Optimized Online Rank Learning for Machine Translation

Taro Watanabe
National Institute of Information and Communication Technology (NICT)
taro.watanabe@nict.go.jp

Tuning for MT

$$\begin{aligned}\hat{e} &= \arg\max_{e} p(e|f;\theta) \\ &= \arg\max_{e} \mathbf{w}^\top \mathbf{h}(f,e)\end{aligned}$$

- MERT(Minimum Error Rate Training) by Och (2003)

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \ell\left(\left\{\arg\max_{e} \mathbf{w}^\top \mathbf{h}(f^i, e)\right\}_{i=1}^{N}, \{\mathbf{e}^i\}_{i=1}^{N}\right)$$

- PRO(Pair-wise Rank Optimization) by Hopkins and May (2010)

$$\arg\min_{\mathbf{w}} \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \ell(\mathbf{w}; D)$$

with hinge-loss:

$$\frac{1}{M(\mathbf{w};D)} \sum_{(f,\mathbf{e})\in D} \sum_{e^*,e'} \max\left\{0, 1 - \mathbf{w}^\top \Phi(f,e^*,e')\right\}$$

$$\begin{aligned} e' &\in \textsc{NBest}(\mathbf{w}; f) \setminus \textsc{Oracle}(\mathbf{w}; f, \mathbf{e}) \\ e^* &\in \textsc{Oracle}(\mathbf{w}; f, \mathbf{e}) \\ \Phi(f,e^*,e') &= \mathbf{h}(f,e^*) - \mathbf{h}(f,e'). \end{aligned}$$

- Batch algorithm: an iterative k-best merging approximation

Online Learning (SGD)

$k = 1, \mathbf{w}_1 \leftarrow \mathbf{0}$
for $t = 1, ..., T$ **do**
  Choose $B_t = \{b_1^t, ..., b_K^t\}$ from D
  for $b \in B_t$ **do**
    Compute n-bests and oracles of b
    Set learning rate η_k
    $\mathbf{w}_{k+\frac{1}{2}} \leftarrow \mathbf{w}_k - \eta_k \nabla(\mathbf{w}_k; b)$
    $\mathbf{w}_{k+1} \leftarrow \min\left\{1, \frac{1/\sqrt{\lambda}}{\|\mathbf{w}_{k+\frac{1}{2}}\|_2}\right\} \mathbf{w}_{k+\frac{1}{2}}$
    $k \leftarrow k + 1$
  end for
end for
return $\mathbf{w}_k$

- Online approximation to the learning objective
- Optimization for sentence-BLEU ≠corpus BLEU!
- Larger sentence-batch for better corpus-BLEU approximation → slower convergence

Optimized Online Learning

- First, suffer gradients from L_2-regularizer

$$\mathbf{w}_{k+\frac{1}{4}} \leftarrow (1 - \lambda\eta_k)\mathbf{w}_k$$

- Second, solve:

$$\begin{aligned} &\arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w} - \mathbf{w}_{k+\frac{1}{4}}\|_2^2 + \eta_k \sum_{(f,\mathbf{e})\in b, e^*, e'} \xi_{f,e^*,e'} \\ &\mathbf{w}^\top \Phi(f,e^*,e') \geq 1 - \xi_{f,e^*,e'} \\ &\xi_{f,e^*,e'} \geq 0. \end{aligned}$$

- Third, Lagrangian dual:

$$\begin{aligned} \arg\min_{\tau_{e^*,e'}} &\frac{1}{2}\Big\|\sum_{(f,\mathbf{e})\in b, e^*, e'} \tau_{e^*,e'} \Phi(f,e^*,e')\Big\|_2^2 \\ &- \sum_{(f,\mathbf{e})\in b, e^*, e'} \tau_{e^*,e'} \left\{1 - \mathbf{w}_{k+\frac{1}{4}}^\top \Phi(f,e^*,e')\right\} \end{aligned}$$

- Finally:

$$\mathbf{w}_{k+\frac{1}{2}} \leftarrow \mathbf{w}_{k+\frac{1}{4}} + \sum_{(f,\mathbf{e})\in b, e^*, e'} \tau_{e^*,e'} \Phi(f,e^*,e')$$

- Note:
  - Update by SGD

$$\mathbf{w}_{k+\frac{1}{2}} \leftarrow \mathbf{w}_{k+\frac{1}{4}} + \sum_{(f,\mathbf{e})\in b, e^*, e'} \frac{\eta_k}{M(\mathbf{w}_k; b)} \Phi(f,e^*,e')$$

  - MIRA solve this:

$$\arg\min_{\mathbf{w}} \frac{\lambda}{2}\|\mathbf{w} - \mathbf{w}_k\|_2^2 + \sum_{(f,\mathbf{e})\in b, e^*, e'} \xi_{f,e^*,e'}$$

Optimized Parallel Learning

$\mathbf{w}^1 \leftarrow \mathbf{0}$
for $t = 1, ..., T$ **do**
  $\mathbf{w}^{t,s} \leftarrow \mathbf{w}^t$
  Each shard learns $\mathbf{w}^{t+1,s}$ using D_s
  $\mathbf{w}^{t+\frac{1}{2}} \leftarrow 1/S \sum_s \mathbf{w}^{t+1,s}$
  $\mathbf{w}^{t+1} \leftarrow (1-\rho)\mathbf{w}^t + \rho \mathbf{w}^{t+\frac{1}{2}}$
end for
return $\mathbf{w}^{T+1}$

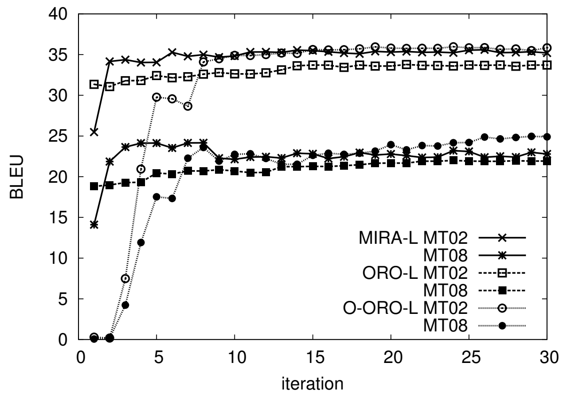
- Each shard learns locally + averaging in each round
- Line search to determine the optimal mixing ρ

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) with hinge-loss/softmax-loss (1,000-best, 30-iterations)

	MT06	MT08
MERT	31.45†	24.13†
PRO	31.76†	24.43†
MIRA-L	31.42†	24.15†
ORO-L_{hinge}	29.76	21.96
O-ORO-L_{hinge}	**32.06**	**24.95**
ORO-L_{softmax}	30.77	23.07
O-ORO-L_{softmax}	31.16†	23.20

Main results by BLEU



Learning curve

	MT06	MT08
MIRA	30.95	23.06
MIRA-L	31.42†	24.15†
$\text{ORO}_{\text{hinge}}$	29.09	21.93
ORO-L_{hinge}	29.76	21.96
$\text{ORO}_{\text{softmax}}$	30.80	23.06
ORO-L_{softmax}	30.77	23.07
O-$\text{ORO}_{\text{hinge}}$	31.15†	23.20
O-ORO-L_{hinge}	**32.06**	**24.95**
O-$\text{ORO}_{\text{softmax}}$	31.40†	23.93†
O-ORO-L_{softmax}	31.16†	23.20

Mixing by line search