Quality Estimation for Machine Translation: different users, different needs

Lucia Specia
L.Specia@wlv.ac.uk
http://pers-www.wlv.ac.uk/~in1316/

JEC Workshop
October 14th 2011
Quality Estimation for Machine Translation: different translators, same needs

Lucia Specia
L.Specia@wlv.ac.uk
http://pers-www.wlv.ac.uk/~in1316/

JEC Workshop
October 14th 2011
Why are you not (yet) using MT?

- Why do you use translation memories?
- Perfect translations?
Outline

- Quality Estimation (QE) for Machine Translation (MT)
- Applications
- General approach
- What aspect of quality we want to estimate and how to represent it
- How we assess quality estimation systems
Goal: given the output of an MT system for a given input, provide an estimate of its quality

Motivations: assessing the quality of translations is

Time consuming, tedious, not worth it

Not always possible

Une interdiction gouvernementale sur la non-UE conjoints étrangers de moins de 21 à venir au Royaume-Uni, qui a été introduit par le Labour en 2008 et vise un partenaire étranger de l'extérieur de l'UE ne pouvait pas se joindre à leurs partenaires au Royaume-Uni si elles étaient moins de 21 ans, est illégale, disent les juges haut.
Main applications:

Is it worth providing this translation to a professional translator for post-editing?

Should this translation be highlighted as “not reliable” to a reader?

Given multiple translation options for a given input, can we select the best one?

Is this sentence good enough for publishing as is?
Different from MT evaluation (BLEU, NIST, etc):

- MT system in use, translating unseen text
- Translation unit: sentence ➔ not about average quality
- Independent from MT system (post-MT)
General approach

1. Decide **which aspect of quality** to estimate
2. Decide **how to represent** this aspect of quality
3. Collect **examples** of translations with different levels of quality
4. Identify and extract **indicators** that represent this quality
5. Apply an algorithm to induce a **model** to predict quality scores for new translations
6. Evaluate this model on new translations
General approach

1. Decide **which aspect of quality** to estimate: “**post-edit effort**”

5. Apply an algorithm to induce a model to predict quality
How is quality defined?

1. **Good vs bad translations**: good for what?
   (Blatz et al., 2003)

2. **MT1 vs MT2**: is MT1 better than MT2. Yes, but is MT1 good enough?
   (Blatz et al., 2003; He et al., 2010)

3. **Perfect vs not perfect translations**: can we publish this translation as is?
   (Soricut and Echihabi, 2010)

   Define “quality” in terms of post-editing effort

4. Which translations are **good enough** for post-
How is quality defined?

What levels of quality can we expect from an MT system?

1. **Perfect**: no post-editing needed at all

2. **Good**: some post-editing needed, but faster/easier than translating from scratch

3. **Bad**: too much post-editing needed, faster/easier to translate from scratch

We expect the machine to estimate this well, but can humans do it well?
How is quality defined?

The court said that the rule was unjustified.

La cour a déclaré que la règle était injustifiée.

"I basically felt like I'd been exiled from my country and in forcing him to leave they'd also forced me to leave," she said.

"J'ai essentiellement ressenti si j'avais été exilé de mon pays et dans le forçant à quitter leur avais m'a aussi forcé de partir", dit-elle.
How is quality defined?

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle! ...
How do humans perform?

Humans are good at identifying perfect translations, as well as terribly bad translations.

But medium quality translations are more difficult: “good enough” depends on the translator.

- Very experienced translators: may prefer only close to perfect translations
- Less experienced translators: may benefit from
How do QE systems perform?

- **Humans**: agreement on **en-es Europarl**: 85% (prof., 2 an.)
- **Humans**: agreement on **en-pt subtitles** of TV series: 850 sentences (non prof., 3 an.)
  - 351 cases (41%) have **full** agreement
  - 445 cases (52%) have **partial** agreement
  - 54 cases (7%) have **null** agreement

- Agreement by score:

<table>
<thead>
<tr>
<th>Score</th>
<th>Full</th>
<th>Partial</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>59%</td>
<td>41%</td>
</tr>
<tr>
<td>3</td>
<td>35%</td>
<td>65%</td>
</tr>
<tr>
<td>2</td>
<td>23%</td>
<td>77%</td>
</tr>
<tr>
<td>1</td>
<td>50%</td>
<td>50%</td>
</tr>
</tbody>
</table>
How do QE systems perform?

Simplify the task, if we know how experienced the translator is: binary problem -> **good**

<table>
<thead>
<tr>
<th>Languages</th>
<th>MT system</th>
<th>Accuracy</th>
<th>Most frequent score</th>
<th>Sentence length</th>
</tr>
</thead>
<tbody>
<tr>
<td>en-es</td>
<td>MT1</td>
<td>70%</td>
<td>52%</td>
<td>36%</td>
</tr>
<tr>
<td>en-es</td>
<td>MT2</td>
<td>77%</td>
<td>74%</td>
<td>21%</td>
</tr>
<tr>
<td>en-es</td>
<td>MT3</td>
<td>66%</td>
<td>57%</td>
<td>30%</td>
</tr>
<tr>
<td>en-es</td>
<td>MT4</td>
<td>94%</td>
<td>94%</td>
<td>70%</td>
</tr>
</tbody>
</table>
How do QE systems perform?

- **Evaluation in terms of classification accuracy → clear**
  - Upper bound = 100%
  - 50% = we are selecting 50% of the bad cases as good / of the good cases as bad

- Is ~70% accuracy enough?

- A different perspective: **precision/recall** by category:
  - How many bad translations the system says are good (**false rate**)
  - How many good the system says are bad (**miss rate**)
How do QE systems perform?

- Selecting only good translations: [3-4] (en-es)

- Number of good translations in the top $n$ translations:

<table>
<thead>
<tr>
<th>$n$</th>
<th>Human</th>
<th>CE</th>
<th>Aborted N</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>80</td>
<td>100</td>
<td>120</td>
<td>140</td>
</tr>
<tr>
<td>200</td>
<td>160</td>
<td>180</td>
<td>200</td>
<td>220</td>
</tr>
<tr>
<td>500</td>
<td>340</td>
<td>360</td>
<td>380</td>
<td>400</td>
</tr>
</tbody>
</table>

Legend:
- Human
- CE
- Aborted N
- Length
Are 2/4 discrete scores enough?

- We want to estimate: 1, 2 or 1, 2, 3, 4
- It’s like saying you can get, from a TM:
  - Only 0% match or 100% match
  - Or the following (fuzzy) match levels: 0%, 50%, 75%, 100%
- Isn’t there anything in between?

Estimate a continuum: a real number in [1, 4]
Estimating a continuous score

- **English-Spanish** Europarl data
  - 4 SMT systems, 4 sets of 4,000 translations

- **Quality score**: 1-4

<table>
<thead>
<tr>
<th>Languages</th>
<th>MT System</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>en-es</td>
<td>MT1</td>
<td>0.653</td>
</tr>
<tr>
<td>en-es</td>
<td>MT2</td>
<td>0.718</td>
</tr>
<tr>
<td>en-es</td>
<td>MT3</td>
<td>0.706</td>
</tr>
<tr>
<td>en-es</td>
<td>MT4</td>
<td>0.603</td>
</tr>
</tbody>
</table>
Is a number in $[1,4]$ informative?

Can we see this number as a fuzzy match level?

• Not really... How much work to do on a 3.2 translation?

Try more objective ways of representing quality:

\[
\text{HTER} = \frac{\# \text{ edits}}{\# \text{ words in post-edited version}}
\]

• Edit distance (HTER): distance (in $[0,1]$) between original MT and post-edited version. What is the proportion of edits (words) will I have to perform to correct
Is a number in $[1, 4]$ informative?

- **Time**: how many seconds will it take to post-edit this sentence?

  - **Time varies** considerably from annotator to annotator.

This annotation is **cheap and easy** to obtain if translators already post-edit MT.
Other ways of representing quality

- **English-Spanish, French-English** news articles
- 1,500-2,500 translations
- Quality scores:
  - Score1 = HTER
  - Score2 = [1-4]
  - Score3 = time
- Annotation tool to collect data from translators
Other ways of representing quality

Results

- Each model trained on examples from a single translator

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>fr-en</td>
<td></td>
</tr>
<tr>
<td>Distance</td>
<td>0.16</td>
</tr>
<tr>
<td>[1-4]</td>
<td>0.66</td>
</tr>
<tr>
<td>Time</td>
<td>0.65</td>
</tr>
<tr>
<td>en-es</td>
<td></td>
</tr>
<tr>
<td>Distance</td>
<td>0.18</td>
</tr>
<tr>
<td>[1-4]</td>
<td>0.55</td>
</tr>
<tr>
<td>Time</td>
<td>1.97</td>
</tr>
</tbody>
</table>
Other ways of representing quality

- So we are almost happy:
  - We can estimate an aspect of quality that is clear and objective (time, distance)

- But do these error metrics say something about how good the QE model is? Or which model is better?
Evaluation by ranking

rank translations by their QE scores (best first)

Based on the quality of the MT system for a small development data, find the percentage of “good enough” translations, using any annotation scheme. E.g. 30% of the translation are good

easure improvement of top 30% according to QE scores:

◆ Compare average quality of full dataset
Evaluation by ranking

<table>
<thead>
<tr>
<th>Languages</th>
<th>Delta [1-4]</th>
<th>Delta Distance [0, 1]</th>
<th>Delta Time (sec/word)</th>
</tr>
</thead>
<tbody>
<tr>
<td>fr-en (70% good)</td>
<td>0.07</td>
<td>-0.02</td>
<td>-0.11</td>
</tr>
<tr>
<td>en-es (40% good)</td>
<td>0.20</td>
<td>-0.06</td>
<td>-0.19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Languages</th>
<th>Delta [1-4]</th>
<th>Delta Distance [0, 1]</th>
<th>Delta Time (sec/word)</th>
</tr>
</thead>
<tbody>
<tr>
<td>fr-en</td>
<td>0.16</td>
<td>-0.04</td>
<td>-0.20</td>
</tr>
<tr>
<td>en-es</td>
<td>0.15</td>
<td>-0.04</td>
<td>-0.26</td>
</tr>
</tbody>
</table>

25%, 50% and 75%
Extrinsic evaluation by ranking

- Measure **post-editing time** to correct top 30% translations selected according to QE scores.
  - Compare it against post-editing time of randomly selected 30% translations.

  - If can't decide on the %, measure **number of words that can be post-edited in a fixed amount of time** from best to worse translations ranked according to QE model.
  - Compare it against number of words post-
Extrinsic evaluation by ranking

Evaluation:

- Model 1 (HTER)
- Model 2 (1-4 scores)
- Model 3 (Time)

2.4K new translations

600 translations
600 translations
600 translations
600 translations

Sorted 600 translations
Sorted 600 translations
Sorted 600 translations

# words?  # words?  # words?  # words?
# Extrinsic evaluation by ranking

- **Post-editing in 1 hour:**

<table>
<thead>
<tr>
<th>MT System / Dataset</th>
<th>Words/second</th>
</tr>
</thead>
<tbody>
<tr>
<td>S6 fr-en HTER (0-1)</td>
<td>0.96</td>
</tr>
<tr>
<td>S6 fr-en [1-4]</td>
<td>0.91</td>
</tr>
<tr>
<td>S6 fr-en time (sec/word)</td>
<td>1.09</td>
</tr>
<tr>
<td>S7 en-es HTER (0-1)</td>
<td>0.41</td>
</tr>
<tr>
<td>S7 en-es [1-4]</td>
<td>0.43</td>
</tr>
<tr>
<td>S7 en-es time (sec/word)</td>
<td>0.57</td>
</tr>
<tr>
<td>S7 en-es no CE</td>
<td>0.32</td>
</tr>
</tbody>
</table>
Extrinsic evaluation by ranking

Summing up:

- The aspect of quality we estimate is clear (time, distance)
- The number of extrinsic ranking-based (esp. extrinsic) evaluations says something about how good a model is
How about other users?

Post-editing time/distance/[1-4] scores have a good (pearson) correlation:

- **Distance** and [1-4] = 0.75 - 0.82
- **Time** and [1-4] = 0.50 - 0.60
  (the smaller values are when scores are given by different translators)

If we correlate post-editing time/distance and [1-4] scores reflecting adequacy (not post-editing effort)

- **Distance** and [1-4] Adequacy = 0.55
- **Time** and [1-4] Adequacy = 0.40
Is this enough?

- Is an accurate QE system at the sentence level enough?

- QE should also indicate, for sentences that are not perfect, what the bad parts are.

  Sub-sentence level QE (error detection in translations) (Xiong et al., 2010): link grammar: mostly words
Conclusions

- It is possible to estimate the quality of MT systems with respect to post-editing needs.

- Measuring and estimating post-editing time seems to be the best way to build and evaluate QE systems.

  - **Translator-dependent** measure: build a model per translator or project the time differences.

  - **Extrinsic** evaluation using time is expensive, not feasible to compare many QE systems.

- Alternative: intrinsic ranking-based measures based on a pre-annotated test set: $\Delta X$, $X = \{1-4\}$, Time, BLEU, etc.
Conclusions

- QE is a relatively new area

- It has a great potential to make MT more useful to end-users:
  - Translation: minimize post-editing time, allow for fair pricing models
  - Localization: keep the “brand” of the product/company
  - Gisting: avoid misunderstandings
  - Dissemination of large amounts of content, e.g.: user reviews
Advertisement:

- **Shared task on QE**
  - Most likely with WMT at NAACL, June 2012
  - **Sentence-level**: classification, regression and ranking

- We will provide:
  - Training sets annotated for quality
  - Baseline feature sets
  - Baseline systems to extract features
  - Test sets annotated for quality
Questions?

Lucia Specia
l.specia@lv.ac.uk
Regression + Confidence Machines to define the splitting point according to expected confidence level.
QE score x MT metrics: Pearson’s correlation across datasets produced by different MT systems:

<table>
<thead>
<tr>
<th>Test set</th>
<th>Training set</th>
<th>Pearson QE and human</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3 en-es</td>
<td>S1 en-es</td>
<td>0.478</td>
</tr>
<tr>
<td></td>
<td>S2 en-es</td>
<td>0.517</td>
</tr>
<tr>
<td></td>
<td>S3 en-es</td>
<td><strong>0.542</strong></td>
</tr>
<tr>
<td></td>
<td>S4 en-es</td>
<td>0.423</td>
</tr>
<tr>
<td>S2 en-es</td>
<td>S1 en-es</td>
<td>0.531</td>
</tr>
<tr>
<td></td>
<td><strong>S2 en-es</strong></td>
<td><strong>0.562</strong></td>
</tr>
<tr>
<td></td>
<td>S3 en-es</td>
<td>0.547</td>
</tr>
<tr>
<td></td>
<td>S4 en-es</td>
<td>0.442</td>
</tr>
</tbody>
</table>

(Journal of MT 2010)
Features

- Adequacy indicators
- Translation
- Fluency indicators
- Source text
- MT system
- Complexity indicators
- Confidence indicators

- Shallow vs linguistically motivated
- MT system dependent vs independent
Source features

- Source sentence length
- Language model of source
- Average number of possible translations per source word
- % of n-grams belonging to different frequency quartiles of the source side of the parallel corpus
- Average source word length
- ...
- ...
Target features

- Target sentence length
- Language model of target
- Proportion of untranslated words
- Grammar checking
- Mismatching opening/closing brackets, quotation symbols
- Coherence of the target sentence
- ...


MT features (confidence)

- SMT model global score and internal features
  - Distortion count, phrase probability, ...
- % search nodes aborted, pruned, recombined ...
- Language model using n-best list as corpus
- Distance to centre hypothesis in the n-best list
- Relative frequency of the words in the translation in the n-best list
- Ratio of SMT model score of the top translation to the sum of the scores of all hypotheses in the n-best list
- ...

Source-target features

- Ratio between source and target sentence lengths
- Punctuation checking (target vs source)
- Correct translation of pronouns
- Matching of phrase/POS tags
- Matching of dependency relations
- Matching of named entities
- Matching of semantic role labels
- Alignment of these and other linguistic markers
- ...

Source-target features
MT system selection

Approach:
- **Train** QE models for each MT system (individually)
- Use all MT systems to **translate** each input segment
- **Estimate** the QE score for each alternative translation
- **Select** the translation with the highest CE score

Experiments:
- **En-Es Europarl [1-4] datasets**, 4 MT systems

Results:
How do QE systems perform?

- Selecting only good translations: [3-4] (en-es)

Average human scores in the top N translations:

![Average scores x TOP N graph](image)