PEMT human evaluation at 100x scale with risk-driven sampling

by Kirill Soloviev

CEO & co-founder

contentquo
**about me**

- **37** years old
- **20** years GILT
- **7** years TQM

- **2002**: linguist
- **2005**: localization PM & loc engineer
- **2009**: l10n pgm manager & director
- **2015**: tech founder
- **2018**: tech CEO
- **2022**: tech CEO
edit distance as a quality metric?!

Custom MT Engines ➔ Post-Editors

Anecdotal quality feedback

Not specific enough
Hard to analyze
Fixing problems is slow

"made lots of corrections"
"this engine is poor"
"raw quality is low"
"takes too long to post-edit"

"Red Pen Syndrome"
Allen 2003, Muzii 2014 et al.

Your MT Team
human evaluation for actionable insights
non-scientific matrix of MT quality evaluation approaches

- **Human** edit distance (subsegment) ***
- **Human** MQM (subsegment)
- **Human** adequacy - fluency (segment)
- **Automatic** w/reference (BLEU)
- **Automatic** non-reference (COMET)

*Not shown to scale!!!*
### Output of Human Quality Evaluations

<table>
<thead>
<tr>
<th>English (en-US)</th>
<th>Estonian (et-EE)</th>
<th>Adequacy</th>
<th>Fluency</th>
<th>MQM Annotation</th>
<th>Edit Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>The quick brown fox jumps over the lazy dog</td>
<td>Kiire <strong>pruun rebane hüppab üle laisa</strong> kaerakoera</td>
<td>4 out of 4</td>
<td>4 out of 4</td>
<td>Accuracy / Major Spelling / Minor</td>
<td>5%</td>
</tr>
</tbody>
</table>
risk-driven sampling

Continuously calculated ED / MTQE

Auto risk scoring and budget-limited sampling of PEMT jobs

Human evaluation

Specific, actionable MT quality error reports

Central issue database

EN>ES: “Acme” translated as “acne” (x4922)

JA>DE: informal “you” instead of formal (x12)

EE>RU: plural instead of singular (x201)

Best Use of Evaluation Budget

MT Team

Improved Custom Engines

Human

Edit Distance
Edit Time
Source Complexity
Previous Engine Scores
Red Pen bias correction
Automatic QC (terms)
MTQE Prediction Error...

Improved Custom Engines

Auto risk scoring and budget-limited sampling of PEMT jobs

Human evaluation

Specific, actionable MT quality error reports

Central issue database

EN>ES: “Acme” translated as “acne” (x4922)

JA>DE: informal “you” instead of formal (x12)

EE>RU: plural instead of singular (x201)
example quality
risk rules

PEMT jobs should be more likely to be picked for human quality evaluation when:

- [x4] Average Edit Distance for language X changed >10% over last 1 month
- [x2] Maximum Edit Distance for engine Y hit 60% twice over 10 translators
- [x0.5] Predicted Edit Distance for MTQE model Z differed >30% from ED
- [x1.5] Post-editor M’s median Edit Distance is >30% different from average
- etc.

+ Budget-guided cutoff point
continuous q.evaluation blueprint for PEMT

**MQM (EP/1000w)**

<table>
<thead>
<tr>
<th>Terms</th>
<th>Mistrans</th>
<th>Om/Ad</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>2.4</td>
<td>2.3</td>
<td>1.4</td>
</tr>
<tr>
<td>0.7</td>
<td>1.7</td>
<td>2.2</td>
<td>1.1</td>
</tr>
<tr>
<td>1.1</td>
<td>1.9</td>
<td>1.7</td>
<td>1.3</td>
</tr>
</tbody>
</table>

**Adequacy Fluency (1.0-4.0)**

<table>
<thead>
<tr>
<th>Wk1</th>
<th>Wk2</th>
<th>Wk3</th>
<th>Wk4</th>
<th>Wk5</th>
<th>Wk6</th>
<th>Wk7</th>
<th>Wk8</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>H</td>
<td>H</td>
<td>M</td>
<td>M</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
</tbody>
</table>

**Actual ED**

<table>
<thead>
<tr>
<th>Wk1</th>
<th>Wk2</th>
<th>Wk3</th>
<th>Wk4</th>
<th>Wk5</th>
<th>Wk6</th>
<th>Wk7</th>
<th>Wk8</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>M</td>
<td>S</td>
<td>M</td>
<td>M</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
</tbody>
</table>

**Predicted ED (MTQE, e.g. COMET or ModelFront)**

<table>
<thead>
<tr>
<th>Wk1</th>
<th>Wk2</th>
<th>Wk3</th>
<th>Wk4</th>
<th>Wk5</th>
<th>Wk6</th>
<th>Wk7</th>
<th>Wk8</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>M</td>
<td>S</td>
<td>M</td>
<td>M</td>
<td>S</td>
<td>M</td>
<td>S</td>
</tr>
</tbody>
</table>
benefits

- Improved ROI on human evaluations
- Faster custom engine improvement
- More reliable fix of Red Pen Syndrome
thank you!

time for Q&A

Email: Kirill.Soloviev@contentquo.com
LinkedIn: https://www.linkedin.com/in/kirillsoloviev/