Knowledge-Aware Meta-learning for Low-Resource Text Classification

Huaxiu Yao[1], Yingxin Wu[2], Maruan Al-Shedivat[4], Eric P. Xing[3,4]

[1]Stanford University, [2]University of Science and Technology of China
[3]MBZUAI, [4]Carnegie Mellon University

Background: Low-resource Text classification

Training data

Politics

Ireland Votes To Repeal Abortion Amendment In Landslide Referendum

Booyah: Obama Photographer Hilariously Trolls Trump's

Entertainments

Jim Carrey Blasts 'Castrato' Adam Schiff And Democrats In New Artwork

Morgan Freeman 'Devastated' That Sexual Harassment Claims Could Undermine Legacy

Test data

Bishop Michael Curry Joins Christian March To White House To 'Reclaim Jesus

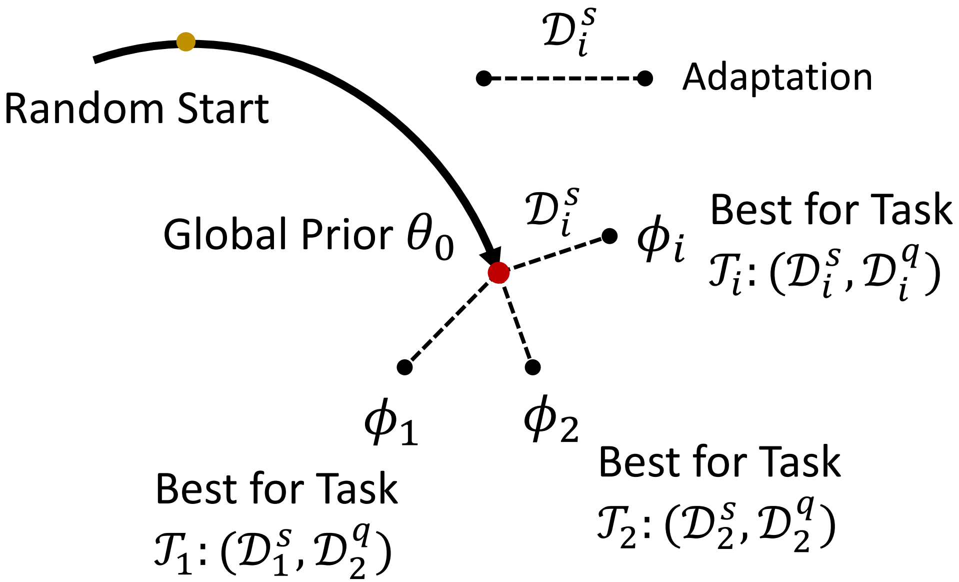
By Politics or Entertainments?

Meta-learning

Task $\mathcal{T}_i$: data $\mathcal{D}_i$; support $\mathcal{D}_i^s$/query set $\mathcal{D}_i^q$ sampled from $\mathcal{D}_i$

ML model: f with initial parameter θ_0

Supervised Adaptation (MAML) [Finn et al. 2017]



Unsupervised Adaptation (ARM) [Zhang et al. 2021]

Sentence Representation

$$f_{\theta^B}\left(x_{i,j}^q\right)$$

BERT parameters

Task Representation

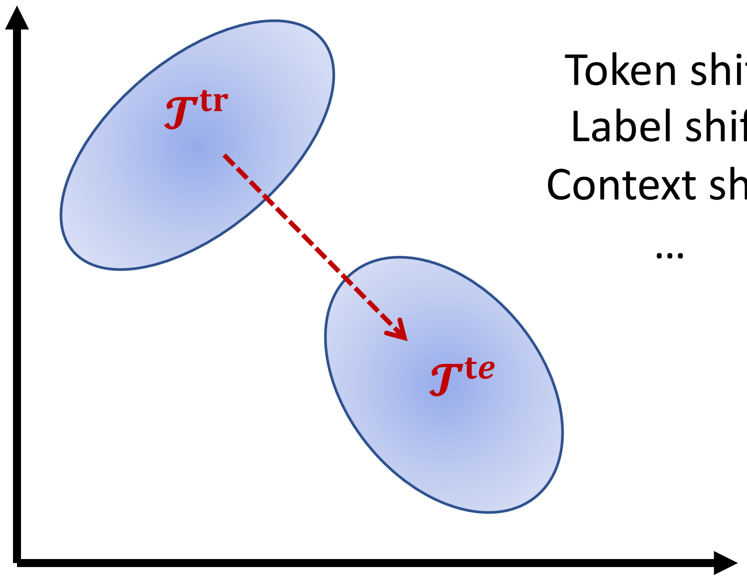
$$c_i = \frac{1}{N^q}\sum_{j=1}^{N^q} f_{\theta^B}\left(x_{i,j}^q\right)$$

$$\theta_\star^B, \theta_\star^c \leftarrow \min_{\theta^B, \theta^c} \frac{1}{n}\sum_i \mathcal{L}\left(f_{\theta^B, \theta^c}; \mathcal{D}_i^q, c_i\right)$$

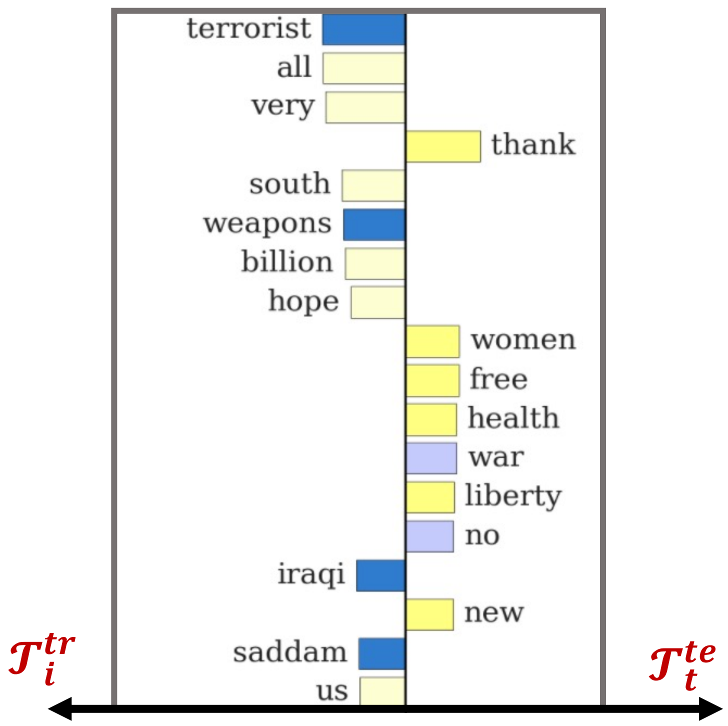
Task-specific parameters

Distribution Shifts between Training and Test Tasks

High-dimensional Distribution



Token shift
Label shift
Context shift
…



Observation-driven: $f = argmin \frac{1}{n}\sum_{i=1}^{n} l(f(x_i), y_i)$

Training task: "Cool weapons and billion ships in the scene"

?

Test task: "demonstrates the liberty of woman"

Can not generalize well to unseen test tasks.

How to Connect Training and Testing Tasks?

Knowledges shared by training and testing tasks

Tobacco use -> Heart failure

Observations

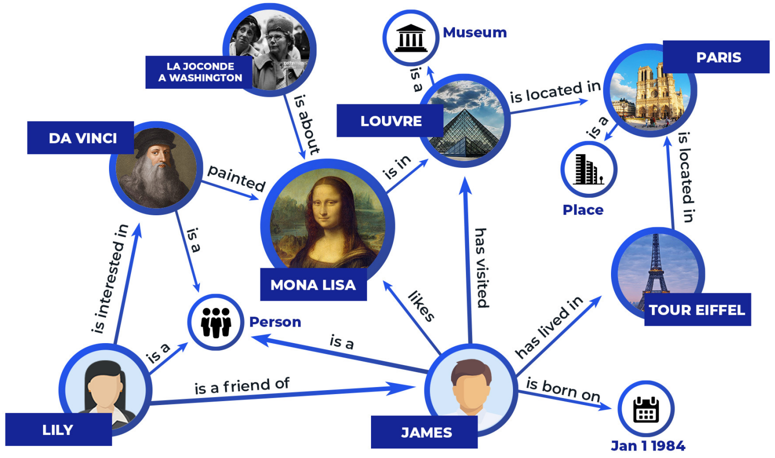
Stefan-Boltzmann Law

$$\frac{P}{A} = e\sigma T^4$$

Convection-Diffusion Eqn.

$$\frac{\partial c}{\partial t} = \nabla \cdot (D \nabla c) - \nabla \cdot (\mathbf{v} c) + R$$

Physics Rules



Knowledge Graph

Knowledge-aware Meta-learning (KGML)

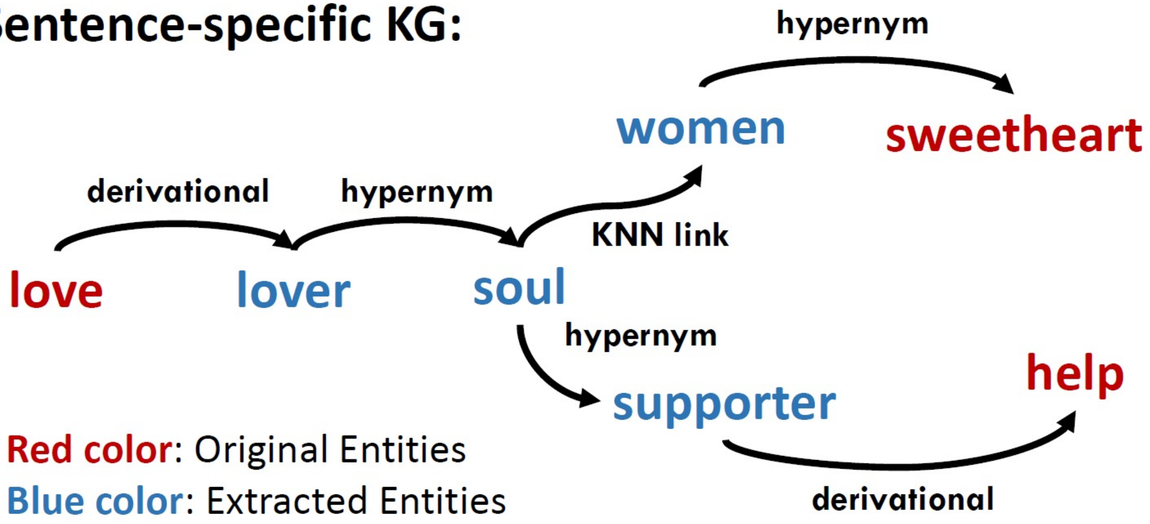
--- Knowledge Extraction & Representation Learning

$\boldsymbol{x_{i,j}}$ **Entity 1**   **Entity 2**   **Entity 3**

Sentence: Thank you for the **love** and **help**, **sweetheart**

Text Encoder → Sentence Representation $f_{\theta^B}(x_{i,j})$

Sentence-specific KG:



GNN → Graph Representation $g_{i,j}$

$$\mathrm{KG} = \mathcal{G}^{knn} \cup \mathcal{G}^{base}$$

Reduce inconnectivity and create rich context

$$\text{Sentence-specific KG} = Extract(\mathrm{KG})$$

Knowledge-aware Meta-learning (KGML)

--- Knowledge Fusion and Overall Framework

Knowledge-aware Representation

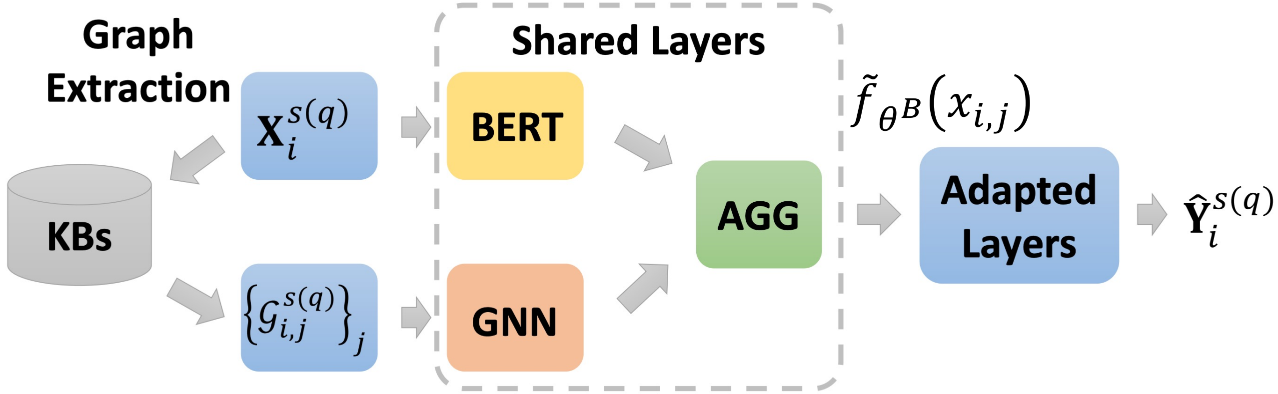
Sentence representation

$$\tilde{f}_{\theta^B}(x_{i,j}) = AGG_{kf}(f_{\theta^B}(x_{i,j}), g_{i,j})$$

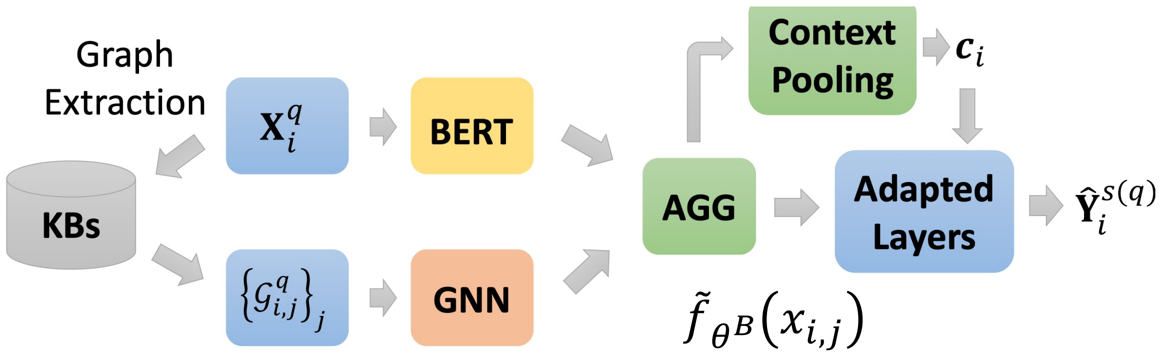
Aggregator

Graph representation

Supervised Adaptation



Unsupervised Adaptation



Empirical Comparison

- Supervised Adaptation
  - Backbone: MAML, Prototypical Network
  - Data
    - Amazon Review – classifier the category of each review
    - Huffpost – classifier the headlines of News

- Unsupervised Adaptation
  - Backbone: Adaptive Risk Minimization (ARM)
  - Data
    - Twitter – federated sentiment classification

Results

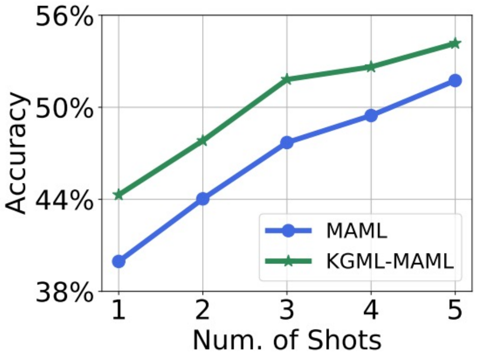
	Supervised Adaptation				Unsupervised Adaptation		
Data	Amazon Review		Huffpost		Data	Twitter	
Shot	1-shot	5-shot	1-shot	5-shot	User Ratio	0.6	1.0
MAML	44.35%	56.94%	39.95%	51.74%	ERM	62.91%	66.05%
ProtoNet	55.32%	73.30%	41.72%	57.53%	UW	63.51%	64.13%
InductNet	45.35%	56.73%	41.35%	55.96%	ARM	60.42%	60.42%
MatchingNet	51.16%	69.89%	41.18%	54.41%	DRNN	63.02%	64.02%
REGRAB	55.07%	72.53%	42.17%	57.66%	-	-	-
KGML-MAML	51.44%	58.81%	**44.29%**	54.16%	**KGML**	**64.92%**	**67.00%**
KGML-ProtoNet	**58.62%**	**74.55%**	42.37%	**58.75%**	-	-	-

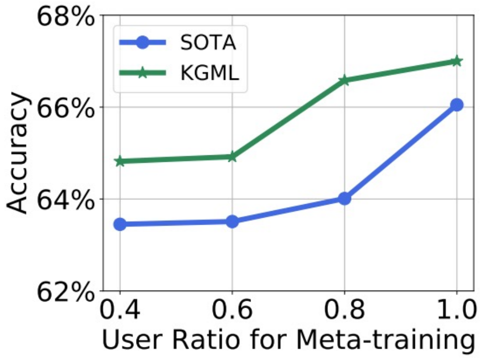
Analysis

Ablation

Ablations	Backbone	Amazon	Huffpost
I. Remove AGG_{kf}	MAML	45.68%	41.55%
	ProtoNet	57.94%	41.71%
II. Remove KNN	MAML	51.07%	41.20%
	ProtoNet	57.80%	41.91%
KGML	MAML	51.44%	**44.29%**
KGML	ProtoNet	**58.62%**	42.37%

Robustness w/ different settings





(a) Huffpost (b) Twitter

SOTA: best baseline

Takeaways & Next

- Bridging training and testing tasks can alleviate the effects of task distribution shifts

- Knowledge graph is a useful domain knowledge to connect training and testing tasks

- What’s Next?
  - Apply KGML to more few-shot NLP tasks

  - More complex few-shot scenarios (e.g., heterogeneous tasks)

Thanks

Q & A