**Overview**

- Good translation quality requires lots of parallel training data
- Only small datasets may be available in some domains
- Fine tuning
  - Train on a large out-of-domain dataset first
  - Continue training on a small in-domain dataset
  - How do we avoid overfitting to the in-domain dataset?

**Regularization**

- Overfitting can be prevented with early stopping
  - Effective, but requires a separate in-domain validation set
- We empirically investigate explicit regularization techniques
  - Variational dropout (Gal and Ghahramani, 2016)
    - Randomly drop activations to zero the same way for each time step
    - Not a specific domain adaptation method
  - MAP-L2 penalization (Chelba and Acero, 2006)
    - Penalize the L2-distance between the weights of the in-domain and out-of-domain models
    - We are the first to apply it to the domain adaptation of neural networks
  - Tuneout
    - For each layer, randomly drop activations towards those computed with the weights of the out-of-domain model

**Experimental setup**

- Language pairs: English-to-German and English-to-Russian
- Out-of-domain data: WMT16 parallel + backtranslated monolingual data
- In-domain data: IWSLT 2015 (En→De) / 2014 (En→Ru)
- Model: GRU sequence-to-sequence with attention
- System: Nematus toolkit with BPE subword segmentation

**Results**

<table>
<thead>
<tr>
<th>System</th>
<th>En→De validation</th>
<th>En→De test (avg.)</th>
<th>En→Ru validation</th>
<th>En→Ru test (avg.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out-of-domain only</td>
<td>27.19</td>
<td>27.76</td>
<td>15.74</td>
<td>16.81</td>
</tr>
<tr>
<td>Early-stopping baseline</td>
<td>30.53</td>
<td>31.20</td>
<td>17.47</td>
<td>18.67</td>
</tr>
<tr>
<td>Early-stopping + dropout</td>
<td>30.63</td>
<td>31.33</td>
<td>17.68</td>
<td>18.80</td>
</tr>
<tr>
<td>Early-stopping + MAP-L2</td>
<td>30.81</td>
<td>31.25</td>
<td>17.77</td>
<td>18.91†</td>
</tr>
<tr>
<td>Early-stopping + tuneout</td>
<td>30.49</td>
<td>30.78†</td>
<td>17.51</td>
<td>18.78</td>
</tr>
<tr>
<td>Early-stopping + dropout + MAP-L2</td>
<td>30.80</td>
<td>31.48†</td>
<td>17.74</td>
<td>19.10†</td>
</tr>
</tbody>
</table>

†: different from the fine-tuning baseline at 5% significance.

**Training curves**

**Findings**

- On full-sized IWSLT training data
  - Dropout and MAP-L2 stabilize training, preventing overfitting
  - Dropout + MAP-L2 significantly improve over Early-stop alone
  - Tuneout did not yield improvements
- We evaluate Dropout + MAP-L2 over different in-domain data sizes (10-206,000)
  - Logarithmic relation between data size and BLEU
  - Even for fixed number of epochs perform equally or better than Early-stop
  - Don’t require held-out validation set
  - Fine-tuning without Early-stop or regularizers underfits or overfits
- We recommend using Dropout + MAP-L2 for fine-tuning, especially for very small amounts of in-domain data

**Links**

- Nematus (includes Dropout and MAP-L2)
  - https://github.com/EdinburghNLP/nematus
- Nematus (Tuneout branch) https://git.io/v7j8Z

---

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreements 644333 (TraMOOC) and 645487 (ModernMT). We also thank Booking.com for their support.