**INTRODUCTION**

Language Models (LMs) for Automatic Speech Recognition (ASR) can benefit from utilizing non-linguistic contextual signals such as application (app) ids.

The vast majority of speech queries lack annotations of such signals, making it challenging to directly train domain-specific LMs.

We propose three domain adaptation schemes to improve the domain-level performance of Long Short-Term Memory (LSTM) LMs in pre-training & fine-tuning stages.

**DATA & BASELINES**

Three utterance-level app signals:
- Google Maps (27.6%)
- Google PlayStore (24.4%)
- YouTube (48.0%)

<table>
<thead>
<tr>
<th>Sets</th>
<th>#(utts))</th>
<th>#(words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOD-Train</td>
<td>257M</td>
<td>1.6B</td>
</tr>
<tr>
<td>OOD-Dev</td>
<td>53K</td>
<td>343K</td>
</tr>
<tr>
<td>DOM-Train</td>
<td>23M</td>
<td>109M</td>
</tr>
<tr>
<td>DOM-Dev</td>
<td>6.8M</td>
<td>31M</td>
</tr>
</tbody>
</table>

Distribution of out-of-domain/domain data

Baseline: 2-layer * 1024-node LSTM LMs, each word is embedded in 1024 dimensions

- One-pass baselines: Train the LMs on OOD-Train only or DOM-Train only
- Fine-tuning baseline: Pre-train on OOD-Train, fine-tune models on DOM-Train

**SCHME I: PREPEND APP ID**

- Initialize LSTM state by contextual app id
- Learn app-embeddings by back-propagation
- 17.0% rel. reduction in domain perplexity

**ADAPTATION STRATEGY**

- Fine-tuning Baseline: Pre-train on OOD-Train only
- Adaptation Baseline: Pre-train on OOD-Train and OOD-Dev

**SCHME II: ADD META-MEMORY**

- Add new layer to embed the app id
- META-MEMORY: Apply affine transformations to the app embedding a
- Add META-MEMORY to LSTM cells (omitting the biases)

\[
\begin{align*}
    f_t &= \sigma(W_{fx}x_t + W_{fh}\hat{h}_{t-1} + W_{fc}c_{t-1} + W_{fa}a) \\
    i_t &= 1 - f_t \\
    c_t &= tanh(W_{cx}x_t + W_{ch}\hat{h}_{t-1} + W_{ca}a) \\
    c_t &= f_t \odot c_{t-1} + i_t \odot c_t \\
    h_t &= \sigma(W_{ax}x_t + W_{ah}\hat{h}_{t-1} + W_{ac}c_t + W_{oa}a) \\
    \hat{h}_t &= c_t \odot tanh(c_t)
\end{align*}
\]

Freezing variants: freeze word embedding layer (and the original LSTM parameters LSTMs*)

Cand variant: only include META-MEMORY in the computation of cell state candidate \(\hat{c}_t\)

**ADAPTATION STRATEGY**

<table>
<thead>
<tr>
<th>Adaptation Strategy</th>
<th>DOM</th>
<th>OOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background LM (OOD Baseline)</td>
<td>82</td>
<td>85</td>
</tr>
<tr>
<td>Adaptation Scheme I</td>
<td>37</td>
<td>602</td>
</tr>
<tr>
<td>Sch II, freeze Embed</td>
<td>41</td>
<td>176</td>
</tr>
<tr>
<td>Sch II, freeze Embed, LSTMs*</td>
<td>44</td>
<td>139</td>
</tr>
<tr>
<td>Cand Variant of Sch II</td>
<td>39</td>
<td>276</td>
</tr>
<tr>
<td>Cand, freeze Embed</td>
<td>48</td>
<td>177</td>
</tr>
<tr>
<td>Cand, freeze Embed, LSTMs*</td>
<td>51</td>
<td>123</td>
</tr>
</tbody>
</table>

**SCHME III: LEARN DOMAIN-PARAMETERS IN ADAPTATION PHASE**

Categorize the model parameters into two sets [2]:
- Pre-training: tune general (non-domain) parameters: Word Embed, LSTMs, and \(W_{OOD,bOOD}\)
- Adaptation: freeze general parameters and tune domain parameters: appEmbed, DNNAdapt, and \(W_D,b_D\)

\[
P(w_t|\text{hist}) = \begin{cases} 
\phi(W_{OOD}h_t + b_{OOD}) \\
\phi(W_{OOD}h_t + b_{OOD} + W_{D}h_t + b_D) 
\end{cases}
\]

Variants of Scheme III: relax freezing constraints to fine-tune \(W_{OOD}\) and \(b_{OOD}\), but multiply their gradients by a factor

**ADAPTATION STRATEGY**

<table>
<thead>
<tr>
<th>Adaptation Strategy</th>
<th>DOM</th>
<th>OOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background LM (OOD Baseline)</td>
<td>82</td>
<td>85</td>
</tr>
<tr>
<td>Adaptation Scheme I</td>
<td>75</td>
<td>85</td>
</tr>
<tr>
<td>Variant 1 (mul 0.25)</td>
<td>62</td>
<td>132</td>
</tr>
<tr>
<td>Variant 2 (mul 0.50)</td>
<td>61</td>
<td>152</td>
</tr>
<tr>
<td>Variant 3 (mul 0.75)</td>
<td>60</td>
<td>152</td>
</tr>
<tr>
<td>Variant 4 (mul 1.00)</td>
<td>51</td>
<td>133</td>
</tr>
</tbody>
</table>

**ASR RESULTS & CONCLUSIONS**

- Adding contextual signals to LSTM LM reduces domain perplexity by 21% relative
- 3% relative reduction in WER on top of an unadapted 5-gram LM
- SxS experiments show significant improvements on sub-domains
- Grouping model parameters into two sets suggests a possible solution to catastrophic forgetting

**REFERENCES**