MT Human Evaluation

Insights & Approaches

Paula Manzur
Agenda

MT Human Evaluations
Key roles, metrics and benefits
Insights on Data Reliability
How to evaluate MT
Ideas to experiment
Recommendations
MT Human Evaluations

MT quality assessment of one or more engines for future implementation in localization workflow and for MT engine improvement. Collaborate with Customer on Quality Expectations

Key roles

Use data to negotiate buy/sell MTPE rates (which need to be aligned with MT quality output) with Customers and Translators – even for baseline engines
## MT Human Evaluations

<table>
<thead>
<tr>
<th>Key metrics</th>
<th>Automatic Metrics (e.g. BLEU, METEOR, TER)</th>
<th>Human Assessment (by error annotation, classification, corrections to the target text)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key benefits</td>
<td>Allow translators (who will become post-editors) to get involved in the validation of the MT system</td>
<td>Allow Customers to make an informed decision on MT implementation with reliable data</td>
</tr>
</tbody>
</table>
MT Human Evaluations
Insights on Data Reliability

MT Automatic Metrics

Objective

Human Assessment

Subjective
MT Human Evaluations
Insights on Data Reliability

- Automatic metrics need a reference, a “golden” human translation – only one “correct” translation is possible otherwise the score will go down.

- Human assessment can be done without a golden reference – more than one “correct” translation possible?

- What makes a translation to be “the correct one” if there are different ways to translate the same sentence? – there might be other options that are “good enough” for the use case.
Humans can disagree without anyone being incorrect
Humans can disagree on a translation without anyone being incorrect.
Humans can disagree on a *machine* translation without anyone being incorrect.
Definition of “amazing goal”: a goal scored directly from corner (Olympic goal)

- For a translation to be “correct” it needs to follow certain rules!
- So what makes a translation “correct”?
- The adherence to the rule (that has been defined for the use case).
When MT is involved, why and where do we (humans) apply rules?

Un gol olímpico es lo más espectacular visto en el fútbol.

An Olympic goal is the most spectacular sight in football.

An Olympic goal is the most amazing sight in soccer.

An Olympic goal is the most amazing thing seen in football.

Olympic goals are the most fantastic sight in soccer.
How to evaluate MT then? Again, with rules!

DQF (Dynamic Quality Framework)
- 2 categories relevant for MT: accuracy and fluency

Evaluation data set (representative of entire content)
200 segments
Order of data should be randomized to eliminate bias
Four evaluators familiar with domain data

Source: TAUS
How much of these guidelines can we follow in practice?

**Quality Evaluation Guidelines**

DQF (Dynamic Quality Framework)
- 2 categories relevant for MT: accuracy and fluency

Evaluation data set (representative of entire content)
200 segments
Order of data should be randomized to eliminate bias
Four evaluators familiar with domain data

- Other categories might be relevant for the use case, such as Compliance and style.
- Is there a “perfect” evaluation data set? Why not a pilot project with Post-Editing in CAT?
- Budget and time might be a constraint. Usually 1 hour as allocated time for error annotation.
- If you randomize data, translators might ask for context. But can include a mix of sentences as long as they’re from the same domain.
- Budget and time constraints again. Usually 2 evaluators is possible, a 3rd could be a Language Specialist on Customer’s side.
Some Ideas to Experiment

A common error from MT is related to Gender Bias:

<table>
<thead>
<tr>
<th>Source</th>
<th>Target – Raw MT</th>
<th>Target – Post Edited</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marie Curie was born in Warsaw. The distinguished scientist received the nobel prize in 1903 and 1911.</td>
<td>Marie Curie nació en Varsovia. El distinguido científico recibió el premio Nobel en 1903 y 1911.</td>
<td>Marie Curie nació en Varsovia. La distinguida científica recibió el premio Nobel en 1903 y 1911.</td>
</tr>
</tbody>
</table>

Diff. between the versions

Marie Curie nació en Varsovia. La distinguida científica recibió el premio Nobel en 1903 y 1911.

In this example, MT is still comprehensible, and mostly usable up to a certain point – general idea can be understood but is not grammatically correct.
Some Ideas to Experiment

During Human Evaluation all is left is to choose an Error Category and Scoring:

**Diff. between the versions**

Marie Curie nació en Varsovia. La distinguida científica El distinguido científico recibió el premio Nobel en 1903 y 1911.

**Primary Issue**  
Language - Grammar, syntax

**Scoring**  
3-Mostly comprehensible and fluent, 1-2 minor issues; mostly usable

---

Evaluators see the errors they fixed and annotate the type of error. This data allow us to assess the level of MT usability to identify efficiency gains.
Recommendations

- Effective research: Make sure quality expectations are clearly defined from start
- Narrow it down to 2 baseline engines
- Use a quality evaluation framework to assess the engines (adjust if needed)
- Perform a full Pilot with Post-Editing, Human Evaluation and (if possible) automatic metrics

Based on gathered data:

- Share results with Language Teams and Customer to collaborate on rates
- Use learning from Evaluations to create post-editing instructions and training (if needed)
Thank You

Paula Manzur
Paula.Manzur@vistatec.com

Vistatec Machine Translation Team
VistatecMT@vistatec.com