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

Huaxiu Yao¹, Yingxin Wu², Maruan Al-Shedivat⁴, Eric P. Xing³,⁴

¹Stanford University, ²University of Science and Technology of China
³MBZUAI, ⁴Carnegie Mellon University
Background: Low-resource Text classification

<table>
<thead>
<tr>
<th>Training data</th>
<th>Test data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politics</td>
<td></td>
</tr>
<tr>
<td>Ireland Votes To Repeal Abortion Amendment In Landslide Referendum</td>
<td>Bishop Michael Curry Joins Christian March To White House To 'Reclaim Jesus</td>
</tr>
<tr>
<td>Booyah: Obama Photographer Hilariously Trolls Trump's</td>
<td></td>
</tr>
<tr>
<td>Entertainments</td>
<td>By Politics or Entertainments?</td>
</tr>
<tr>
<td>Jim Carrey Blasts 'Castrato' Adam Schiff And Democrats In New Artwork</td>
<td></td>
</tr>
<tr>
<td>Morgan Freeman 'Devastated' That Sexual Harassment Claims Could Undermine Legacy</td>
<td></td>
</tr>
</tbody>
</table>
Meta-learning

Task $T_i$: 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 $\theta_0$

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

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

Sentences Representation

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

Task Representation

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

BERT parameters

$$\theta^B_*, \theta^c_* \leftarrow \min_{\theta^B, \theta^c} \frac{1}{n} \sum_i \mathcal{L}(f_{\theta^B, \theta^c}; \mathcal{D}_i^q, c_i)$$

Task-specific parameters
Distribution Shifts between Training and Test Tasks

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

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

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

Physics Rules

Convection-Diffusion Eqn.

\[ \frac{\partial c}{\partial t} = \nabla \cdot (D\nabla c) - \nabla \cdot (vc) + R \]

Knowledge Graph
Knowledge-aware Meta-learning (KGML)

--- Knowledge Extraction & Representation Learning

Text Encoder

Sentence Representation $f_{θ^B}(x_{i,j})$

GNN

Graph Representation $g_{i,j}$

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

Sentence-specific KG:

- **love**
- **lover**
- **soul**
- **women**
- **sweetheart**
- **supporter**

Red color: Original Entities
Blue color: Extracted Entities

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

Reduce inconnectivity and create rich context

Sentence-specific KG = Extract(KG)
Knowledge-aware Meta-learning (KGML)

--- Knowledge Fusion and Overall Framework

Supervised Adaptation

Unsupervised Adaptation

Knowledge-aware Representation

Sentence representation

Aggregator

Graph representation

Aggregator

Knowledge-aware Representation

\[ \tilde{f}_{\theta_B}(x_{i,j}) = AGK_{\theta_f}(f_{\theta_B}(x_{i,j}), g_{i,j}) \]
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

<table>
<thead>
<tr>
<th>Data Shot</th>
<th>Supervised Adaptation</th>
<th></th>
<th>Unsupervised Adaptation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Amazon Review</td>
<td>HuffPost</td>
<td>Data User Ratio</td>
<td>Twitter</td>
</tr>
<tr>
<td></td>
<td>1-shot</td>
<td>5-shot</td>
<td>1-shot</td>
<td>5-shot</td>
</tr>
<tr>
<td>MAML</td>
<td>44.35%</td>
<td>56.94%</td>
<td>39.95%</td>
<td>51.74%</td>
</tr>
<tr>
<td>ProtoNet</td>
<td>55.32%</td>
<td>73.30%</td>
<td>41.72%</td>
<td>57.53%</td>
</tr>
<tr>
<td>InductNet</td>
<td>45.35%</td>
<td>56.73%</td>
<td>41.35%</td>
<td>55.96%</td>
</tr>
<tr>
<td>MatchingNet</td>
<td>51.16%</td>
<td>69.89%</td>
<td>41.18%</td>
<td>54.41%</td>
</tr>
<tr>
<td>REGRAB</td>
<td>55.07%</td>
<td>72.53%</td>
<td>42.17%</td>
<td>57.66%</td>
</tr>
<tr>
<td>KGML-MAML</td>
<td>51.44%</td>
<td>58.81%</td>
<td><strong>44.29%</strong></td>
<td>54.16%</td>
</tr>
<tr>
<td>KGML-ProtoNet</td>
<td><strong>58.62%</strong></td>
<td><strong>74.55%</strong></td>
<td>42.37%</td>
<td><strong>58.75%</strong></td>
</tr>
</tbody>
</table>
## Analysis

### Ablation

<table>
<thead>
<tr>
<th>Ablations</th>
<th>Backbone</th>
<th>Amazon</th>
<th>Huffpost</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Remove $\text{AGG}_{k_f}$</td>
<td>MAML</td>
<td>45.68%</td>
<td>41.55%</td>
</tr>
<tr>
<td></td>
<td>ProtoNet</td>
<td>57.94%</td>
<td>41.71%</td>
</tr>
<tr>
<td>II. Remove KNN</td>
<td>MAML</td>
<td>51.07%</td>
<td>41.20%</td>
</tr>
<tr>
<td></td>
<td>ProtoNet</td>
<td>57.80%</td>
<td>41.91%</td>
</tr>
<tr>
<td>KGML</td>
<td>MAML</td>
<td>51.44%</td>
<td>44.29%</td>
</tr>
<tr>
<td>KGML</td>
<td>ProtoNet</td>
<td>58.62%</td>
<td>42.37%</td>
</tr>
</tbody>
</table>

### Robustness w/ different settings

![Graphs](image_url)
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