Translation Quality Assessment: Evaluation and Estimation

Lucia Specia

University of Sheffield
l.specia@sheffield.ac.uk

9 September 2013
“MT Evaluation is better understood than MT” (Carbonell and Wilks, 1991)
Outline

1. Translation Quality
2. QTLaunchPad Project
3. Quality Estimation
4. State of the art in QE
5. Open issues
6. Conclusions
Overview

- What does **quality** mean?
  - Fluent?
  - Adequate?
  - Easy to post-edit?
Overview

- **What does quality mean?**
  - Fluent?
  - Adequate?
  - Easy to post-edit?

- **Quality for whom?**
  - End-user
  - MT-system (tuning)
  - Post-editor
  - Other applications (e.g., CLIR)
Overview

- What does **quality** mean?
  - Fluent?
  - Adequate?
  - Easy to post-edit?
- Quality for **whom**?
  - End-user
  - MT-system (tuning)
  - Post-editor
  - Other applications (e.g. CLIR)
- Quality for **what**?
  - Internal communications
  - Dissemination (publishing)
  - Gisting (Google Translate)
  - Draft translations (light vs heavy post-editing)
  - MT system improvement (diagnosis)
ref  Do **not** buy this product, it’s their craziest invention!
sys  Do buy this product, it’s their craziest invention!
ref Do **not** buy this product, it’s their craziest invention!
sys Do *buy* this product, it’s their craziest invention!

- **Severe** if end-user does not speak source language
- **Trivial** to post-edit by translators

The battery lasts 6 hours and it can be fully recharged in 30 minutes. Six-hours battery, 30 minutes to full charge last.
Overview

ref Do **not** buy this product, it’s their craziest invention!
sys Do buy this product, it’s their craziest invention!

- **Severe** if end-user does not speak source language
- **Trivial** to post-edit by translators

ref The **battery lasts** 6 hours and it can be **fully recharged** in **30 minutes**.
sys Six-hours **battery**, **30 minutes** to **full charge last**.
Overview

ref Do **not** buy this product, it’s their craziest invention!
sys Do buy this product, it’s their craziest invention!

- **Severe** if end-user does not speak source language
- **Trivial** to post-edit by translators

ref The **battery lasts 6 hours** and it can be **fully recharged** in 30 minutes.
sys Six-hours **battery**, 30 minutes to **full charge last**.

- **Ok** for gisting - meaning preserved
- **Very costly** for post-editing if style is to be preserved
Overview

How do we measure quality?

- **Human metrics**: error counts (which?), ranking, acceptability, 1-N fluency/adequacy
- **Automatic metrics** based on human references: (BLEU, METEOR, TER, etc.)
- **Semi-automatic metrics** based on post-editions: HTER, PE time, eye-tracking, etc.
Overview

How do we **measure** quality?

- **Human metrics**: error counts (which?), ranking, acceptability, 1-N fluency/adequacy
- **Automatic metrics** based on human references: (BLEU, METEOR, TER, etc.)
- **Semi-automatic metrics** based on post-editions: HTER, PE time, eye-tracking, etc.
- **Automatic metrics** without references: **quality estimation**
QTLaunchPad project

http://www.qt21.eu/launchpad/

- Multidimensional Quality Metrics (MQM) based on a specification
- Machine and human translation quality
- Manual and (semi-)automatic assessment
- Takes quality of source text into account
- MT system improvement, gisting, dissemination, etc.
Multidimensional Quality Metrics (MQM)

Issues selected based on a given **specification** (dimensions):

- Language/locale
- Subject field/domain
- Terminology
- Text Type
- Audience
- Purpose
- Register
- Style
- Content correspondence
- Output modality, ...
**Issue types (core):**

- **Accuracy**
  - Terminology*
  - Mistranslation
  - Omission*
  - Addition*
  - Untranslated*

- **Fluency**
  - Register*
  - Style*
  - Inconsistency
  - (Content)
  - Spelling*
  - Typography*
  - Grammar*
  - Locale violation*

- **Verity**
  - Unintelligible
  - Completeness
  - Legal requirements
  - Locale applicability

Translation Quality Assessment: Evaluation and Estimation
**Multidimensional Quality Metrics (MQM)**

**Issue types**: http://www.qt21.eu/launchpad/content/high-level-structure-0

**Combining issue types**:

\[ TQ = 100 - AccP - (FluP_T - FluP_S) - (VerP_T - VerP_S) \]
Overview

- **Quality estimation** (QE): metrics that provide an estimate on the quality of unseen translations
Overview

- **Quality estimation** (QE): metrics that provide an estimate on the quality of unseen translations
- Quality defined by the data
Overview

- **Quality estimation** (QE): metrics that provide an estimate on the quality of unseen translations
- Quality defined by the data

Quality = **Can we publish it as is?**
Overview

- **Quality estimation** (QE): metrics that provide an estimate on the quality of unseen translations
- Quality defined by the data

Quality = **Can we publish it as is?**

Quality = **Can a reader get the gist?**
Overview

- **Quality estimation** (QE): metrics that provide an estimate on the quality of unseen translations
- Quality defined by the data

Quality = Can we publish it as is?

Quality = Can a reader get the gist?

Quality = Is it worth post-editing it?
Overview

- **Quality estimation** (QE): metrics that provide an estimate on the quality of unseen translations

  - Quality defined by the **data**

  
  \[
  \text{Quality} = \begin{align*}
  &\text{Can we publish it as is?} \\
  &\text{Can a reader get the gist?} \\
  &\text{Is it worth post-editing it?} \\
  &\text{How much effort to fix it?}
  \end{align*}
  \]
Overview

- **Quality estimation** (QE): metrics that provide an estimate on the quality of unseen translations
- Quality defined by the data

Quality = **Can we publish it as is?**

Quality = **Can a reader get the gist?**

Quality = **Is it worth post-editing it?**

Quality = **How much effort to fix it?**

Quality = **What’s this translation’s MQM score?**
Translation Quality Assessment: Evaluation and Estimation

Framework

X: examples of source & translations

Feature extraction

Features

Y: Quality scores for examples in X

Machine Learning

QE model
Translation Quality Assessment: Evaluation and Estimation

Framework

MT system → Translation for $x_t'$ → Feature extraction

Source Text $x_s'$ → Features

Quality score $y'$ → QE model
Main components to build a QE system:

1. Definition of quality: **what to predict**
2. (Human) labelled **data** (for quality)
3. **Features**
4. Machine learning **algorithm**
Definition of quality

- Predict 1-N **absolute** scores for adequacy/fluency
- Predict 1-N **absolute** scores for post-editing effort
- Predict average post-editing **time** per word
- Predict **relative** rankings
- Predict **relative** rankings for same source
- Predict **percentage of edits** needed for sentence
- Predict word-level **edits** and its types
- Predict **BLEU**, etc. scores for document
Datasets

- **SHEF** (several): http://staffwww.dcs.shef.ac.uk/people/L.Specia/resources.html


Features

- Source text
- MT system
- Translation

Adequacy indicators

- Complexity indicators
- Confidence indicators
- Fluency indicators
** Goal: ** framework to explore features for QE

- **Feature extractors** for 150+ features of all types: Java
- **Machine learning**: Gaussian Processes & scikit-learn toolkit (Python), with wrappers for a number of algorithms, grid search, feature selection

Open source: [http://www.quest.dcs.shef.ac.uk/](http://www.quest.dcs.shef.ac.uk/)
Outline

1. Translation Quality
2. QTLaunchPad Project
3. Quality Estimation
4. State of the art in QE
5. Open issues
6. Conclusions
Shared Task

- **WMT12-13** – with Radu Soricut & Christian Buck
Shared Task

- **WMT12-13** – with Radu Soricut & Christian Buck
- **Sentence-** and **word-level** estimation of **PE effort**
Shared Task

- **WMT12-13** – with Radu Soricut & Christian Buck
- **Sentence-** and **word-level** estimation of **PE effort**
- Datasets and **language pairs:**

<table>
<thead>
<tr>
<th>Quality</th>
<th>Year</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-5 subjective scores</td>
<td>WMT12</td>
<td>en-es</td>
</tr>
<tr>
<td>Ranking all sentences best-worst</td>
<td>WMT12/13</td>
<td>en-es</td>
</tr>
<tr>
<td>HTER scores</td>
<td>WMT13</td>
<td>en-es</td>
</tr>
<tr>
<td>Post-editing time</td>
<td>WMT13</td>
<td>en-es</td>
</tr>
<tr>
<td>Word-level edits: change/keep</td>
<td>WMT13</td>
<td>en-es</td>
</tr>
<tr>
<td>Word-level edits: keep/delete/replace</td>
<td>WMT13</td>
<td>en-es</td>
</tr>
<tr>
<td>Ranking 5 MTs per source</td>
<td>WMT13</td>
<td>en-es; de-en</td>
</tr>
</tbody>
</table>
Shared Task

- **WMT12-13** – with Radu Soricut & Christian Buck
- **Sentence- and word-level** estimation of **PE effort**
- Datasets and **language pairs**:

<table>
<thead>
<tr>
<th>Quality</th>
<th>Year</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-5 subjective scores</td>
<td>WMT12</td>
<td>en-es</td>
</tr>
<tr>
<td>Ranking all sentences best-worst</td>
<td>WMT12/13</td>
<td>en-es</td>
</tr>
<tr>
<td>HTER scores</td>
<td>WMT13</td>
<td>en-es</td>
</tr>
<tr>
<td>Post-editing time</td>
<td>WMT13</td>
<td>en-es</td>
</tr>
<tr>
<td>Word-level edits: change/keep</td>
<td>WMT13</td>
<td>en-es</td>
</tr>
<tr>
<td>Word-level edits: keep/delete/replace</td>
<td>WMT13</td>
<td>en-es</td>
</tr>
<tr>
<td>Ranking 5 MTs per source</td>
<td>WMT13</td>
<td>en-es; de-en</td>
</tr>
</tbody>
</table>

- Evaluation metric:

\[
\text{MAE} = \frac{\sum_{i=1}^{N} |H(s_i) - V(s_i)|}{N}
\]
Baseline system

Features:

- number of tokens in the source and target sentences
- average source token length
- average number of occurrences of words in the target
- number of punctuation marks in source and target sentences
- LM probability of source and target sentences
- average number of translations per source word
- % of source 1-grams, 2-grams and 3-grams in frequency quartiles 1 and 4
- % of seen source unigrams
Baseline system

**Features:**

- number of tokens in the source and target sentences
- average source token length
- average number of occurrences of words in the target
- number of punctuation marks in source and target sentences
- LM probability of source and target sentences
- average number of translations per source word
- % of source 1-grams, 2-grams and 3-grams in frequency quartiles 1 and 4
- % of seen source unigrams

**SVM regression** with RBF kernel with the parameters $\gamma$, $\epsilon$ and $C$ optimised using a grid-search and 5-fold cross validation on the training set
## Results - scoring sub-task (WMT12)

<table>
<thead>
<tr>
<th>System ID</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDLLW_M5P_bestDeltaAvg</td>
<td>0.61</td>
<td>0.75</td>
</tr>
<tr>
<td>UU_best</td>
<td>0.64</td>
<td>0.79</td>
</tr>
<tr>
<td>SDLLW_SVM</td>
<td>0.64</td>
<td>0.78</td>
</tr>
<tr>
<td>UU_bltk</td>
<td>0.64</td>
<td>0.79</td>
</tr>
<tr>
<td>Loria_SVM_linear</td>
<td>0.68</td>
<td>0.82</td>
</tr>
<tr>
<td>UEdin</td>
<td>0.68</td>
<td>0.82</td>
</tr>
<tr>
<td>TCD_M5P-resources-only*</td>
<td>0.68</td>
<td>0.82</td>
</tr>
<tr>
<td><strong>Baseline bb17 SVR</strong></td>
<td>0.69</td>
<td>0.82</td>
</tr>
<tr>
<td>Loria_SVM_rbf</td>
<td>0.69</td>
<td>0.83</td>
</tr>
<tr>
<td>SJTU</td>
<td>0.69</td>
<td>0.83</td>
</tr>
<tr>
<td>WLV-SHEF_FS</td>
<td>0.69</td>
<td>0.85</td>
</tr>
<tr>
<td>PRHLT-UPV</td>
<td>0.70</td>
<td>0.85</td>
</tr>
<tr>
<td>WLV-SHEF_BL</td>
<td>0.72</td>
<td>0.86</td>
</tr>
<tr>
<td>DCU-SYMC_unconstrained</td>
<td>0.75</td>
<td>0.97</td>
</tr>
<tr>
<td>DFKI_grcfs-mars</td>
<td>0.82</td>
<td>0.98</td>
</tr>
<tr>
<td>DFKI_cfs-plsreg</td>
<td>0.82</td>
<td>0.99</td>
</tr>
<tr>
<td>UPC_1</td>
<td>0.84</td>
<td>1.01</td>
</tr>
<tr>
<td>DCU-SYMC_constrained</td>
<td>0.86</td>
<td>1.12</td>
</tr>
<tr>
<td>UPC_2</td>
<td>0.87</td>
<td>1.04</td>
</tr>
<tr>
<td>TCD_M5P-all</td>
<td>2.09</td>
<td>2.32</td>
</tr>
</tbody>
</table>
## Results - scoring sub-task \textbf{(WMT13)}

<table>
<thead>
<tr>
<th>System ID</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHEF FS</td>
<td>12.42</td>
<td>15.74</td>
</tr>
<tr>
<td>SHEF FS-AL</td>
<td>13.02</td>
<td>17.03</td>
</tr>
<tr>
<td>CNGL SVRPLS</td>
<td>13.26</td>
<td>16.82</td>
</tr>
<tr>
<td>LIMSI</td>
<td>13.32</td>
<td>17.22</td>
</tr>
<tr>
<td>DCU-SYMC combine</td>
<td>13.45</td>
<td>16.64</td>
</tr>
<tr>
<td>DCU-SYMC alltypes</td>
<td>13.51</td>
<td>17.14</td>
</tr>
<tr>
<td>CMU noB</td>
<td>13.84</td>
<td>17.46</td>
</tr>
<tr>
<td>CNGL SVR</td>
<td>13.85</td>
<td>17.28</td>
</tr>
<tr>
<td>FBK-UEdin extra</td>
<td>14.38</td>
<td>17.68</td>
</tr>
<tr>
<td>FBK-UEdin rand-svr</td>
<td>14.50</td>
<td>17.73</td>
</tr>
<tr>
<td>LORIA inctrain</td>
<td>14.79</td>
<td>18.34</td>
</tr>
<tr>
<td><strong>Baseline bb17 SVR</strong></td>
<td>14.81</td>
<td>18.22</td>
</tr>
<tr>
<td>TCD-CNGL open</td>
<td>14.81</td>
<td>19.00</td>
</tr>
<tr>
<td>LORIA inctraincont</td>
<td>14.83</td>
<td>18.17</td>
</tr>
<tr>
<td>TCD-CNGL restricted</td>
<td>15.20</td>
<td>19.59</td>
</tr>
<tr>
<td>CMU full</td>
<td>15.25</td>
<td>18.97</td>
</tr>
<tr>
<td>UMAC</td>
<td>16.97</td>
<td>21.94</td>
</tr>
</tbody>
</table>
Outline

1. Translation Quality
2. QTLaunchPad Project
3. Quality Estimation
4. State of the art in QE
5. Open issues
6. Conclusions
Agreement between annotators

- **Absolute value judgements**: difficult to achieve consistency even in highly controlled settings
  - 30% of initial dataset discarded
  - Remaining annotations had to be scaled
Agreement between annotators

- **Absolute value judgements**: difficult to achieve consistency even in highly controlled settings
  - 30% of initial dataset discarded
  - Remaining annotations had to be scaled

- **More objective absolute scores**
  - Post-editing time, HTER, edits
  - Also subject to huge variance (WPTP 2013, Wisniewski et al., MT-Summit 2013)
  - Multi-task learning to address this variance (Cohn and Specia, ACL 2013)
Agreement between annotators

- **Absolute value judgements**: difficult to achieve consistency even in highly controlled settings
  - 30% of initial dataset discarded
  - Remaining annotations had to be scaled

- **More objective absolute scores**
  - Post-editing time, HTER, edits
  - Also subject to huge variance (WPTP 2013, Wisniewski et. al, MT-Summit 2013)
  - Multi-task learning to address this variance (Cohn and Specia, ACL 2013)

- **Relative scores**
  - Different task altogether
  - WMT13: better results than reference-based metrics
Annotation costs

**Active learning** to select subset of instances to be annotated (Beck et al., ACL 2013)
**Feature selection** to identify relevant info for dataset (Shah et al., MT Summit 2013)
Curse of dimensionality

**Feature selection** to identify relevant info for dataset (Shah et al., MT Summit 2013)

Common feature set identified, but nuanced subsets for specific datasets
How to use estimated PE effort scores?

Do users prefer detailed estimates (sub-sentence level) or an overall estimate for the complete sentence or not seeing bad sentences at all?

- Too much information vs hard-to-interpret scores
How to use estimated PE effort scores?

Do users prefer **detailed estimates** (sub-sentence level) or an **overall estimate** for the complete sentence or **not seeing** bad sentences at all?

- Too much information vs hard-to-interpret scores
- IBM’s *Goodness* metric

**Source**

أنت مختلف تماماً عن زيد وعمر فلا تحشر نفسك في سراديب التقليد والمحاكاة والذوبان

**MT output**

you totally different from zaid amr, and not to deprive yourself in a basement of imitation and assimilation.

**We predict and visualize**

you **totally different from** zaid amr, and **not to deprive yourself in a basement of imitation and** assimilation.
How to use estimated PE effort scores?

Do users prefer detailed estimates (sub-sentence level) or an overall estimate for the complete sentence or not seeing bad sentences at all?

- Too much information vs hard-to-interpret scores
- IBM’s Goodness metric

MATECAT project investigating it
Feature engineering

Two families of features missing in current work:

- Can we benefit from contextual, document-wide information?
- Can we predict human translation quality?
Feature engineering

Two families of features missing in current work:

- Can we benefit from contextual, **document-wide information**?
- Can we predict **human translation quality**?

**Don’t panic, you can help!**

Two (sub-) QuEst projects at MTM-2013 :-)
Conclusions

- (Machine) Translation evaluation is still an open problem
Conclusions

- (Machine) Translation evaluation is still an open problem
- Different purposes/users, different needs, different notions of **quality**
(Machine) Translation evaluation is still an open problem

Different purposes/users, different needs, different notions of quality

**Quality estimation**: learning of these different notions
Conclusions

- (Machine) Translation evaluation is still an open problem
- Different purposes/users, different needs, different notions of quality
- Quality estimation: learning of these different notions
- Estimates have been used in real applications
  - Ranking translations: filter out bad quality translations
  - Selecting translations from multiple MT systems
(Machine) Translation evaluation is still an open problem
Different purposes/users, different needs, different notions of quality

Quality estimation: learning of these different notions
Estimates have been used in real applications
- Ranking translations: filter out bad quality translations
- Selecting translations from multiple MT systems

Commercial interest
Conclusions

- (Machine) Translation evaluation is still an open problem
- Different purposes/users, different needs, different notions of quality
- Quality estimation: learning of these different notions
- Estimates have been used in real applications
  - Ranking translations: filter out bad quality translations
  - Selecting translations from multiple MT systems
- Commercial interest
  - SDL LW: TrustScore
  - Multilizer: MT-Qualifier
Conclusions

- (Machine) Translation evaluation is still an open problem
- Different purposes/users, different needs, different notions of **quality**
- **Quality estimation**: learning of these different notions
- Estimates have been used in **real applications**
  - Ranking translations: filter out bad quality translations
  - Selecting translations from multiple MT systems
- **Commercial** interest
  - SDL LW: TrustScore
  - Multilizer: MT-Qualifier
- Interesting **open issues**: join the QuEst projects!