Hypotheses

NMT involves a fixed training procedure where each sentence is sampled once during each epoch:

- Some sentences are well-learned during the initial few epochs.
- Some sentences were not well learned until 10-30 epochs.
- Training these two type sentences together results in a wastage of time.

We propose a Dynamic Sentence Sampling (DSS) method:

- We use the training cost difference as the criterion to measure which sentence has been well-learned.
- We propose two sentence sampling mechanisms: Weighted Sampling (WS) and Review Mechanism (RM).

Criterion

The training cost of a sentence pair \((x, y)\) from corpus \(D\) during the \(i\)th iteration can be calculated as:

\[
\text{cost}^i_{(x,y)} = -\log P(y|x, \theta). \tag{1}
\]

We adopt the ratio of differences (\(\text{dif}\)) between training costs of two training iterations to be the criterion,

\[
\text{dif}^i_{(x,y)} = \frac{\text{cost}^{i-1}_{(x,y)} - \text{cost}^i_{(x,y)}}{\text{cost}^{i-1}_{(x,y)}}. \tag{2}
\]

Dynamic Sentence Sampling (DSS)

(1) Weighted Sampling (WS)

Weighted sampling without any replacement was used to select a small subset, such as 80% of the entire corpus, as the corpus \(D_{ws}^{i+1}\) to perform the subsequent iteration.

\[
J_{ws} = \sum_{(x,y) \in D_{ws}} -\log P(y|x, \theta). \tag{3}
\]

(2) Review Mechanism (RM)

To prevent the loss of the knowledge that was obtained from the \(D_{low}\) group during NMT, a small percentage \(\lambda\), such as 10%, of the \(D_{low}\) group is sampled as the knowledge to be reviewed.

\[
J_{rm} = \sum_{(x,y) \in D_{high}} -\log P(y|x, \theta) + \sum_{(x,y) \in \lambda D_{low}} -\log P(y|x, \theta). \tag{4}
\]

Results and Analyses

We adopted attentional RNN based NMT by Nematus.

<table>
<thead>
<tr>
<th></th>
<th>Dev (NIST02)</th>
<th>Test (NIST03-08)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBSMT</td>
<td>33.15</td>
<td>29.66</td>
</tr>
<tr>
<td>Vanilla NMT</td>
<td>38.48</td>
<td>35.08</td>
</tr>
<tr>
<td>Random Sampling</td>
<td>38.35</td>
<td>34.62</td>
</tr>
<tr>
<td>Kocmi (Curriculum Learning)</td>
<td>38.51</td>
<td>35.19</td>
</tr>
<tr>
<td>Wees (Dynamic tuning)</td>
<td><strong>39.16</strong></td>
<td><strong>35.62</strong></td>
</tr>
<tr>
<td>Zhang (Boosting)</td>
<td>39.08</td>
<td>35.57</td>
</tr>
<tr>
<td>DSS-WS</td>
<td>39.54+</td>
<td>36.85++</td>
</tr>
<tr>
<td>DSS-RM</td>
<td><strong>39.89++</strong></td>
<td><strong>37.33++</strong></td>
</tr>
</tbody>
</table>

Discussions

We would like to investigate what would happen if:

- Train on larger/extreme-large corpora.
- Keep training for longer time.
- There is noisy/low-quality data in the corpus.