Tuning for MT

\[ \hat{e} = \arg \max_e p(e|f; \theta) \]
\[ = \arg \max_e w^T h(f, e) \]

- MERT (Minimum Error Rate Training) by Och (2003)
  \[ \hat{w} = \arg \min_w \ell(\left\{ \arg \max_e w^T h(f_i, e) \right\}, \{e_i\}^N_{i=1}) \]
- PRO (Pair-wise Rank Optimization) by Hopkins and May (2010)
  \[ \arg \min_w \frac{1}{2} \|w\|_2^2 + \ell(w; D) \]
with hinge-loss:
\[ \frac{1}{M(w; D)} \sum_{(f,e) \in D} \max \{ 0, 1 - w^T \Phi(f, e, e') \} \]
\[ e' \in \text{NBEST}(w; f) \setminus \text{ORACLE}(w; f, e) \]
\[ \Phi(f, e, e') = h(f, e) - h(f, e'). \]
- Batch algorithm: an iterative k-best merging approximation

Optimized Online Learning

- First, suffer gradients from L2-regularizer
  \[ w_{k+\frac{1}{2}} := (1 - \lambda \eta_k) w_k \]
- Second, solve:
  \[ \arg \min_w \frac{1}{2} \|w - w_{k+\frac{1}{2}}\|_2^2 + \eta_k \sum_{(f,e) \in \Phi(f, e, e')} \xi_{f,e^*,e'} \]
\[ w_k^T \Phi(f, e^*, e') \geq 1 - \xi_{f,e^*,e'} \]
\[ \xi_{f,e^*,e'} \geq 0. \]
- Third, Lagrangian dual:
  \[ \arg \min_{\tau, e^*, e'} \frac{1}{2} \sum_{(f,e) \in \Phi(f, e, e')} \tau_{e^*, e'} \Phi(f, e^*, e') \]
\[ - \sum_{(f,e) \in \Phi(f, e, e')} \left\{ 1 - w_{k+\frac{1}{2}}^T \Phi(f, e^*, e') \right\} \]
- Finally:
  \[ w_{k+\frac{1}{2}} := w_{k+\frac{1}{2}} + \sum_{(f,e) \in \Phi(f, e, e')} \tau_{e^*, e'} \Phi(f, e^*, e') \]

Note:
- Update by SGD
  \[ w_{k+\frac{1}{2}} := w_{k+\frac{1}{2}} + \sum_{(f,e) \in \Phi(f, e, e')} \eta_k \frac{M(w_k; b)}{M(w_k; D)} \Phi(f, e^*, e') \]
- MIRA solve this:
  \[ \arg \min_{w} \frac{1}{2} \|w - w_k\|_2^2 + \sum_{(f,e) \in \Phi(f, e, e')} \xi_{f,e^*,e'} \]

Optimized Parallel Learning

- Online approximation to the learning objective
- Optimization for sentence-BLEU ≠ corpus BLEU!
- Larger sentence-batch for better corpus-BLEU approximation → slower convergence
- Each shard learns \( w^{t+1} \) using \( D_s \)
  \[ w^{t+\frac{1}{2}} := 1/S \sum w^{t+1,s} \]
\[ w^{t+1} := (1 - \rho) w^t + \rho w^{t+\frac{1}{2}} \]
- end for
- return \( w^{T+1} \)

Experiments

- NIST08 Chinese-to-English translation task
- MT02/MT06/MT08 for tuning/development testing/test
- SCFG / # of features = 14 / batch size = 16 / cores = 8
- MERT/PRO/MIRA and Online Rank Optimization (ORO)

Main results by BLEU

<table>
<thead>
<tr>
<th></th>
<th>MT06</th>
<th>MT08</th>
</tr>
</thead>
<tbody>
<tr>
<td>MERT</td>
<td>31.45†</td>
<td>24.13†</td>
</tr>
<tr>
<td>PRO</td>
<td>31.76†</td>
<td>24.43†</td>
</tr>
<tr>
<td>MIRA-L</td>
<td>31.42†</td>
<td>24.15†</td>
</tr>
<tr>
<td>ORO-hinge</td>
<td>29.76</td>
<td>21.96</td>
</tr>
<tr>
<td>O-ORO-hinge</td>
<td>32.06</td>
<td>24.95</td>
</tr>
<tr>
<td>O-ORO-softmax</td>
<td>30.77</td>
<td>23.07</td>
</tr>
<tr>
<td>O-ORO-L-softmax</td>
<td>31.16</td>
<td>23.20</td>
</tr>
</tbody>
</table>

Learning curve

Mixing by line search