Structured Ramp Loss Minimization for Machine Translation

Kevin Gimpel  
Noah A. Smith

LEARNING IN MACHINE TRANSLATION

Why is learning in MT different from other tasks?

References often unreachable, so surrogate references are used instead (e.g., BLEU-oracles on k-best lists; Och & Ney 2003)

How are learning algorithms affected?

Loss functions are changed:

\[
\text{Perceptron} \quad \text{Latent perceptron loss:} \quad \text{loss}_{\text{perc}} = - \max_{y \in \{+1,-1\}} \text{score}(x, y, h) + \max_{y \in \{+1,-1\}} \
\text{score}(x, y, h) - \text{cost}(y, y')
\]

\[
\text{Latent perceptron with } k \text{-best BLEU oracle (Liang et al., 2006):} \quad \text{loss}_{\text{perc blu}} = - \max_{y \in \{+1,-1\}} \text{score}(x, y, h) + \max_{y \in \{+1,-1\}} \
\text{score}(x, y, h) - \text{cost}(y, y')
\]

\[
\text{MIRA for MT (Chiang et al., 2008; 2009):} \quad \text{loss}_{\text{mira}} = \log \sum_{y \in \{+1,-1\}} \exp \left( \text{score}(x, y, h) \right)
\]

\[
\text{Log loss for MT (Och & Ney, 2002):} \quad \text{loss}_{\text{log}} = \log \sum_{y \in \{+1,-1\}} \exp \left( \text{score}(x, y, h) \right) + \log \sum_{y \in \{+1,-1\}} \exp \left( \text{score}(x, y, h) \right)
\]

\[
\text{Algorithm 1. Rampion} \quad \text{Campanula rapunculus} \quad \text{"A hardy biennial, cultivated for the use of its fleshy roots in salads, either boiled or in a raw state, generally the latter; the leaves are also used in winter salads" (Nicholson, 1884)}
\]

EXPERIMENTS

Small-Feature Experiments

Moses phrase-based MT system, 14 default features, default Moses initialization, 3 runs of each algorithm

<table>
<thead>
<tr>
<th>Method</th>
<th>Urdu-English</th>
<th>Chinese-English</th>
<th>Arabic-English</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>MERT</td>
<td>24.5 ± 0.1</td>
<td>35.7 ± 0.3</td>
<td>55.0 ± 1.7</td>
<td>36.6</td>
</tr>
<tr>
<td>PRO</td>
<td>24.2 ± 0.1</td>
<td>36.3 ± 0.1</td>
<td>55.6 ± 0.3</td>
<td>36.6</td>
</tr>
<tr>
<td>RAMPION</td>
<td>24.5 ± 0.1</td>
<td>36.4 ± 0.4</td>
<td>55.5 ± 0.3</td>
<td>36.8</td>
</tr>
</tbody>
</table>

Large-Feature Experiments

14 default Moses features + 7,200 additional features:

1k most frequent bilingual word pairs 200 most frequent unigrams, 1k most frequent bigrams, 1k most frequent trigrams

4k top trigger pairs, ranked by mutual information (Rosenfeld, 1996)

<table>
<thead>
<tr>
<th>Method</th>
<th>Urdu-English</th>
<th>Chinese-English</th>
<th>Arabic-English</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>MERT</td>
<td>24.5 ± 0.1</td>
<td>35.7 ± 0.3</td>
<td>55.0 ± 1.7</td>
<td>36.6</td>
</tr>
<tr>
<td>PRO</td>
<td>24.2 ± 0.1</td>
<td>36.3 ± 0.1</td>
<td>55.6 ± 0.3</td>
<td>36.6</td>
</tr>
<tr>
<td>RAMPION</td>
<td>24.5 ± 0.1</td>
<td>36.4 ± 0.4</td>
<td>55.5 ± 0.3</td>
<td>36.8</td>
</tr>
</tbody>
</table>

REFERENCES


