Understanding and Implementing Effective Translation Quality Evaluation Techniques

Stephen Doherty & Federico Gaspari

Centre for Next Generation Localisation
Dublin City University, Ireland

October 3rd, 2013
Outline

1. Critical overview of commonly used human and automatic translation evaluation models and metrics;

2. Effective implementation of translation quality assessment methods and tools;

3. Further information, training materials and opportunities.
Human Evaluation
Commonly Models & Metrics

• Overview of models and metrics used for evaluating human and machine translation and localisation:
  o based on publicly available documentation, software or demos.

• Three parts:
  o manual evaluation metrics and tools for human translation;
  o automatic evaluation tools for human translation;
  o quality metrics for machine translation (manual, automatic, diagnostic).

• Some key terms:
  o **error**: an incorrect aspect of a translation that may be corrected;
  o **metric**: a method for assessing the translation quality;
  o **quality**: the degree to which a translation meets the threshold for errors set in a metric;
  o **validity** is key (i.e. relationship between error-metric-quality is circular!).
A human reviewer scores the translation by checking it and looking for errors;

Error rate vs. rubric models;

**Error rate models** (more common):

- based on error counting, a subtractive model of translation quality;
- a score indicates the quality, usually as a percentage;
- errors subtract from the overall score (“perfect” translation 100%);
- difference in the importance of errors is recognised (e.g. minor spelling mistake vs. completely mistranslated term):
  - **severity**: how serious the error is, usually based on a 2- or 3-tier system, according to the perceived/anticipated impact of the error on the target text reader;
  - **weight**: a numerical multiplier corresponding to how important the error is.
Manual evaluation metrics and tools for human translation (cont’d)

- The SDL TMS Classic Model has neither severity nor weight, counting all the errors with an undifferentiated value of 1;
- The LISA QA Model features 3 severity levels, but no weighting;
- The SAE J2450 Model combines both, with different point values for error type and severity.

- Negative quality scores are possible, with no pre-determined lower bound;
- Scores are subtracted from the total number of words; if the score falls below a certain minimum threshold (say, 95%), the translation is rejected.
• Rubric models (less common):
  o These models have an additive nature;
  o Starting from zero, points are added if the translation meets requirements specified in the metric;
  o Quality is achieved by adding up positive factors:
    • e.g. on a 0-10 quality scale, 1 point is added for adherence to specified terminology, another point for matching a pre-defined style sheet, etc.
  o Rubric models seem to be still uncommon in commercial translation and localisation, being largely confined to academia. Hence, they are not discussed further here, but they might become more common in the future.
  o In theory, blended (error / rubric) models are possible, but not (yet) popular.
LISA QA Model, version 3.1

- Developed by the Localization Industry Standards Association (LISA) in the 1990s;
- “LISA QA Model” can refer both to the software product and to the abstract model, which is also implemented in other localisation tools;
- Intended for use by human reviewers in assessing a localised text or product; most of its categories are not suitable for automatic assessment as they rely on human judgment;
- It addresses both linguistic errors and other types of errors (e.g. layout, functional errors, etc.), and includes a number of process categories;
- It was never released as a standard, although it enjoyed the status of a de facto industry standard around the time of LISA’s insolvency (March 2011).
Scores are usually expressed as a percentage, and calculated based on a sample of the text (up to a 100% sample, but generally smaller).

The Model consists of a set of 20, 25 or 123 error categories, depending on how they are counted, i.e.

- software user interface: 25 categories in 7 broad groups with some repetition;
- documentation: 20 categories, with some repetition under different headings:
  - the documentation also presents a hierarchy of error types, with 123 terminal nodes.
Error categories in the user interface:

<table>
<thead>
<tr>
<th>Documentation</th>
<th>Help</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc Language (7)</td>
<td>Help Formatting (4)</td>
<td>Software Formatting (5)</td>
</tr>
<tr>
<td>Mistranslation</td>
<td>Index</td>
<td>Graphics</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Layout</td>
<td>Alignment</td>
</tr>
<tr>
<td>Terminology</td>
<td>Typography</td>
<td>Sizing</td>
</tr>
<tr>
<td>Language</td>
<td>Graphics</td>
<td>Truncation/Overlap</td>
</tr>
<tr>
<td>Style</td>
<td>Help Formatting – Asian (4)</td>
<td>Character Formatting</td>
</tr>
<tr>
<td>Country</td>
<td>Index</td>
<td></td>
</tr>
<tr>
<td>Consistency</td>
<td>(Layout)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Graphics</td>
<td></td>
</tr>
<tr>
<td>Doc Formatting (6)</td>
<td>(Typography)</td>
<td></td>
</tr>
<tr>
<td>TOC</td>
<td>Graphics</td>
<td></td>
</tr>
<tr>
<td>Index</td>
<td>(Typography)</td>
<td></td>
</tr>
<tr>
<td>Layout</td>
<td>Localizable Text</td>
<td></td>
</tr>
<tr>
<td>Typography</td>
<td>HyperText functionality, jumps, popups</td>
<td></td>
</tr>
<tr>
<td>Graphics</td>
<td>Graphics</td>
<td></td>
</tr>
<tr>
<td>Call Outs and Captions</td>
<td>Call Outs and Captions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Double/Single Size</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Punctuation Marks</td>
<td></td>
</tr>
<tr>
<td>Doc Formatting – Asian (8)</td>
<td>TOC</td>
<td>Software Functionality Testing (5)</td>
</tr>
<tr>
<td></td>
<td>Index</td>
<td>Localizable Text</td>
</tr>
<tr>
<td></td>
<td>(Layout)</td>
<td>Dialogue Functionality</td>
</tr>
<tr>
<td></td>
<td>Graphics</td>
<td>Menu Functionality</td>
</tr>
<tr>
<td></td>
<td>(Typography)</td>
<td>Hotkeys/Accelerators</td>
</tr>
<tr>
<td></td>
<td>Localizable Text</td>
<td>Jumps/Links</td>
</tr>
</tbody>
</table>

21 total 8 total 10 total
TOTAL: 39 total (25 unique)
Division into top-level error categories in the documentation:

- General language-related error categories:
  - Language, Style, Country;

- Document-related error categories:
  - Document Language, Document Formatting, Document Functional;

- Software-specific error categories:
  - Software Language, Software Formatting, Software Functional.
Comparison of error categories in the LISA QA Model software vs. documentation:

<table>
<thead>
<tr>
<th>In both UI and documentation</th>
<th>In UI only</th>
<th>In documentation only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>Alignment</td>
<td>Arts &amp; Design</td>
</tr>
<tr>
<td>Call Outs and Captions</td>
<td>Character Formatting</td>
<td>Compatibility</td>
</tr>
<tr>
<td>Consistency</td>
<td>Dialogue Functionality</td>
<td>Cross-references</td>
</tr>
<tr>
<td>Country</td>
<td>Double/Single Size</td>
<td>Delivery</td>
</tr>
<tr>
<td>Graphics</td>
<td>Hotkeys/Accelerators</td>
<td>Functionality</td>
</tr>
<tr>
<td>Index</td>
<td>Hyper text functionality, jumps, popups</td>
<td>Interface</td>
</tr>
<tr>
<td>Language</td>
<td>Jumps/Links</td>
<td>Output</td>
</tr>
<tr>
<td>Style</td>
<td>Layout</td>
<td>Printing</td>
</tr>
<tr>
<td>Table of Contents (TOC)</td>
<td>Localizable Text</td>
<td>References</td>
</tr>
<tr>
<td>Typography</td>
<td>Menu Functionality</td>
<td>Terminology</td>
</tr>
<tr>
<td></td>
<td>Mistranslation</td>
<td>Typography</td>
</tr>
<tr>
<td></td>
<td>Punctuation Marks</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sizing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Terminology</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Truncation/Overlap</td>
<td></td>
</tr>
</tbody>
</table>

Total: 10  Total: 15  Total: 11
LISA QA Model, version 3.1

- LISA QA Model user interface

Three severity levels:
- Minor: 1 point
- Major: 5 points
- Critical: 10 points

Minimum threshold: 13
• The chosen text sample is scored manually by a reviewer: if it remains above a pre-defined threshold, the translation is deemed to pass; if it falls below the predetermined threshold, the translation is deemed to fail;

• The software only counts the errors: it does not flag them for change or provide any sort of audit trail.
SAE J2450

• Originally developed by the Society of Automotive Engineers (SAE) to meet the needs of the automotive industry for quality checking translations;

• Not as widely used in the broader translation and localisation industry as the LISA QA Model, but still one of the most common error metrics in use;

• It has the advantage of simplicity, but focuses strictly on linguistic quality;

• It is designed for use by human assessors (usually examining a sample from a larger text):
  - 7 error categories;
  - 2 severity levels with point weights.
The choice between Minor and Major weights and which category to select are determined through two “meta-rules”:

- if in doubt on the error category (ambiguity), choose the first one in the list;
- if in doubt as to whether the error is minor or major, consider it major.
The SDL TMS Classic Model was developed for human review in SDL products and is intended for human rather than automatic checking;

It foregoes any weighting or severity mechanisms, and consists of 7 error categories, each with a numerical value of 1:

- Overall Grading;
- Translation Accuracy;
- Grammar;
- Spelling;
- Style;
- Terminology;
- Technical Accuracy.
• ATA has issued formal criteria for the evaluation of certification exams;

• A point-based scoring system is used, based on 3 top-level error types (these are further sub-divided):
  o errors concerning the form of the exam;
  o translation/strategic/transfer errors - negative impact on understanding/use of target text;
  o mechanical errors - negative impact on overall quality of target text.

• For further information, see “ATA’s Framework for Standardized Error Marking Explanation of Error Categories”:
  o [www.atanet.org/certification/aboutexams_error.php](http://www.atanet.org/certification/aboutexams_error.php)
Summary of human translation quality assessment metrics

- They all rely on bilingual assessors detecting discrepancies in meaning and/or form between source and target text;
- Many of the error categories are actually general language errors (e.g. misspellings, grammatical errors, etc.) that do not require bilingual knowledge;
- Only one error category is addressed by all the models (terminology errors), i.e. there is little overlap of error categories between models and metrics.
Summary of human translation quality assessment metrics

• The issue of **validity:**
  - To what extent do errors alone (in error rate, subtractive, models) provide an accurate measure of (perceived) translation quality?

• Additional issue of **reliability:**
  - Formal quality assessment methods are useful if they help quality assessment become reliable, i.e. its results are consistent and repeatable:
    - what is the role played by the subjectivity (experience, training, etc.) of the assessors?
    - reliability may be in inverse relationship with the complexity of the metric.

• To avoid subjectivity and the high costs generally involved in human translation quality assessment, automatic evaluation tools can be used.
Quality Metrics for Machine Translation
Quality metrics for machine translation

- Manual MT evaluation metrics;
- Automatic MT evaluation metrics;
- MT error classification and diagnostic MT evaluation.
Manual MT evaluation metrics

• Adequacy and fluency (Koehn, 2009: 218):
  o **Adequacy** – does the MT output provide the same meaning as the input? How well does the MT output preserve the meaning of the input?
  o **Fluency** – does the output read well? How well does the MT output read?
  o Such criteria can be evaluated by human judges, who need to be instructed as to how these notions are defined in practice for a specific evaluation task, e.g.
    ▪ fluency and adequacy can be judged according to N-point scale scores;
    ▪ comparison of the MT output to a target-language reference, as opposed to the input text in the source language, so that monolingual assessors can perform the evaluation;
    ▪ adequacy and fluency can also be used to rank MT output (e.g. sentences).

• Utility- and task-based evaluation:
  o Raw MT output evaluated, e.g.
    ▪ in terms of subsequent post-editing effort to obtain a publishable translation (productivity gains and comparative evaluations);
    ▪ in terms of success in finding information or performing a task (instructions manual).
• Arguably the most popular automatic MT evaluation metric;
• Like others, it is based on the premise that MT quality can be evaluated on the basis of how similar MT output is to human translation, considered as a reference;
• It needs a corpus of human translations for the same source text that is to be translated by means of MT;
• It measures the percentage of matching n-grams in both translations, normalised by the number of n-grams in the MT output;
• It is a precision-oriented metric, with a brevity penalty to prevent very short translations (which miss many of the reference n-grams) to score very high;
• Having multiple reference translations helps;
• Not comparable across datasets.
Automatic MT evaluation metrics

NIST

• A modification of BLEU addressing some of its limitations, in particular:
  o the way n-grams are weighed;
  o the way n-gram scores are averaged;
  o the way the brevity penalty is computed.

• NIST assigns greater value than BLEU to information-rich n-grams (the more rare ones in the corpus).
Automatic MT evaluation metrics

METEOR

- It includes both precision and recall components, and an explicit fragmentation score to account for word order;
- It enhances token matching considering stemming, synonymy and paraphrase lookup;
- If multiple references are available, the matches are computed against each reference separately, and the best match is selected;
- METEOR allows the unigram matches to be exact word matches, or generalized to stems, synonyms and paraphrases, provided that the required language resources are available;
- Due to the reliance on specific language resources, METEOR can only be applied in full to certain languages, due to the limited availability of synonymy and paraphrase modules.
Automatic MT evaluation metrics

Error Rates

- *Edit or error rate measures* compute the amount of changes that must be applied to the MT output in order to transform it into a reference translation, hence indicating the required post-editing effort:
  - **Word Error Rate (WER)** is based on the Levenshtein edit distance and computes the minimum number of *substitutions*, *deletions* and *insertions* that have to be performed to convert MT output into the human reference;
  - **Position-independent Word Error Rate** does not penalise reorderings;
  - **Translation Edit Rate (TER)** also considers *shifts* of word sequences, and all edits have equal cost;
  - **Translation Edit Rate Plus (TERp)** is an extension of TER which is language-dependent and requires specific language resources;
  - **Human-targeted Translation Edit Rate (HTER)** is a semi-automatic variation of TER: the references are built as human-corrected versions of the MT output to guarantee that the edit rate is measured as the minimum of edits necessary to make the system output a fluent and adequate translation.
Summary of automatic MT evaluation metrics

- Automatic MT evaluation metrics do not provide any insights into the severity of the errors produced by the system;
- The correlation of the scores provided by automatic MT evaluation metrics with human judgement can vary as a function of the language pair, translation direction, text type, domain, size of the datasets, number (and expertise) of the evaluators, etc.
- What do they mean?
- Are they truly objective and consistent?
- What does this mean in practice and industry usage?
Implementation of Evaluation Metrics
Automatic quality evaluation tools for human translation

• In this scenario, error candidates are automatically detected and must then be reviewed by a human to take further action.

• **Acrocheck**
  o Commercial software package developed by Acrolinx for authoring assistance and checking of texts for potential issues;
  o It is used primarily in the authoring phase to identify issues in source texts that may cause problems for readers and in translation;
  o It addresses general language issues with a few translation-specific rules (separate rules for specific languages);
  o No check on consistency between source and target or to identify issues caused by the translation process.
Automatic quality evaluation tools for human translation (cont’d)

- **ApSIC XBench**
  - Commercial tool distributed under open license terms that include free use in non-commercial settings;
  - It performs a number of functions besides quality assurance, and includes quality checks on 3 levels: basic (e.g. untranslated segments), content (e.g. repeated word) and checklists (project-specific checklist).

- **CheckMate**
  - Component of the open-source Okapi toolkit;
  - It is intended for automatic checking in an XLIFF environment;
  - It does not impose any scoring metric, but it identifies possible errors for human action;
  - It includes 23 error categories (e.g. “missing target translation unit”, “extra leading white space”, “target same as source”);
  - Some error categories can be user-defined or rely on user-supplied glossary file.
Automatic quality evaluation tools for human translation (cont’d)

• **LanguageTool**
  - Open-source, community-maintained grammar/language checker;
  - Currently, 29 languages are supported to varying degrees;
  - Independent communities develop language-specific rule sets;
  - Available as a stand-alone Java tool and as a plugin for OpenOffice.

• **QA Distiller**
  - Commercial tool from Yamagata Europe;
  - It identifies classes of errors for human evaluation and provides tools for fixing them;
  - It has a hierarchical list of error classes divided into 5 broad categories:
    - formatting (language-independent and language-dependent);
    - omissions (e.g. empty or partial translations);
    - inconsistencies;
    - terminology;
    - general.
• **XLIFF:doc**
  - XLIFF-based format created by the Interoperability Now! group to promote interoperability of localization-related data;
  - This is not really a tool, but it defines data categories for quality assurance (e.g. date, inconsistency, number, omission, punctuation, tags, terminology, etc.);
  - 3 levels of severity are provided: *error*, *warning* and *non-error*. 
Overview of automatic translation quality assessment metrics

- One open issue concerns which (types of) potential errors can be detected automatically (+) and which ones cannot (-), for example:
  - Terminology/Glossary adherence (+);
  - Omission (+);
  - Punctuation (+);
  - Consistency (+).
  - Mistranslation (-);
  - Style (-);
  - Accuracy (-);
  - Faithfulness (-).

- More broadly, errors can be broken down into the following groups:
  - Language-related errors (comparing source and target text), i.e.
    - categories requiring human checking/error detection;
    - categories that can be checked automatically (e.g. formal errors such as incorrect dates, untranslated segments).
  - Non-linguistic errors (e.g. related to graphics, user interface design, functionality):
    - formatting;
    - functionality;
    - project.
MT error classification and diagnostic MT evaluation

- Limitations of automatic MT metrics;
- Drawbacks of manual evaluation;
- MT error classification and diagnostic MT evaluation provide more fine-grained insights into the linguistic performance of systems;
- May focus on detecting frequent and/or serious classes of errors in MT output;
- Can be used to improve the system itself via re-training, etc.;
- Can lessen repetitive corrections, e.g. in post-editing;
- Collection of valuable data – human in the loop processes.
Vilar et al. (2006) present a hierarchical error classification, that can be more indicative of the quality of MT systems:

- **Errors**
  - **Missing Words**
    - Content Words
    - Filler Words
  - **Word Order**
    - Word Level
      - Local Range
      - Long Range
    - Phrase Level
      - Local Range
      - Long Range
  - **Incorrect Words**
    - Sense
      - Wrong Lexical Choice
      - Incorrect Disambiguation
    - Incorrect Form
    - Extra Words
    - Style
    - Idioms
  - **Unknown Words**
    - Unknown Stem
    - Unseen Forms
  - **Punctuation**
Available from:


- Easy to adapt for basic classifications;
- Useful for MT but limited for human translation and requires a lot of human resources;
- Prone to subjective comprehension and use of the metric;
- Requires more detailed, perhaps customised, in-house criteria.
And then there were three…

- **Rubrics:**
  - evaluation on a numerical scale based on set criteria.

- **Full evaluation and error analysis:**
  - evaluation of the entire text by a human evaluator – sometimes essential due to legal requirements.

- **Sampling:**
  - taking a portion of the translated text, either randomly or by specific sections, to decide if it meets requirements and, consequently, the entire text.
Multidimensional Quality Metrics

• Basic of the project, e.g. language, client, field, tech., audience;
• MQM recommends metric(s) based on these & additional criteria that can be added – it then gives:
  o A predefined or customised metric(s);
  o Weighting and specification file that can be shared across projects, etc.
• Provides in-line mark-up and full reporting features;
• Can be implemented within a workflow or used in an open-source tool – www.translate5.net
• Measures quality as formula of adequacy, fluency, and verity
• Takes source text and its quality into account;
• Provides a more fine-grained tier, e.g. for error analysis, re-training, training, etc.
MQM

Issue Types

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Verity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terminology*</td>
<td>Legal requirements</td>
</tr>
<tr>
<td>Mistranslation</td>
<td>Locale applicability</td>
</tr>
<tr>
<td>Omission*</td>
<td></td>
</tr>
<tr>
<td>Addition*</td>
<td></td>
</tr>
<tr>
<td>Untranslated*</td>
<td></td>
</tr>
<tr>
<td>Completeness</td>
<td></td>
</tr>
</tbody>
</table>

Fluency

<table>
<thead>
<tr>
<th>(Content)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Register*</td>
</tr>
<tr>
<td>Style*</td>
</tr>
<tr>
<td>Inconsistency</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(Mechanical)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spelling*</td>
</tr>
<tr>
<td>Typography*</td>
</tr>
<tr>
<td>Grammar*</td>
</tr>
<tr>
<td>Locale violation*</td>
</tr>
</tbody>
</table>

Unintelligible
Training Materials & Resources
Tools for MT error analysis and diagnostic MT evaluation

- **Addicter:**
  - open-source tool that automatically detects and displays common translation errors;

- **AMEANA:**
  - open-source tool for error analysis in NLP tasks, specifically geared towards morphologically rich languages; it provides a detailed error analysis and is easily adaptable to any language for which a morphological analyser is available;

- **BLAST:**
  - open-source tool for error analysis in MT output, designed to be flexible: it can be used with any MT system, language pair and kind of error;

- **DELiC4MT:**
  - free open-source language-independent toolkit for the diagnostic evaluation of MT systems based on linguistic checkpoints, i.e. a user-defined set of linguistic phenomena;
Tools for MT error analysis and diagnostic MT evaluation (cont’d)

• **Hjerson:**
  o tool for automatic classification of errors in MT output, which supports the detection of five word-level error classes: morphological, reordering, omissions, unnecessary additions and lexical errors;

• **TerrorCat:**
  o this tool uses frequencies of automatically obtained translation error categories as a basis to conduct the pairwise comparison of translation hypotheses, which, in turn, is used to generate a score for every translation;

• **Woodpecker:**
  o proprietary tool supporting diagnostic evaluation based on linguistic checkpoints, i.e. particularly interesting or challenging linguistic phenomena for MT processing identified by the MT user or developer (e.g. ambiguous words, complex POS-n-gram constructs, specific collocations that are difficult to translate, etc.).
Other Online Resources

• Toolkits of Several Metrics:
  o Asiya - http://asiya.lsi.upc.edu

• Specific Metrics:
  o Edit Distance - http://www.let.rug.nl/kleiweg/lev/

• More information on MQM, quality estimation, and industry reports on translation technology, including evaluation and training - http://www.qt21.eu/launchpad/content/training
Thank you for your attention!

stephen.doherty@dcu.ie            fgaspari@computing.dcu.ie

www.qt21.eu

QTLaunchPad on