Weakly Supervised Data Augmentation through Prompting for Dialogue Understanding
Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Andy Rosenbaum, Seokhwan Kim, Yang Liu, Zhou Yu, Dilek Hakkani-Tur

**Background & Challenges**

High quality conversational data is scarce.
- **Data augmentation** techniques can be used to expand limited training data
- Prompting large pre-trained language models appears to be a viable method of generating coherent synthetic dialogue turns
- But: uncontrolled generation means no guarantee on utterance “correctness”

(neutral) Alice: You’re going to set up your own law office, aren’t you?
(neutral) Bob: Yes. After so many years of hard work, I’d rather I had an office of my own.
(happy) Alice: If you need help, don’t hesitate to ask me.
(happy) Bob: I’ll be very glad if you would help.
(happy) Alice: I’d like to wish you every success in your new venture.
(happy) Bob: Thank you. I wish I would.
(happy) Alice:

Generated Responses:
1. Good luck to you. Let’s do lunch soon, Bob.
2. It’s such a rare pleasure to meet such an ideal partner in your work.
3. You know, you seem quite different.

**Refining Data with Weak Supervision**

WeakDAP workflow:
- Prefix prompt to generate synthetic turns
- Weakly classify each prospective synthetic instance’s label, rejecting silver data if mismatching desired label with high confidence
- Repeat until convergence

![WeakDAP Workflow Diagram](image)

**Results: Emotion Classification**

- Varied seed and synthetic data sizes in DailyDialog emotion classification task
- Prompting yields large improvements at 2x

![Emotion Classification Graph](image)

**Results: Cross-lingual Intent Detection**

<table>
<thead>
<tr>
<th>FBTOD ES</th>
<th>( Acc_{LR} )</th>
<th>( F1_{LR} )</th>
<th>( Acc_{HR} )</th>
<th>( F1_{HR} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1% No Aug.</td>
<td>0.572</td>
<td>0.164</td>
<td>0.572</td>
<td>0.164</td>
</tr>
<tr>
<td>1% Prompt</td>
<td>0.737</td>
<td>0.316</td>
<td>0.776</td>
<td>0.359</td>
</tr>
<tr>
<td>+ WeakDAP</td>
<td><strong>0.834</strong></td>
<td><strong>0.495</strong></td>
<td><strong>0.831</strong></td>
<td><strong>0.528</strong></td>
</tr>
<tr>
<td>5% No Aug.</td>
<td>0.845</td>
<td>0.417</td>
<td>0.845</td>
<td>0.417</td>
</tr>
<tr>
<td>5% Prompt</td>
<td>0.953</td>
<td>0.641</td>
<td>0.954</td>
<td>0.682</td>
</tr>
<tr>
<td>+ WeakDAP</td>
<td><strong>0.957</strong></td>
<td><strong>0.715</strong></td>
<td><strong>0.962</strong></td>
<td><strong>0.732</strong></td>
</tr>
<tr>
<td>10% No Aug.</td>
<td>0.942</td>
<td>0.588</td>
<td>0.942</td>
<td>0.588</td>
</tr>
<tr>
<td>10% Prompt</td>
<td>0.973</td>
<td>0.772</td>
<td>0.973</td>
<td>0.791</td>
</tr>
<tr>
<td>+ WeakDAP</td>
<td><strong>0.979</strong></td>
<td><strong>0.905</strong></td>
<td><strong>0.976</strong></td>
<td><strong>0.846</strong></td>
</tr>
<tr>
<td>100% No Aug.</td>
<td>0.988</td>
<td>0.889</td>
<td>0.988</td>
<td>0.889</td>
</tr>
</tbody>
</table>

- Few-shot Spanish Intent Detection - FBTOD
- Augmentation using mixed Spanish and English in-context examples
- HR/LR = High/Low-Resource English in-context example pool

Link to paper: [WeakDAP](#)

WeakDAP can be adapted to any large pre-trained large language model, and any classifier for a given dialogue understanding task