Adaptive Quality Estimation for Machine Translation

Marco Turchi¹, Antonios Anastasopoulos³, José G.C. de Souza¹,², Matteo Negri¹

¹FBK – Fondazione Bruno Kessler, Trento, Italy, ²University of Trento, Italy, ³National Technical University of Athens, Greece

turchi, desouza, negri}@fbk.eu, anastasopoulos.ant@gmail.com

**Task: Machine Translation Quality Estimation (QE)**

Given a (source, target) pair, predict the quality of the target without reference translations.

(One) application scenario: assess at run time the quality of MT suggestions in a Computer-assisted translation (CAT) environment.

**Problem: adaptability**

Since:
- The notion of MT output quality is highly subjective
- Each translation job has its own specificities

...QE components should be capable to self-adapt to:
- the behavior of specific users
- differences between training and test data

**Solution: online learning**

Idea:
- Learn stepwise (either from scratch or by refining an existing model) from user feedback

User = human translator
Feedback = distance between predicted labels and “true labels”...calculated from MT post editions.

**Experimental setup**

Training/test data with different label distributions
- WMT12 QE shared task – EN/ES (artificial data partitions)
  - 1,832 training, 422 test sentences
- MateCat data – EN/IT (user and user+domain changes)
  - Legal (164 sentences) & Information Technology (280 sentences)
  - 8 professional translators

Comparison (Mean Absolute Error) between:
- **Adaptive**: built on top of an existing model
- **Empty**: only learns from the test set
- **Batch**: only learns from the training set
- **Baseline (μ)**: label with the mean HTER calculated on training

**Collecting and exploiting user feedback**

(source, target) ⇒ (source, target, post-edited target) ⇒ TERC⁴pp ⇒ HTER

(source, target, HTER) ⇒

**Predicting QE scores**

(source, target, HTER) ⇒ feature extractor QuEst² ⇒ learning algorithm ⇒ prediction

Features: 17 “baseline” QuEst features

Algorithms: SVR³, OnlineSVR⁴, Passive Aggressive Perceptron⁵

³ QuEst - http://www.quest.dcs.shef.ac.uk
⁴ OnlineSVR - http://www2.imperial.ac.uk/~gmontana/onlinesvr.htm
⁵ sofi/ml - https://code.google.com/p/solfi-ml

**Results on MateCat data (IT Vs Legal, Rad Vs Cons)**

<table>
<thead>
<tr>
<th>Train/Test</th>
<th>Label Distribution</th>
<th>Method</th>
<th>T</th>
<th>MAE</th>
<th>MAE</th>
<th>MAE</th>
<th>Alg</th>
<th>MAE</th>
<th>Alg</th>
</tr>
</thead>
<tbody>
<tr>
<td>L Cons</td>
<td>IT rad</td>
<td>HTER</td>
<td>24.5</td>
<td>26.4</td>
<td>27</td>
<td>18.2</td>
<td>OSVR</td>
<td>16.6</td>
<td>OSVR</td>
</tr>
<tr>
<td>IT rad</td>
<td>L Cons</td>
<td>MAE</td>
<td>24</td>
<td>25.4</td>
<td>19.7</td>
<td>OSVR</td>
<td>12.5</td>
<td>OSVR</td>
<td></td>
</tr>
<tr>
<td>L Cons</td>
<td>L rad</td>
<td>MAE</td>
<td>20.5</td>
<td>21.2</td>
<td>14.5</td>
<td>PA</td>
<td>12.5</td>
<td>OSVR</td>
<td></td>
</tr>
<tr>
<td>L Cons</td>
<td>L rad</td>
<td>Alg</td>
<td>19.4</td>
<td>21.2</td>
<td>16.1</td>
<td>PA</td>
<td>11.3</td>
<td>OSVR</td>
<td></td>
</tr>
<tr>
<td>IT Cons</td>
<td>L Cons</td>
<td>Alg</td>
<td>13.5</td>
<td>17.3</td>
<td>15.7</td>
<td>OSVR</td>
<td>12.5</td>
<td>OSVR</td>
<td></td>
</tr>
<tr>
<td>IT rad</td>
<td>L Cons</td>
<td>OSVR</td>
<td>12.8</td>
<td>19.2</td>
<td>17.5</td>
<td>OSVR</td>
<td>16.6</td>
<td>OSVR</td>
<td></td>
</tr>
<tr>
<td>L Cons</td>
<td>IT Cons</td>
<td>OSVR</td>
<td>12.7</td>
<td>17.6</td>
<td>15.1</td>
<td>OSVR</td>
<td>15.5</td>
<td>OSVR</td>
<td></td>
</tr>
<tr>
<td>IT rad</td>
<td>IT Cons</td>
<td>OSVR</td>
<td>9.6</td>
<td>16.8</td>
<td>13.6</td>
<td>PA</td>
<td>15.5</td>
<td>OSVR</td>
<td></td>
</tr>
<tr>
<td>L Cons</td>
<td>IT Cons</td>
<td>OSVR</td>
<td>8.3</td>
<td>12.3</td>
<td>10.7</td>
<td>OSVR</td>
<td>11.3</td>
<td>OSVR</td>
<td></td>
</tr>
<tr>
<td>IT rad</td>
<td>L Cons</td>
<td>OSVR</td>
<td>6.8</td>
<td>17</td>
<td>16.2</td>
<td>OSVR</td>
<td>16.6</td>
<td>OSVR</td>
<td></td>
</tr>
<tr>
<td>L rad</td>
<td>L Cons</td>
<td>OSVR</td>
<td>5.0</td>
<td>15.4</td>
<td>14.7</td>
<td>OSVR</td>
<td>15.5</td>
<td>OSVR</td>
<td></td>
</tr>
<tr>
<td>It rad</td>
<td>L Cons</td>
<td>OSVR</td>
<td>2.2</td>
<td>10.6</td>
<td>10.8</td>
<td>OSVR</td>
<td>11.3</td>
<td>OSVR</td>
<td></td>
</tr>
</tbody>
</table>

**Take home messages**

- Real-world scenarios raise new, interesting challenges for QE
- Training/test data homogeneity, users’ individual preferences, etc.
- Adaptability as a crucial capability (not only for CAT)
- Even in the same domain different user may show high ΔHTER
- Online learning from user corrections as a way to overcome the limitations of batch strategies
- Use “empty” models (with OSVR) with highly heterogeneous data
- Use our open source tool!

http://hlt.fbk.eu/technologies/aqet