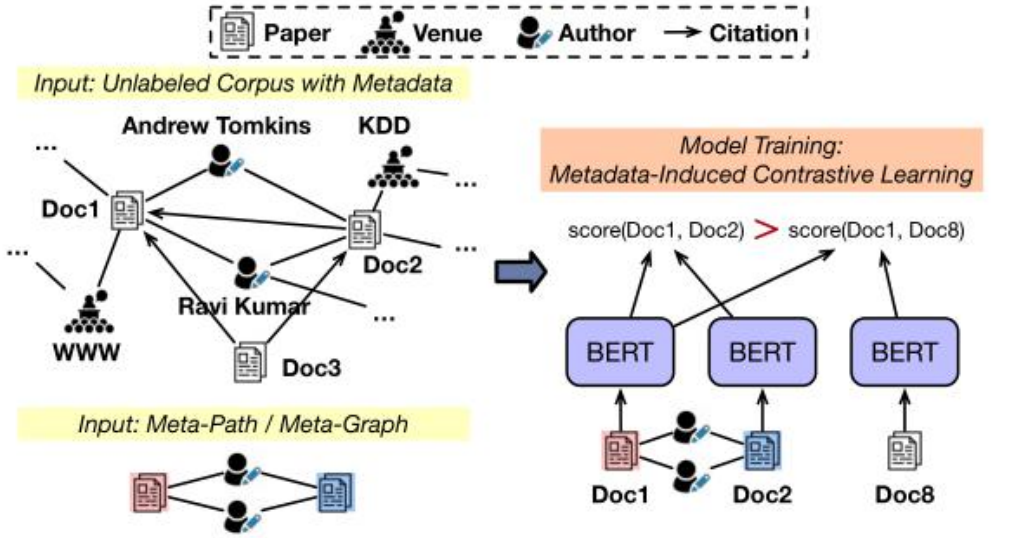
(a) Label "Webgraph" from Microsoft Academic (https://academic.microsoft.com/topic/2777569578/).

(b) Label "Betacoronavirus" from PubMed (https://meshb.nlm.nih.gov/record/ui?ui=D000073640).

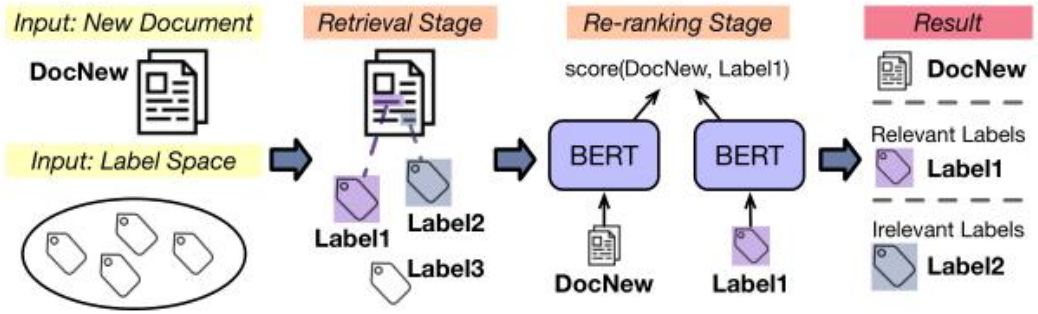
Figure 1: Two examples of labels with name(s) and description from Microsoft Academic [49] and PubMed [24].

- We propose a zero-shot LMTC framework that utilizes document metadata. The framework does not require any labeled training data and only relies on label surface names and descriptions during inference.
- We propose a novel metadata-induced contrastive learning method.Different from previous contrastive learning approaches which manipulate text only, we exploit metadata information to produce contrastive training pairs.
- We conduct extensive experiments on two large-scale datasets to demonstrate the effectiveness of the proposed MICoL framework.

Approach

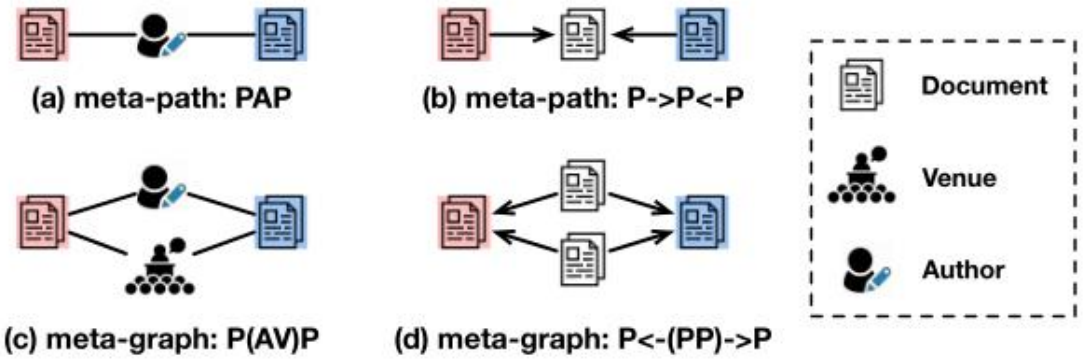


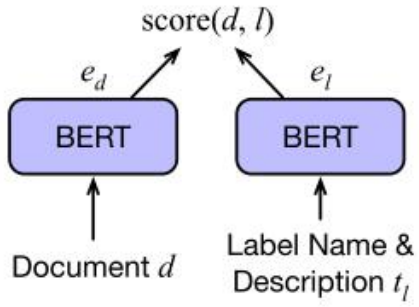
(a) Training phase of MICoL.



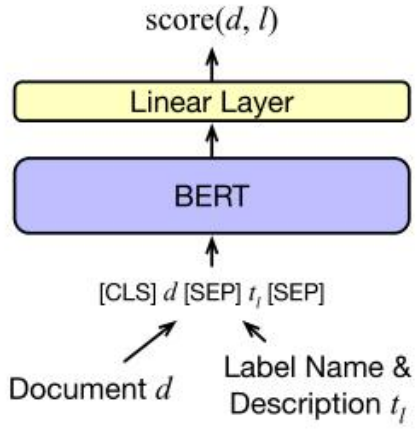
(b) Inference phase of MICoL.

Figure 2: Overview of the proposed MICoL framework.



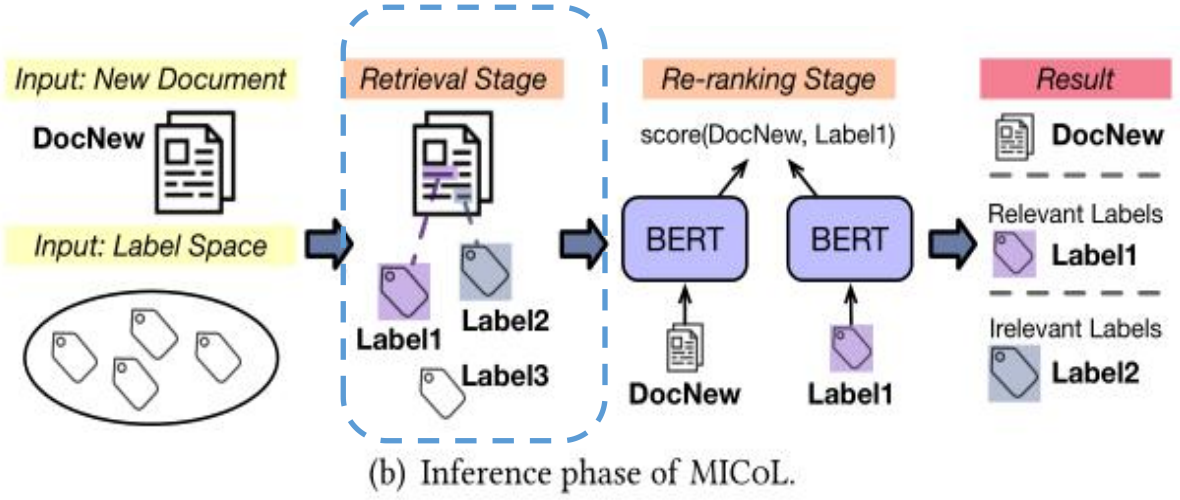


(a) Bi-Encoder



(b) Cross-Encoder

Approach



(b) Inference phase of MICoL.

$$\mathrm{BM25}(d,l) = \sum_{w \in d \cap t_l} \mathrm{IDF}(w) \frac{\mathrm{TF}(w,t_l)\cdot(k_1+1)}{\mathrm{TF}(w,t_l)\cdot k_1\left(1-b+b\frac{|\mathcal{L}|}{avgdl}\right)}. \quad (1)$$

$$avgdl = \frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}}|t_l|$$

$$C_{\mathrm{BM25}}(d) = \{l \mid l \in \mathcal{L},\ \mathrm{BM25}(d,l) > \eta\}. \quad (2)$$

$$C(d) = C_{\mathrm{exact}}(d) \cup C_{\mathrm{BM25}}(d)$$

Approach

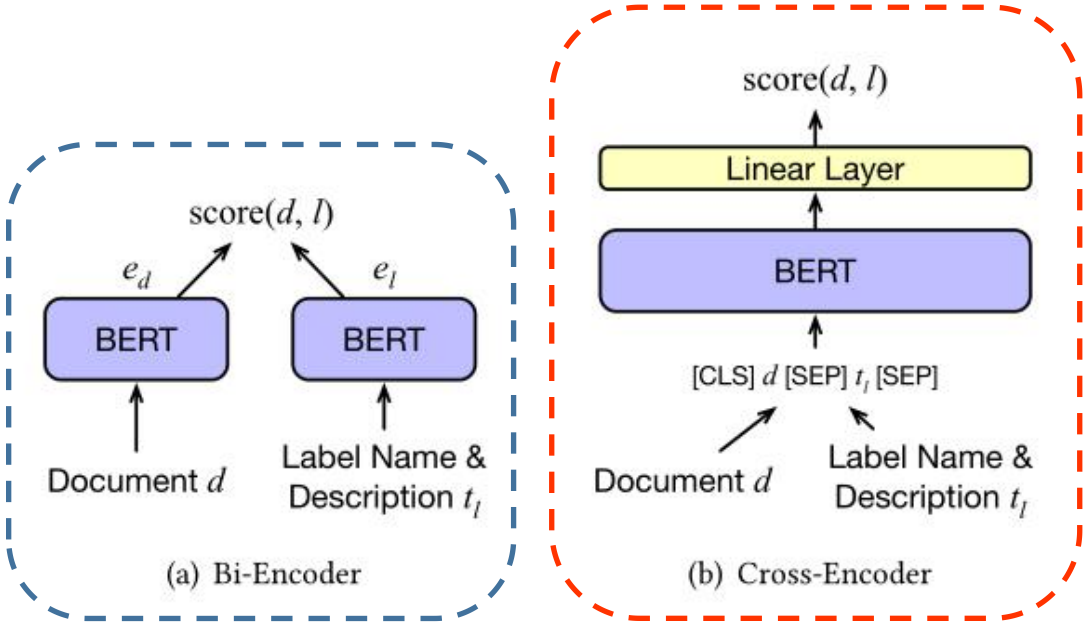


Figure 4: Two architectures that compute the similarity between a document d and a candidate label l.

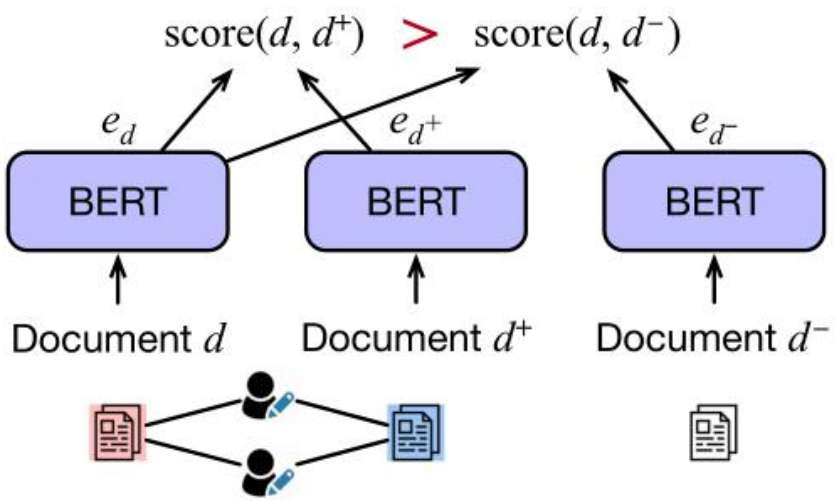
$$\boldsymbol{e}_d = \text{BERT}(d), \quad \boldsymbol{e}_l = \text{BERT}(t_l). \tag{3}$$

$$score(d, l) = \cos(\boldsymbol{e}_d, \boldsymbol{e}_l). \tag{4}$$

$$\boldsymbol{e}_{d||t_l} = \text{BERT}(d \,||\, t_l). \tag{5}$$

$$score(d, l) = \boldsymbol{w}^\top \boldsymbol{e}_{d||t_l}, \tag{6}$$

Approach



(a) Bi-Encoder fine-tuning

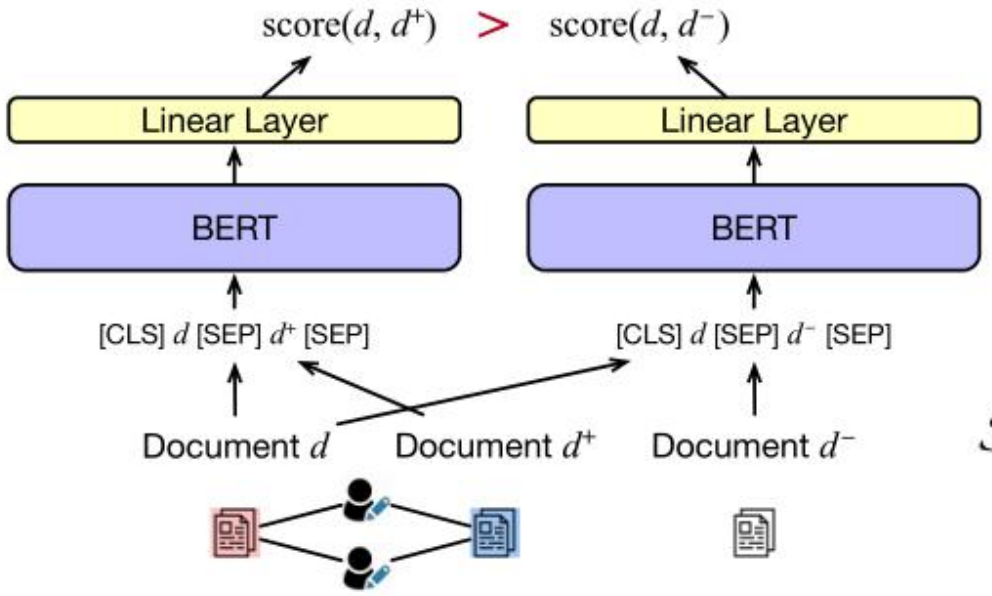
$$\boldsymbol{e}_d = \text{BERT}(d), \quad \boldsymbol{e}_{d^+} = \text{BERT}(d^+), \quad \boldsymbol{e}_{d_i^-} = \text{BERT}(d_i^-). \tag{7}$$

$$-\log \frac{\exp(\cos(\boldsymbol{e}_d, \boldsymbol{e}_{d^+})/\tau)}{\exp(\cos(\boldsymbol{e}_d, \boldsymbol{e}_{d^+})/\tau) + \sum_{i=1}^{N} \exp(\cos(\boldsymbol{e}_d, \boldsymbol{e}_{d_i^-})/\tau)}, \tag{8}$$

$$\mathcal{J}_{\text{Bi}} = \mathop{\mathbb{E}}_{\substack{d^+ \in N_M(d) \\ d_i^- \sim \mathcal{D}}} \left[-\log \frac{\exp(\cos(\boldsymbol{e}_d, \boldsymbol{e}_{d^+})/\tau)}{\exp(\cos(\boldsymbol{e}_d, \boldsymbol{e}_{d^+})/\tau) + \sum_{i=1}^{N} \exp(\cos(\boldsymbol{e}_d, \boldsymbol{e}_{d_i^-})/\tau)} \right]. \tag{9}$$

[1]Following previous studies [45, 47], we view $P_1 \rightarrow P_2 \leftarrow P_3$ as a "directed path" from P_1 to P_3 by explaining it as $P_1 \xrightarrow{\text{cites}} P_2 \xrightarrow{\text{is cited by}} P_3$. In this way, $P \rightarrow P \leftarrow P$ can be defined as a meta-path according to Definition 2.2. Similarly, we view PAP as a "directed path" $P \xrightarrow{\text{writes}} A \xrightarrow{\text{is written by}} P$. Using the same explanation, both $P(AV)P$ and $P \leftarrow (PP) \rightarrow P$ in Figure 3 can be viewed as a DAG, thus they are meta-graphs according to Definition 2.3.

Approach



(b) Cross-Encoder fine-tuning

$$\boldsymbol{e}_{d||d^+} = \text{BERT}(d \,||\, d^+), \qquad \boldsymbol{e}_{d||d_i^-} = \text{BERT}(d \,||\, d_i^-),$$
$$score(d, d^+) = \boldsymbol{w}^\top \boldsymbol{e}_{d||d^+}, \qquad score(d, d_i^-) = \boldsymbol{w}^\top \boldsymbol{e}_{d||d_i^-}. \tag{10}$$

$$\mathcal{J}_{\text{Cross}} = \mathop{\mathbb{E}}_{\substack{d^+ \in \mathcal{N}_{\mathcal{M}}(d) \\ d_i^- \sim \mathcal{D}}} \left[-\log \frac{\exp(score(d, d^+))}{\exp(score(d, d^+)) + \sum_{i=1}^{N} \exp(score(d, d_i^-))} \right]. \tag{11}$$

Experiments

Table 1: Dataset statistics.

Dataset	#Training (Unlabeled)	#Testing	#Labels	Labels/Doc	Words/Doc
MAG-CS [49]	634,874	70,533	15,808	5.59	126.55
PubMed [24]	808,692	89,854	17,963	7.80	199.14

Experiments

Table 2: P@k and NDCG@k scores of compared algorithms on MAG-CS and PubMed. Bold: the highest score of zero-shot approaches. *: MICoL (Cross-Encoder, $P \rightarrow P \leftarrow P$) is significantly better than this algorithm with p-value < 0.05. **: MICoL (Cross-Encoder, $P \rightarrow P \leftarrow P$) is significantly better than this algorithm with p-value < 0.01.

	Algorithm	MAG-CS [49]					PubMed [24]				
		P@1	P@3	P@5	NDCG@3	NDCG@5	P@1	P@3	P@5	NDCG@3	NDCG@5
Zero-shot	Doc2Vec [31]	0.5697**	0.4613**	0.3814**	0.5043**	0.4719**	0.3888**	0.3283**	0.2859**	0.3463**	0.3252**
	SciBERT [2]	0.6440**	0.5030**	0.4011**	0.5545**	0.5061**	0.4427**	0.3572**	0.3031**	0.3809**	0.3510**
	ZeroShot-Entail [61]	0.6649**	0.5003**	0.3959**	0.5570**	0.5057**	0.5275**	0.4021	**0.3299**	0.4352	**0.3913**
	SPECTER [8]	0.7107**	0.5381**	0.4184**	0.5979**	0.5365**	0.5286**	0.3923**	0.3181**	0.4273**	0.3815**
	EDA [53]	0.6442**	0.4939**	0.3948**	0.5471**	0.5000**	0.4919	0.3754*	0.3101*	0.4058*	0.3667*
	UDA [57]	0.6291**	0.4848**	0.3897**	0.5362**	0.4918**	0.4795**	0.3696**	0.3067**	0.3986**	0.3614**
	MICoL (Bi-Encoder, $P \rightarrow P \leftarrow P$)	0.7062*	0.5369*	0.4184*	0.5960*	0.5355*	0.5124**	0.3869*	0.3172*	0.4196*	0.3774*
	MICoL (Bi-Encoder, $P \leftarrow (PP) \rightarrow P$)	0.7050*	0.5344*	0.4161*	0.5937*	0.5331*	0.5198**	0.3876*	0.3172*	0.4215*	0.3786*
	MICoL (Cross-Encoder, $P \rightarrow P \leftarrow P$)	**0.7177**	**0.5444**	**0.4219**	**0.6048**	**0.5415**	**0.5412**	**0.4036**	0.3257	**0.4391**	0.3906
	MICoL (Cross-Encoder, $P \leftarrow (PP) \rightarrow P$)	0.7061	0.5376	0.4187	0.5964	0.5357	0.5218	0.3911	0.3172*	0.4249	0.3794
Supervised	MATCH [68] (10K Training)	0.4423**	0.2851**	0.2152**	0.3375**	0.3003**	0.6915	0.3869*	0.2785**	0.4649	0.3896
	MATCH [68] (50K Training)	0.6215**	0.4280**	0.3269**	0.4987**	0.4489**	0.7701	0.4716	0.3585	0.5497	0.4750
	MATCH [68] (100K Training)	0.8321	0.6520	0.5142	0.7342	0.6761	0.8286	0.5680	0.4410	0.6405	0.5626
	MATCH [68] (Full, 560K+ Training)	0.9114	0.7634	0.6312	0.8486	0.8076	0.9151	0.7425	0.6104	0.8001	0.7310

$$\text{P@}k = \frac{1}{k}\sum_{i=1}^{k} y_{d,\text{rank}(i)} \cdot$$

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{y_{d,\text{rank}(i)}}{\log(i+1)}, \quad \text{NDCG@}k = \frac{\text{DCG@}k}{\sum_{i=1}^{\min(k,||\boldsymbol{y}_d||_0)} \frac{1}{\log(i+1)}}.$$

Experiments

Table 3: PSP@k and PSN@k scores of compared algorithms on MAG-CS and PubMed. Bold, *, and **: the same meaning as in Table 2. We also show the ratio PSP@1/P@1. The higher PSP@k/P@k is, the more infrequent the correctly predicted labels are.

	Algorithm	MAG-CS [49]						PubMed [24]					
		PSP@1	PSP@3	PSP@5	PSN@3	PSN@5	$\frac{\text{PSP@1}}{\text{P@1}}$	PSP@1	PSP@3	PSP@5	PSN@3	PSN@5	$\frac{\text{PSP@1}}{\text{P@1}}$
Zero-shot	Doc2Vec [31]	0.4287**	0.4623**	0.4656**	0.4450**	0.4425**	0.75	0.2717**	0.2948**	0.3029**	0.2856**	0.2879**	0.70
	SciBERT [2]	0.4668**	0.4958**	0.4843**	0.4788**	0.4667**	0.72	0.3149**	0.3231**	0.3221**	0.3174**	0.3131**	0.71
	ZeroShot-Entail [61]	0.4796**	0.4892**	0.4759**	0.4777**	0.4644**	0.72	0.3617**	0.3498**	0.3389**	0.3492**	0.3378**	0.69
	SPECTER [8]	0.5304	0.5334*	0.5059*	0.5223	0.4988*	0.75	0.3907**	0.3638**	0.3442**	0.3666**	0.3489**	0.74
	EDA [53]	0.4916**	0.4968**	0.4821**	0.4859**	0.4708**	0.76	0.3572*	0.3451*	0.3334*	0.3442*	0.3322*	0.73
	UDA [57]	0.4850**	0.4907**	0.4771**	0.4797**	0.4654**	0.77	0.3547**	0.3423**	0.3311**	0.3416**	0.3298**	0.74
	MICoL (Bi-Encoder, $P \rightarrow P \leftarrow P$)	0.5176	0.5311	0.5065	0.5175	0.4963	0.73	0.3676**	0.3559**	0.3423*	0.3550**	0.3418**	0.72
	MICoL (Bi-Encoder, $P \leftarrow (PP) \rightarrow P$)	0.5160	0.5281	0.5037	0.5150	0.4940	0.73	0.3780**	0.3589*	0.3423*	0.3597**	0.3450**	0.73
	MICoL (Cross-Encoder, $P \rightarrow P \leftarrow P$)	**0.5375**	**0.5415**	**0.5118**	**0.5302**	**0.5052**	0.75	**0.4105**	**0.3807**	**0.3558**	**0.3841**	**0.3625**	0.76
	MICoL (Cross-Encoder, $P \leftarrow (PP) \rightarrow P$)	0.5326	0.5363	0.5087	0.5249	0.5013	0.75	0.3871	0.3664	0.3462	0.3677	0.3496	0.74
Supervised	MATCH [68] (10K Training)	0.1978**	0.1807**	0.1712**	0.1850**	0.1764**	0.45	0.2840**	0.2138**	0.1870**	0.2332**	0.2139**	0.41
	MATCH [68] (50K Training)	0.2854**	0.2830**	0.2738**	0.2838**	0.2780**	0.46	0.3201**	0.2715**	0.2532**	0.2848**	0.2713**	0.42
	MATCH [68] (100K Training)	0.4271**	0.4750**	0.4737*	0.4624**	0.4635**	0.51	0.3576**	0.3579**	0.3456*	0.3584**	0.3507**	0.43
	MATCH [68] (200K Training)	0.4695**	0.5401	0.5530	0.5217	0.5325	0.54	0.3732**	0.3988	0.3905	0.3913	0.3882	0.44
	MATCH [68] (Full, 560K+ Training)	0.5501	0.6397	0.6627	0.6171	0.6345	0.60	0.4371	0.5188	0.5200	0.4978	0.5011	0.48

$$\frac{1}{p_l} = 1 + C(N_l + B)^{-A}, \quad \text{PSP@}k = \frac{1}{k}\sum_{i=1}^{k}\frac{y_{d,\text{rank}(i)}}{p_{d,\text{rank}(i)}}. \qquad \text{PSDCG@}k = \sum_{i=1}^{k}\frac{y_{d,\text{rank}(i)}}{p_{d,\text{rank}(i)}\log(i+1)}, \quad \text{PSN@}k = \frac{\text{PSDCG@}k}{\sum_{i=1}^{\min(k,\|\boldsymbol{y}_d\|_0)}\frac{1}{\log(i+1)}}.$$

Experiments

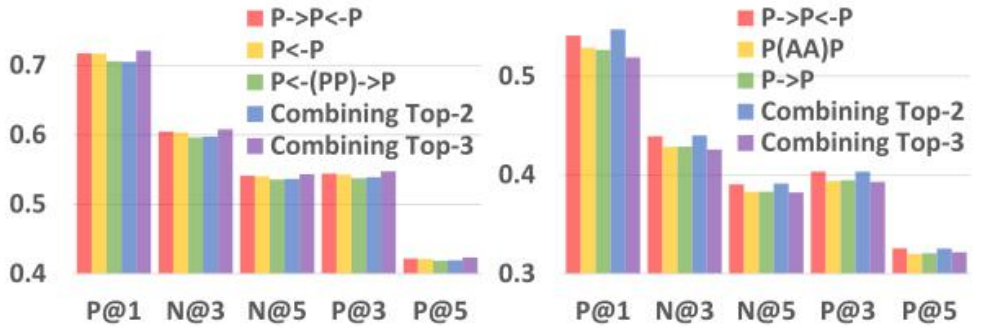
Table 4: P@k and NDCG@k scores of MICoL using different meta-paths/meta-graphs. Bold: the best model. *: significantly worse than the best model with p-value < 0.05. **: significantly worse than the best model with p-value < 0.01. All meta-paths and meta-graphs, except PVP, can improve the classification performance upon unfine-tuned SciBERT.

Algorithm	MAG-CS [49]					PubMed [24]				
	P@1	P@3	P@5	NDCG@3	NDCG@5	P@1	P@3	P@5	NDCG@3	NDCG@5
Unfine-tuned SciBERT	0.6599**	0.5117**	0.4056**	0.5651**	0.5136**	0.4371**	0.3544**	0.3014**	0.3775**	0.3485**
MICoL (Bi-Encoder, PAP)	0.6877**	0.5285**	0.4143**	0.5852**	0.5280**	0.4974**	0.3818**	0.3154*	0.4122**	0.3727**
MICoL (Bi-Encoder, PVP)	0.6589**	0.5123**	0.4063**	0.5656**	0.5145**	0.4440**	0.3507**	0.2966**	0.3761**	0.3458**
MICoL (Bi-Encoder, $P \rightarrow P$)	0.7094	0.5391	0.4190	0.5982	0.5367	0.5200*	0.3903*	0.3195	0.4240*	0.3808*
MICoL (Bi-Encoder, $P \leftarrow P$)	0.7095*	0.5374*	0.4178*	0.5970*	0.5356*	0.5195**	0.3905*	0.3192	0.4240*	0.3806*
MICoL (Bi-Encoder, $P \rightarrow P \leftarrow P$)	0.7062*	0.5369*	0.4184*	0.5960*	0.5355*	0.5124**	0.3869*	0.3172*	0.4196*	0.3774*
MICoL (Bi-Encoder, $P \leftarrow P \rightarrow P$)	0.7039*	0.5379*	0.4187*	0.5963*	0.5356*	0.5174**	0.3886*	0.3187*	0.4220*	0.3795*
MICoL (Bi-Encoder, $P(AA)P$)	0.6873**	0.5272**	0.4130**	0.5840**	0.5269**	0.4963**	0.3794**	0.3139**	0.4101**	0.3711**
MICoL (Bi-Encoder, $P(AV)P$)	0.6832**	0.5263**	0.4135**	0.5823**	0.5263**	0.4894**	0.3743**	0.3099**	0.4045**	0.3664**
MICoL (Bi-Encoder, $P \rightarrow (PP) \leftarrow P$)	0.7015**	0.5334**	0.4160**	0.5920**	0.5322**	0.5163**	0.3879*	0.3172*	0.4211*	0.3781*
MICoL (Bi-Encoder, $P \leftarrow (PP) \rightarrow P$)	0.7050*	0.5344*	0.4161*	0.5937*	0.5331*	0.5198**	0.3876*	0.3172*	0.4215*	0.3786*
MICoL (Cross-Encoder, PAP)	0.7034*	0.5355	0.4168	0.5943	0.5337	0.5212**	0.3921*	0.3207	0.4255*	0.3818*
MICoL (Cross-Encoder, PVP)	0.6720*	0.5203*	0.4103*	0.5750*	0.5210*	0.4668**	0.3633**	0.3051**	0.3908**	0.3574**
MICoL (Cross-Encoder, $P \rightarrow P$)	0.7033*	0.5391	0.4201	0.5971*	0.5365*	0.5266	0.3946	0.3207	0.4286	0.3830
MICoL (Cross-Encoder, $P \leftarrow P$)	0.7169	0.5430	0.4214	0.6033	0.5406	0.5265	0.3924	0.3186	0.4268	0.3811
MICoL (Cross-Encoder, $P \rightarrow P \leftarrow P$)	**0.7177**	**0.5444**	**0.4219**	**0.6048**	**0.5415**	**0.5412**	**0.4036**	**0.3257**	**0.4391**	**0.3906**
MICoL (Cross-Encoder, $P \leftarrow P \rightarrow P$)	0.7045	0.5356*	0.4168*	0.5944*	0.5336*	0.5243*	0.3932*	0.3190*	0.4271*	0.3814*
MICoL (Cross-Encoder, $P(AA)P$)	0.7028	0.5351	0.4171	0.5939	0.5338	0.5290*	0.3937	0.3201	0.4285*	0.3830
MICoL (Cross-Encoder, $P(AV)P$)	0.7024*	0.5354*	0.4177	0.5940*	0.5343*	0.5164**	0.3897*	0.3195*	0.4225*	0.3797*
MICoL (Cross-Encoder, $P \rightarrow (PP) \leftarrow P$)	0.7076*	0.5379*	0.4188	0.5971*	0.5363*	0.5186	0.3924*	0.3184*	0.4254*	0.3800*
MICoL (Cross-Encoder, $P \leftarrow (PP) \rightarrow P$)	0.7061	0.5376	0.4187	0.5964	0.5357	0.5218	0.3911	0.3172*	0.4249	0.3794

Experiments

Table 5: Metadata-related statistics of the training corpus.

Dataset	MAG-CS [49]	PubMed [24]
# Authors	762,259	2,068,411
# Author–Paper Edges	2,047,166	5,391,314
# Venues	105	150
# Venue–Paper Edges	634,874	808,692
# Paper→Paper Edges	1,219,234	3,615,220



(a) MAG-CS, Cross-Encoder (b) PubMed, Cross-Encoder

Figure 6: P@k and NDCG@k scores of top-3 performing meta-paths/meta-graphs and the *combination* of them.

Thank you !