Unbabel’s workflow: Crowdsourcing model and scalable quality management

Helena Moniz
Computational Linguist Researcher
What is quality and why does it matter in a AI and Crowd pipeline?

What are the quality processes to firstly assess then assure?

How can we create trust in a crowd ecosystem?
Unbabel Pipeline

MACHINE
- Job
- Machine Translation

Q.E.
- Quality Estimation
- Low Q.E.
- Re-Eval

COMMUNITY
- Crowd
- High Q.E.
- Customer

Unbabel Pipeline
Distributed Pipeline

Quality Estimation
Dynamic Factors on Quality

• Initial Text (MT)
• Editors Pool and Assignment
• Content types
• Interfaces
• Language pairs (arguable, but true!)
• Periodic Quality Assessment for Quality Assurance:
  • Crowd evaluation
  • Jobs annotation
Editors Pool

- **Testing Phase**: Editors are tested when they sign up.
- **Training Content**: Editors get ratings for the tasks.
- **Paid Content**: Best editors are selected.
- **Expert**: Evaluators and annotators are included.

Evaluators

Annotators
Editor’s on-boarding

• Language guidelines
• Usability guidelines
• Evaluation criteria
• Legal & ethical issues
Crowd Assessment

- **Evaluation task profile**
  - Multiple choice tests with usually no feedback provided (not us!)
  - **Scoring with feedback, so editors may learn and be evaluated again**

- **Inter-rater agreement**
  - several metrics
  - Editors automatic evaluation
  - Gold tasks for both evaluators and editors
Crowd Assessment

• Evaluators profile and training
  • Usability / interface training
  • Templates to provide constructive feedback
  • Forum to discuss problematic issues
Evaluation Error Criteria

- **Accuracy**
  - Addition
  - Omission
  - Mistranslation
    - False Friend
    - Overly Literal
    - Lexical Selective
    - Shouldn’t have been translated
  - Untranslated

- **Fluency**
  - Grammar
    - Agreement
    - Tense/Aspect
    - POS
  - Word order
  - Typography
    - Capitalization
    - Diacritics
    - Hyphenation
    - Orthography
    - Punctuation
    - Whitespace

- **Style**
  - Grammatical Register
  - Lexical Register
  - Wrong Language Variety
  - Company Terminology
How Editors are Evaluated
Editors Profiling
Editor Assignment

Regular distribution

Smart distribution

old rating
3.8

Improved rating
4.6
Editors Pool

- **Expert**
  - Evaluators and annotators

- **Paid Content**
  - Best editors
  - Editors get ratings for the tasks

- **Training Content**

- **Testing Phase**
  - Editors are tested when they sign up
Error typologies

- Localization Quality Evaluation (LQE, Google)
- Dynamic Quality Framework (DQF, TAUS), normalised with MQM
- Multidimensional Quality Metric (MQM, QT21)
MQM (QT21)

- 7 major categories:
  - Accuracy
  - Fluency
  - Style
  - Design
  - Locale
  - Verity
  - Fine-grained subcategorisation into 120 parameters
  - Subset of 60 parameters is used at Unbabel
MQM Core

Translation Quality

Accuracy
- Addition
- Mistranslation
- Omission
- Untranslated

Fluency
- Grammar
- Grammatical register
- Inconsistency
- Spelling
- Typography
- Unintelligible

Design
- Locale convention
- Style
- Terminology

Completeness
- Legal requirements
- Locale-specific content

Verity
- Legal requirements
- Locale-specific content
MQM (QT21)

• 4 levels of severity:
  • neutral, minor, major, and critical.

• **Formula:**
  
  \[ = 100 - \text{SUM}((1 \times \text{MINORS}) + (5 \times \text{MAJORS}) + (10 \times \text{CRITICALS})) / \#\text{Words} \times 100 \]

  • 95% as professional level translation
Annotation at Unbabel

- **Profile of the annotator:**
  - PhD on Linguistics and/or Translation Studies

- **Inter-annotator agreement**
  - training usability / parameters
  - gold tasks
  - feedback

- **Recent research studies:**
  - positive impact of Decision Trees (tested with EN > DE)
  - severity levels harder for agreement than the type of errors
Error Typology for Annotation
Deep Annotations
Annotation at Unbabel

• **Annotation output**
  • monitor quality in all steps / customer / domain /interfaces / QE and MT produced
  • feedback for editors (linguistic tips)
  • correlation with evaluation scores
  • feed NLP modules
Smartcheck

External NLP Services
- Spell Check
- Syntax Parser
- Word Aligner

Spelling
Register
Formality
Consistency
Client Rule
Smartcheck

External NLP Services
- Spell Check
- Syntax Parser
- Word Aligner

Annotated
- Annotation Tool

Eval
- Spelling
- Register
- Formality
- Consistency
- Client Rule

Learn
Pos-Editing Interfaces

Please fix the remaining issues before submitting

- Spelling: ... registrar **novamente** seus...
- Inconsistency Issues:
  - Sed do eiusmod tempor incididunt ut labore et doler magna aliqua. Ut enim ad minim veniam.

3 issues detected

In that way we will be able to register again your tastes and provide you with awesome things.

Dessa forma, **seremos capazes** de registrar **novamente** seus gostos e fornecê-lo com sugestões impressionantes.
Thank you!

helena@unbabel.com