Quality Estimation for Machine Translation

Lucia Specia
L.Specia@sheffield.ac.uk
http://dce.shef.ac.uk/~lucia

GALA Conference
March 27th 2012
Quality Estimation - what is it about

- **Quality assessment is important:**
  - Human & machine translation (MT)
  - MT: system development & system in use

- **Goal:** given translations (segments) produced by an MT system, estimate its quality **PRIOR** to any post-editing/human analysis
  - This estimate should be **meaningful** for the task at hand
Quality Estimation - what it is not about

Different from:

- MT evaluation metrics – not a similarity metric: how close the MT is to good translation(s) (string matching)

  Although the northern wind shrieked across the sky, but was still very clear.

  Although the northern wind shrieked across the sky, but was still very clear.

  However, the sky remained clear under the strong north wind.

  Although a north wind was howling, the sky remained clear and blue.

  The sky was still crystal clear, though the north wind was howling.

- TM fuzzy match level – not based on source text only

- (Linguistic) error analysis – at least not yet
Quality Estimation for MT

- Some applications:
  - Is it worth providing this translation to a professional translator for revision?
  - 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?
General approach

1. Decide **which aspect(s) of quality** to estimate and **how to represent** it

2. Collect **examples** of translations with different levels of quality

3. Identify and extract **indicators** that reflect quality

4. Apply an algorithm to induce a **model** to predict quality scores for new translations
Quality indicators

Source text

MT system

Translation

Complexity indicators

Confidence indicators

Fluency indicators
How is quality defined?

In terms of post-editing needs

- How much **effort** it will be necessary to edit a segment
- How much **time** it will take to edit a segment
- How **far** the machine translation is from a correct translation
  - What percentage of words will need to be corrected
## Variants of estimates

- **Post-editing effort**: scores in [1-4]

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Requires complete retranslation</td>
</tr>
<tr>
<td>2</td>
<td>A lot of post-editing needed</td>
</tr>
<tr>
<td>3</td>
<td>A little post-editing needed</td>
</tr>
<tr>
<td>4</td>
<td>Fit for purpose</td>
</tr>
</tbody>
</table>
Variants of estimates

- **Edit distance (HTER)**: proportion of edits (words) that need to be performed to correct a translation
  - Deletion, insertion, substitution and shift
Variants of estimates

- **Time**: seconds to correct a translation
  - **Time varies** considerably from annotator to annotator

This annotation is **cheap** and **easy** to obtain if translators already use MT.
Extrinsic evaluation by ranking

- Post-editing in 1 hour:

<table>
<thead>
<tr>
<th></th>
<th>fr-en</th>
<th>en-es</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTER (0-1)</td>
<td>0.96</td>
<td>0.41</td>
</tr>
<tr>
<td>Effort [1-4]</td>
<td>0.91</td>
<td>0.43</td>
</tr>
<tr>
<td>time (sec/word)</td>
<td>1.09</td>
<td>0.57</td>
</tr>
<tr>
<td>no QE</td>
<td>0.75</td>
<td>0.32</td>
</tr>
</tbody>
</table>
Conclusions

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

- QE has the potential to make MT more useful to end-users.
Where to

- Better understanding of what dimensions of quality are relevant for different tasks
  - Estimate those dimensions
- Directly estimate the cost to post-edit a segment
- Move towards error detection (and automated post-editing) ???
  - “What” needs to be corrected
  - Depending on the dimensions chosen, QE and error detection will be more or less similar
Quality Estimation for Machine Translation

Lucia Specia
L.Specia@sheffield.ac.uk
http://dce.shef.ac.uk/~lucia

GALA Conference
March 27th 2012
General approach

1. Decide **which aspect(s) of quality** to estimate
2. Decide **how to represent** this aspect of quality
3. Collect **examples** of translations with different levels of quality
Extrinsic evaluation by ranking

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

2.4 K new translations

600 translations

Model 1 (HTER)

Sorted 600 translations

# words?

# words?

# words?

# words?
Evaluating machine translation

- **TER**: Translation Error Rate
- Edit distance between hypotheses and references
- Considers ‘shifts’ to measure “the amount of work needed to correct the translations”

**REF:** SAUDI ARABIA denied this week information published in the AMERICAN new york times

**HYP:** [this week] the saudis denied information published in the ***** new york times

1 Shift, 2 Substitutions, 1 Deletion
4 Edits (TER = 4/13 = 31%)