• Neural Language Models (NLMs): able to model long contexts of
  – Target words: standard NLMs (Bengio et al., 2003).
  – Source + target words: joint NLMs (Devlin et al., 2014).

• Only 1- or 2-hidden-layer NLMs have been used in MT.
  – (Schwenk, 2010), (Vaswani et al., 2013): no effect on layers.
  – (Schwenk et al., 2012), (Devlin et al., 2014): a small gain with 2 layers.

Train deep NLMs with wisdom from past works.

No clear results on whether deeper models are better!

Deep Neural Language Models in Machine Translation
Source context
She \textit{jogs along the bank of the river} in the morning
Elle fait du \textit{jogging le long de la rivière} ...
Target context

Chinese\textarrow{\rightarrow}English MT Experiments

• Data: bitext from the DARPA BOLT program.
  – 11.2M sent pairs (281M Chinese and 307M English words).

• NLM Training: 11-gram src context, 4-gram tgt history
  – Top 40K most frequent words.
  – 256-dim embeddings, 512-dim hidden layers.
  – Train 4 epochs: 10-14 days on Tesla K40 (1000 target words/s).

• Task: use NLMs to rerank n-best lists of phrasal MT
  – Strong baseline: similar to (Green et al., 2014)
    • Dense features + sparse features for rules, word pairs/classes.
    • 3 LMs over English bitext, 16.38 word corpus (word/class).
  – Discriminative reranker: on 1000-best output
    • MERT on all dense features, the decoder score, an NLM score.

Results

• NLMs: more layers, better perplexities / normalization.
  – Self-norm weight $\alpha=0.1$, validation: 585 sents, test: 1124 sents.

<table>
<thead>
<tr>
<th>Models</th>
<th>Perplexity Valid</th>
<th>Normalization Test</th>
<th>(log Z)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 layer</td>
<td>9.39</td>
<td>8.99</td>
<td>0.51</td>
</tr>
<tr>
<td>2 layers</td>
<td>9.20</td>
<td>8.96</td>
<td>0.50</td>
</tr>
<tr>
<td>3 layers</td>
<td>8.64</td>
<td>8.13</td>
<td>0.43</td>
</tr>
<tr>
<td>4 layers</td>
<td>8.10</td>
<td>7.71</td>
<td>0.35</td>
</tr>
</tbody>
</table>

• Reranking: deeper models give bigger gains
  – 3/4 layers: +0.9 BLEU over baseline, +0.5 BLEU over 1/2 layers.
  – test1: dev10wb syscomtune, test2: p1r6_dev.

Domain Adaptation

• Adapt from web-forum domain to sms-chat:
  – Use existing models: finetune on out-of-domain data.
  – Sms-chat corpus: 146K sent pairs.

• Similar trends observed:
  – 3 layers: +3.1 BLEU over baseline, +0.5 BLEU over 1/2 layers.
  – 4 layers: overfit training data.
  – test: p2r2smscht syscomtune

Analysis

• Learning curve: deeper NLMs are better than 1-layer NLMs.
  – Gaps between models towards the end are 40.1, 1.1, 2.0 (in perplexities).

Conclusion

• Adapt from web-forum domain to sms-chat:
  – Use existing models: finetune on out-of-domain data.
  – Sms-chat corpus: 146K sent pairs.

• Similar trends observed:
  – 3 layers: +3.1 BLEU over baseline, +0.5 BLEU over 1/2 layers.
  – 4 layers: overfit training data.
  – test: p2r2smscht syscomtune

• Learning curve: deeper NLMs are better than 1-layer NLMs.
  – Gaps between models towards the end are 40.1, 1.1, 2.0 (in perplexities).

• Demonstrate how to adapt NLMs to out-domain conditions:
  – Deep NLMs with 3 layers yield substantial gains.
  – 4-layer models unfortunately overfit. Future: more regularization.

• Demonstrate how to adapt NLMs to out-domain conditions:
  – Deep NLMs with 3 layers yield substantial gains.
  – 4-layer models unfortunately overfit. Future: more regularization.