Towards Fairer Evaluations of Commercial MT Systems on BTEC

Hervé Blanchon
GETA-CLIPS
herve.blanchon@imag.fr
Outline

- Why did we embark?
- Which Systems did we embark with?
- Chinese-English unlimited data track
- Japanese-English unlimited data track
- Competitive evaluation
- Some Q&A
- Final personal comments
Why did we embark?

- View actual data and associated evaluation results, to:
  - Follow the state of the art techniques
  - Clarify our ideas about some problems either of generally proposed evaluation settings or of the communication around the results
Which systems, pairs and tracks?

- Widespread and available for the language pairs
  - Systran Web & Systran Prof. Premium V5
- Rule-based approach
- Unlimited data track
- Systran Web already been used a baseline system

<table>
<thead>
<tr>
<th>Runs</th>
<th>C-E</th>
<th>J-E</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_1</td>
<td>J_1</td>
<td>Systran Web V5</td>
</tr>
<tr>
<td>C_2</td>
<td>J_2</td>
<td>Systran PP v5 with original dictionaries</td>
</tr>
<tr>
<td>C_3</td>
<td>J_3</td>
<td>Systran PP v5 with original and user dictionaries</td>
</tr>
</tbody>
</table>
Subjective evaluation (\textbf{C_3})

- Non-native English > Fluency > Disfluent English
- Adequacy ≈ much of the meaning is expressed

Objective evaluation

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>GMT</th>
<th>NIST</th>
<th>PER</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_3</td>
<td>0.1620</td>
<td>1</td>
<td>6.0061</td>
<td>0.5429</td>
<td>0.6581</td>
</tr>
<tr>
<td>C_1</td>
<td>0.1600</td>
<td>3</td>
<td>5.9143</td>
<td>0.5423</td>
<td>0.6474</td>
</tr>
<tr>
<td>C_2</td>
<td>0.1620</td>
<td>1</td>
<td>6.0039</td>
<td>0.5429</td>
<td>0.6581</td>
</tr>
</tbody>
</table>

Same results for each version!
## Japanese-English (results)

### Subjective evaluation (J_3)
- non-native English > **Fluency** > disfluent English
- much > **Adequacy** > little

### Objective evaluation

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>GMT</th>
<th>NIST</th>
<th>PER</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>J_3</td>
<td>0.1320</td>
<td>0.5687</td>
<td>5.6476</td>
<td>0.5978</td>
<td>0.7304</td>
</tr>
<tr>
<td>J_2</td>
<td>0.1311</td>
<td>0.5672</td>
<td>5.6096</td>
<td>0.6012</td>
<td>0.7349</td>
</tr>
<tr>
<td>J_1</td>
<td>0.0810</td>
<td>0.5116</td>
<td>4.1935</td>
<td>0.7179</td>
<td>0.8726</td>
</tr>
</tbody>
</table>

*Systems are ordered according to expectation*
Bad translation when subject is omitted
- ここで 降ります。It gets off here.

Euphemistic utterance が translated by "but"
- 両替 を したい の です が。It is to like to exchange but.

Question word order
- 入場 料 は いくら です か。Is admission fee how much?

Requests or invitations
- 一緒 に 行き ましょう。It will go together.

...
## Competitive evaluation

### Observed results for Japanese-English

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>GMT</th>
<th>NIST</th>
<th>PER</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>JE_1</td>
<td>0.6306</td>
<td>0.7967</td>
<td>10.7201</td>
<td>0.2333</td>
<td>0.2631</td>
</tr>
<tr>
<td>JE_3</td>
<td>0.6190</td>
<td>0.8243</td>
<td>11.2541</td>
<td>0.2492</td>
<td>0.3056</td>
</tr>
<tr>
<td>JE_4</td>
<td>0.3970</td>
<td>0.6722</td>
<td>7.8893</td>
<td>0.4202</td>
<td>0.4857</td>
</tr>
<tr>
<td>J_3</td>
<td>0.1320</td>
<td>0.5687</td>
<td>5.6476</td>
<td>0.5978</td>
<td>0.7304</td>
</tr>
</tbody>
</table>

- From rough scores Systran is 4th
- What does it mean knowing that subjective evaluation is bad?
- Is this ranking relevant?
Competitive evaluation

- Observed results for Human Japanese-English

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>GMT</th>
<th>NIST</th>
<th>PER</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>JE_1</strong></td>
<td>0.6306</td>
<td>0.7967</td>
<td>10.7201</td>
<td>0.2333</td>
<td>0.2631</td>
</tr>
<tr>
<td><strong>JE_3</strong></td>
<td>0.6190</td>
<td>0.8243</td>
<td>11.2541</td>
<td>0.2492</td>
<td>0.3056</td>
</tr>
<tr>
<td><strong>J_4</strong></td>
<td>0.4691</td>
<td>0.7777</td>
<td>9.9189</td>
<td>0.3236</td>
<td>0.3711</td>
</tr>
<tr>
<td><strong>JE_4</strong></td>
<td>0.3970</td>
<td>0.6722</td>
<td>7.8893</td>
<td>0.4202</td>
<td>0.4857</td>
</tr>
<tr>
<td><strong>J_3</strong></td>
<td>0.1320</td>
<td>0.5687</td>
<td>5.6476</td>
<td>0.5978</td>
<td>0.7304</td>
</tr>
</tbody>
</table>

- Perfect human minimal post-edition does not over-score MT
- What does it mean knowing that subjective evaluation should be good?
Q&A (actual questions from IR people)

Q: Is-it NORMAL?
A: YES of course! We are NOT evaluating translation QUALITY but SIMILARITY between candidate translations and references translations!

Q: BUT, references are produced by humans!!!
A: Yes, But … the post-editor may have produced translations having different style or wording compared with the references!
Q: Objective evaluation is said to be good because it results correlates with subjective evaluation?
R: Correlation is still a hot topic! Sometimes the correlation is good, sometimes it is not the case.

Q: Getting better, only mean that the system produces translations that resemble better to the references? Is that why post-edition is not ranked first?
R: Yes.
Concluding personal comments

- Systran, as it is, cannot be used as a baseline system for comparative, competitive evaluation, at least for English to Japanese on the BTEC corpus
- Other language pairs have to be examined
Concluding personal comments

Objective evaluation techniques do not evaluate translation quality, they evaluate the capacity of the system to mimic the reference. Then:
- good scores mean good mimicking
- bad scores mean nothing on their own

These techniques may be well suited for “systems that learn” from the data but not for others and the comparison is meaningless.