Multi-Task Minimum Error Rate Training for SMT

Patrick Simianer, Katharina Wäschle, Stefan Riezler

Department of Computational Linguistics
University of Heidelberg, Germany
Multi-Task Learning

- Multi-task learning aims at learning several different tasks simultaneously,
  - addressing **commonalities** through **shared parameters**
  - and modeling **differences** through **task-specific parameters**.

- Predestined application: Patent translation over classes of patents w.r.t. International Patent Classification (IPC)
  - **commonalities**: highly specialized legal jargon not found in everyday language, rigid textual structure including highly formulaic language.
  - **differences**: technological terminology specific to IPC class.
IPC Sections

A Human Necessities
B Performing Operations; Transporting
C Chemistry; Metallurgy
D Textiles; Paper
E Fixed Constructions
F Mechanical Engineering; Lighting; Heating; Weapons; Blasting
G Physics
H Electricity
Goal and Approach

**Goal:** Learn a translation system that performs well across several different patent sections, thus benefits from shared information, and yet is able to address the specifics of each patent section.

**Approach:** Machine learning approach to trading off optimality of parameter vectors for each task-specific model and closeness of these model parameters to average parameter vector across models.
Assume specific setting: Not enough data for training generative SMT pipeline on all tasks, however, enough data for tuning for each specific task.

In other words: How much gain is there in extending the standard tuning technique of minimum error rate training (MERT) to **multi-task MERT** for SMT.

Also apply techniques for parameter averaging from distributed learning to a version of **averaged MERT**.
Parallel Patent Data

- Extract bilingual abstract and claims sections from the EP and WO parts for German-to-English translation.
- Sentence splitting and tokenizing with Europarl tools\(^1\).
- Sentence alignment with Gargantua 1.0b\(^2\).

\(^1\)http://www.statmt.org/europarl/
\(^2\)http://sourceforge.net/projects/gargantua/
Distribution of IPC sections for de-en abstracts and claims

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>266,521</td>
<td>21.81%</td>
</tr>
<tr>
<td>B</td>
<td>384,517</td>
<td>31.47%</td>
</tr>
<tr>
<td>C</td>
<td>372,903</td>
<td>30.52%</td>
</tr>
<tr>
<td>D</td>
<td>50,579</td>
<td>4.14%</td>
</tr>
<tr>
<td>E</td>
<td>54,396</td>
<td>4.45%</td>
</tr>
<tr>
<td>F</td>
<td>149,370</td>
<td>12.22%</td>
</tr>
<tr>
<td>G</td>
<td>291,671</td>
<td>23.87%</td>
</tr>
<tr>
<td>H</td>
<td>228,147</td>
<td>18.67%</td>
</tr>
</tbody>
</table>
Parallel data for de-en patent translation

<table>
<thead>
<tr>
<th># parallel sents</th>
<th>train</th>
<th>dev</th>
<th>devtest</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg. # tokens de</td>
<td>32,329,745</td>
<td>59,376</td>
<td>60,061</td>
<td>59,930</td>
</tr>
<tr>
<td>avg. # tokens en</td>
<td>36,005,763</td>
<td>69,584</td>
<td>70,700</td>
<td>70,331</td>
</tr>
</tbody>
</table>
Multi-task learning objective

**Objective:** Minimize task-specific loss functions $l_d$ under regularization of task-specific parameter vectors $w_d$ towards an average parameter vector $w_{avg}$.

\[
\min_{w_1, \ldots, w_D} \sum_{d=1}^{D} l_d(w_d) + \lambda \sum_{d=1}^{D} \| w_d - w_{avg} \|_p^p
\]  
(1)
Prediction:

Task-specific weight vectors $w_d \in \{w_1, \ldots, w_D\}$ that have been adjusted to trade off task-specificity (small $\lambda$) and commonality (large $\lambda$).

or: Average weight vector $w_{\text{avg}}$ as a global model.
Average MERT

\[
\text{AvgMERT}(w^{(0)}, D, \{c_d\}_{d=1}^D): \\
\text{for } d = 1, \ldots, D \text{ parallel do} \\
\quad \text{for } t = 1, \ldots, T \text{ do} \\
\quad \quad w_d^{(t)} = \text{MERT}(w_d^{(t-1)}, c_d(w_d)) \\
\quad \text{end for} \\
\text{end for} \\
\text{return } w_{\text{avg}} = \frac{1}{D} \sum_{d=1}^D w_d^{(T)}
\]

- Apply ideas from distributed learning (Zinkevich et al. NIPS’10) by basing the distribution strategy on task-specific partitions of data.
regularization: Set $p=1$ in equation 1 to obtain an $\ell_1$ regularizer.

clipping: Weight vector $w_d$ is moved towards the average weight vector $w_{\text{avg}}$ by adding or subtracting the penalty $\lambda$ for each weight component $w_d[k]$, and clipped when it crosses the average.

code: Script wrapper around the MERT implementation of Bertoldi et al. 2009; licensed unter the LGPL; online at http://www.cl.uni-heidelberg.de/statnlpgroup/mmert/.
Multi-task MERT

\[
\text{MMERT}(w^{(0)}, D, \{c_d\}_{d=1}^D): \\
\text{for } t = 1, \ldots, T \text{ do} \\
\quad w^{(t)} = \frac{1}{D} \sum_{d=1}^D w^{(t-1)} \\
\quad \text{for } d = 1, \ldots, D \text{ parallel do} \\
\quad \quad w^{(t)}_d = \text{MERT}(w^{(t-1)}_d, c_d(w_d)) \\
\quad \quad \text{for } k = 1, \ldots, K \text{ do} \\
\quad \quad \quad \text{if } w^{(t)}_d[k] - w^{(t)}_{\text{avg}}[k] > 0 \text{ then} \\
\quad \quad \quad \quad w^{(t)}_d[k] = \max(w^{(t)}_{\text{avg}}[k], w^{(t)}_d[k] - \lambda) \\
\quad \quad \quad \text{else if } w^{(t)}_d[k] - w^{(t)}_{\text{avg}}[k] < 0 \text{ then} \\
\quad \quad \quad \quad w^{(t)}_d[k] = \min(w^{(t)}_{\text{avg}}[k], w^{(t)}_d[k] + \lambda) \\
\quad \quad \quad \text{end if} \\
\quad \quad \text{end for} \\
\quad \text{end for} \\
\text{end for} \\
\text{return } w^{(T)}_1, \ldots, w^{(T)}_D, w^{(T)}_{\text{avg}}
Experimental Setup

- Open-source Moses SMT system (Koehn et al. 2007); MERT implementation of Bertoldi et al. 2009.
- All systems use same phrase tables and language models, trained on 1M parallel data pooled from all IPC sections.
- *ind.* systems are tuned on each IPC section separately.
- *pooled* system is tuned on 2K sentences pooled from 250 sentences from each IPC section.
- AvgMERT and MMERT are algorithms described above.
- \( w_{\text{avg}} \) is global model produced as by-product in multi-task learning.
Experimental Evaluation

- All systems evaluated on 8 test sets, each consisting of 2K sentences from a separate IPC domain.
- Statistical significance of pairwise result differences assessed by $p$-values smaller than 0.05 using Approximate Randomization test (Riezler & Maxwell 2005).
- Statistically significant improvement over $ind$ indicated by $\ast$
- Statistically significant improvement over $pooled$ indicated by $+$
- Statistically significant improvement over AvgMERT indicated by $\#$
## Experimental Results

<table>
<thead>
<tr>
<th>section</th>
<th>ind</th>
<th>pooled</th>
<th>AvgMERT</th>
<th>MMERT</th>
<th>$w_{avg}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.5187</td>
<td>0.5199</td>
<td><strong>0.5213</strong> *</td>
<td>0.5195#</td>
<td>0.5196#</td>
</tr>
<tr>
<td>B</td>
<td>0.4877</td>
<td>0.4885</td>
<td>0.4908**+</td>
<td>0.4911*</td>
<td><strong>0.4921</strong> *#</td>
</tr>
<tr>
<td>C</td>
<td>0.5214</td>
<td>0.5175</td>
<td>0.5199**+</td>
<td><strong>0.5218</strong> #</td>
<td>0.5162**#</td>
</tr>
<tr>
<td>D</td>
<td>0.4724</td>
<td>0.4730</td>
<td>0.4733</td>
<td><strong>0.4736</strong></td>
<td>0.4734</td>
</tr>
<tr>
<td>E</td>
<td>0.4666</td>
<td>0.4661</td>
<td>0.4679**+</td>
<td>0.4669</td>
<td><strong>0.4685</strong> *</td>
</tr>
<tr>
<td>F</td>
<td>0.4794</td>
<td>0.4801</td>
<td>0.4811*</td>
<td>0.4821*</td>
<td><strong>0.4830</strong> *#</td>
</tr>
<tr>
<td>G</td>
<td>0.4596</td>
<td>0.4576</td>
<td>0.4607+</td>
<td>0.4606</td>
<td><strong>0.4610</strong> *</td>
</tr>
<tr>
<td>H</td>
<td>0.4573</td>
<td>0.4560</td>
<td>0.4578</td>
<td><strong>0.4581</strong></td>
<td><strong>0.4581</strong></td>
</tr>
</tbody>
</table>
Discussion

- *pooled* shows no s.s. improvement over *ind*.
- Best results (**bold face**) achieved by AvgMERT, MMERT, or $w_{avg}$.
- Best results are small, but statistically significant improvements over *ind* and *pooled*.
- Significant degradation on section C ("chemistry") by averaging techniques due to exceptional character of chemical formulae and compound names.
- Interpretation of small improvements with a grain of salt, however, hope for larger improvements with larger feature sets.