Validating Quality Estimation in a CAT Workflow: Speed, Cost and Quality Trade-off

Fernando Alva-Manchego, Lucia Specia, Sara Szoc, Tom Vanallemeersch, Heidi Depraetere

MT Summit 2021
Outline

- APE-QuEst: A Quality Gate for Machine Translation
- Evaluation Protocol
- Experiments and Results
- Conclusions
How to speed up this process taking advantage of acceptable quality MT outputs?
I am not aware of this debt I would like this debt validated or removed from my credit report.

Je n’ai pas connaissance de cette dette que je souhaiterais que cette dette soit validée ou retirée de mon rapport de crédit.
CAT Workflow with MT + QE

Original Documents → Machine Translation → Quality Estimation → Acceptable Quality Translations → Human Translator → High Quality Translations
A Quality Gate for Machine Translation

Quality Gate

- **blocked**
  - human corrections
- **medium**
  - automatic corrections
- **passed**
  - directly to end user

Machine Translation

MT output is scored and sent to 3 separate workflows

How to determine QE scores thresholds for each workflow?

How do these thresholds affect speed/time/cost?
Experimental Setup: Machine Translation

- **eTranslation**
  - Neural MT models for more than 24 languages
  - Targeted mainly at European public administrations

- **Experiments include** English → Dutch, English → French

<table>
<thead>
<tr>
<th>Generic data</th>
<th>Domain-specific data</th>
</tr>
</thead>
<tbody>
<tr>
<td>We have read the new book.</td>
<td>Nous avons lu le nouveau livre.</td>
</tr>
<tr>
<td>They have described what happened there.</td>
<td>Ils ont décrit ce qui s’est passé là.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>I monitor my credit report, more frequently now as we’re attempting to buy our first house.</td>
<td>Je surveille mon rapport de crédit, plus souvent maintenant car nous essayons d’acheter notre première maison.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Experimental Setup: Quality Estimation

- Translation Quality Estimation at the *sentence level*

- Prediction of quality score: 1 - HTER score
  - HTER: Human Translation Edit Rate

1 = Everything was changed

0 = Nothing was changed

github.com/Unbabel/OpenKiwi

github.com/TharinduDR/TransQuest
Experimental Setup: Quality Estimation

- Trained language-specific models by fine-tuning Multilingual BERT
- Training dataset from Ive et al. (2020)
  - Tuples: (source, MT, human post-edition, target)
  - Legal domain
  - Size:
    - 11,249 for English-Dutch (EN-NL)
    - 9,989 for English-French (EN-FR)

<table>
<thead>
<tr>
<th>Model</th>
<th>EN-NL</th>
<th>MAE</th>
<th>EN-FR</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ive et al. (2020)</td>
<td>0.38</td>
<td>0.14</td>
<td>0.58</td>
<td>0.14</td>
</tr>
<tr>
<td>Ours</td>
<td>0.51</td>
<td>0.10</td>
<td>0.69</td>
<td>0.10</td>
</tr>
</tbody>
</table>
Research Questions

● When compared to the Traditional workflow, does the Quality Gate workflow help to improve speed (i.e. time to get to final translation) and reduce cost (how many translations need HPE)?

● When compared to the MT-Only workflow, does the Quality Gate workflow help to improve translation quality?
Evaluation Protocol: Measurable Criteria

- **Quality**: Percentage of sentences considered of acceptable quality by independent human raters

- **Cost**: Percentage of sentences that require human post-editing, versus being fit for purpose

- **Speed**: Time required for human post-edition
Evaluation Protocol: Use Cases

- Texts sampled from a European public administration handling consumer complaints

<table>
<thead>
<tr>
<th>Use Case 1 (Assimilation)</th>
<th>Use Case 2 (Dissemination)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consumer Complaints (informal)</td>
<td>Privacy Statement (formal)</td>
</tr>
<tr>
<td>Content needs to be <strong>understood</strong></td>
<td>Content needs to be <strong>published</strong></td>
</tr>
<tr>
<td>Is the translation Acceptable / Not Acceptable ?</td>
<td>Is the translation Publishable / Not Publishable ?</td>
</tr>
<tr>
<td>966 English source sentences</td>
<td>114 English source sentences</td>
</tr>
</tbody>
</table>
Data Collection: Human Post-Edits*

- All MT outputs were post-edited (in each use case and target language)
- Post-editors were experienced professional translators
- For each target language, three post-editors were hired, and each sentence was post-edited once
- Sentences were post-edited within their document context

*For confidentiality reasons, the example originates from a comparable, publicly available dataset (US Government)
Data Collection: Acceptability Ratings

- All MT outputs and HPEs were rated
- Raters were professional translators
  - They were not informed of whether the sentences being judged were an MT output or HPE
- For each target language and use case, two raters scored each translation (either MT or HPE) once
- Sentences were rated within their document context
Results: Predicted Scores

- High quality gains compared to MT only workflow
- Cost / time savings compared to traditional workflow
- Less impact on dissemination dataset due to high quality in-domain MT output
Results: Predicted vs Oracle Scores

<table>
<thead>
<tr>
<th>Lang</th>
<th>Threshold</th>
<th>Assimilation</th>
<th>Dissemination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Time (%)</td>
<td>Quality (%)</td>
</tr>
<tr>
<td></td>
<td>Traditional</td>
<td>100.00</td>
<td>97.67</td>
</tr>
<tr>
<td></td>
<td>QE &lt; 0.90</td>
<td>64.18</td>
<td>83.87</td>
</tr>
<tr>
<td></td>
<td>QE &lt; 0.80</td>
<td>8.80</td>
<td>59.16</td>
</tr>
<tr>
<td></td>
<td>MT-Only</td>
<td>0.00</td>
<td>54.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lang</th>
<th>Threshold</th>
<th>Assimilation</th>
<th>Dissemination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Time (%)</td>
<td>Quality (%)</td>
</tr>
<tr>
<td></td>
<td>Traditional</td>
<td>100.00</td>
<td>97.67</td>
</tr>
<tr>
<td></td>
<td>QE &lt; 0.90</td>
<td>90.04</td>
<td>97.09</td>
</tr>
<tr>
<td></td>
<td>QE &lt; 0.80</td>
<td>80.44</td>
<td>94.04</td>
</tr>
<tr>
<td></td>
<td>MT-Only</td>
<td>0.00</td>
<td>54.07</td>
</tr>
</tbody>
</table>

- Oracle scores show that there is still much room for improving QE predictions
Conclusions

- Evidence of the benefits of introducing QE into a CAT workflow
- **APE-QuEst Quality Gate**: use QE scores to determine if MT outputs can be used as-is (acceptable quality) or if they require post-edition (unacceptable quality)
- **Trade-off study**: establish thresholds on the QE scores
  - We collected human post-edits and acceptability ratings from real use case scenarios
  - Quality Gate can **obtain similar levels of quality** to the current human-only workflow, for all use cases and target languages explored
Interface to the Quality Gate

MT output / Human post-edition / APE output

QE score

Source: QE existing translation 2019-11-21 55.55555
Thank you!