This translation is not too bad:
An analysis of post-editor choices in a machine translation post-editing task

Maarit Koponen
University of Helsinki
maarit.koponen@helsinki.fi

2nd Workshop on Post-Editing Technology and Practice
MT Summit, Nice, September 2, 2013
Background and the purpose of this study

Post-editing is an inherently subjective task – meaning can be expressed in more than one way.

Questions explored in this study:

1. How often do post-editors agree on the best MT suggestion?
2. How often do they produce the same PE version?
3. How do individual editors differ from each other?
4. What kinds of differences are observed in the PE versions?
Introduction

Background and the purpose of this study

Post-editing is an inherently subjective task – meaning can be expressed in more than one way.

Questions explored in this study:

1. How often do post-editors agree on the best MT suggestion?
2. How often do they produce the same PE version?
3. How do individual editors differ from each other?
4. What kinds of differences are observed in the PE versions?
The post-editing experiment

Material

Pilot evaluation material from EU MOLTO evaluation of a rule-based, multi-lingual controlled language generation and MT tool.

- 139 English source segments (2–15 words per segment, total 827)
- 3 Finnish MT versions (MOLTO GF, Google, Bing)
- Post-editing data from 11 post-editors

Post-editing was carried out by 11 native Finnish speakers using the Appraise tool (Federmann 2012).
The post-editing experiment

Material

Pilot evaluation material from EU MOLTO evaluation of a rule-based, multi-lingual controlled language generation and MT tool.

- 139 English source segments (2–15 words per segment, total 827)
- 3 Finnish MT versions (MOLTO GF, Google, Bing)
- Post-editing data from 11 post-editors

Post-editing was carried out by 11 native Finnish speakers using the Appraise tool (Federmann 2012).

Instructions for the post-editing task:

1. Select the MT suggestion you consider the best.
2. Edit the suggestion as necessary, or accept as-is.
3. If none of the suggestions are acceptable, translate from scratch.
Post-editor agreement of best MT and acceptability

- The same MT suggestion is selected by all. All accept without editing.
- The same MT suggestion is selected by all. None accept without editing.
- The same MT suggestion is selected by all. Some accept without editing.
Post-editor agreement of best MT and acceptability

- The same MT suggestion is selected by all. All accept without editing.
- The same MT suggestion is selected by all. None accept without editing.
- The same MT suggestion is selected by all. Some accept without editing.
- Different MT suggestions are selected. All accept without editing.
- Different MT suggestions are selected. None accept without editing.
- Different MT suggestions are selected. Some accept without editing.
## Results: Selection and acceptance

<table>
<thead>
<tr>
<th></th>
<th>MT suggestions selected</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Same</td>
<td>Different</td>
<td>Total</td>
</tr>
<tr>
<td>All accept</td>
<td>44</td>
<td>15</td>
<td>59</td>
</tr>
<tr>
<td>None accept</td>
<td>1</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Some accept</td>
<td>33</td>
<td>41</td>
<td>74</td>
</tr>
<tr>
<td>Total</td>
<td>78</td>
<td>61</td>
<td>139</td>
</tr>
</tbody>
</table>

M. Koponen (University of Helsinki)
This translation is not too bad
WPTP-2, Sept 2, 2013
## Results: Different PE versions per segment

<table>
<thead>
<tr>
<th></th>
<th>Number of PE versions</th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>≥ 4</td>
</tr>
<tr>
<td><strong>Same MT</strong></td>
<td>All accept</td>
<td>44</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>None accept</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Some accept</td>
<td>0</td>
<td>21</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td><strong>Different MT</strong></td>
<td>All accept</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>None accept</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Some accept</td>
<td>3</td>
<td>11</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>48</td>
<td>47</td>
<td>23</td>
<td>21</td>
</tr>
</tbody>
</table>
Versions vs segment length
Comparison of individual editors

Individual editors were compared in terms of

- Number of segments edited
- HTER: number of edited tokens divided by number of PE tokens – calculated using TERplus (Snover et al. 2009)
- Post-editing time recorded by Appraise (Federmann 2012)
- Number of MT selections differing from the most common choice
- Number of PE versions differing from the most common choice
Results: Number of sentences edited by each editor

<table>
<thead>
<tr>
<th>Editor</th>
<th>Number of sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>FI01</td>
<td>40</td>
</tr>
<tr>
<td>FI02</td>
<td>30</td>
</tr>
<tr>
<td>FI04</td>
<td>20</td>
</tr>
<tr>
<td>FI05</td>
<td>30</td>
</tr>
<tr>
<td>FI06</td>
<td>30</td>
</tr>
<tr>
<td>FI08</td>
<td>40</td>
</tr>
<tr>
<td>FI09</td>
<td>20</td>
</tr>
<tr>
<td>FI10</td>
<td>30</td>
</tr>
<tr>
<td>FI11</td>
<td>10</td>
</tr>
<tr>
<td>FI13</td>
<td>20</td>
</tr>
<tr>
<td>FI14</td>
<td>40</td>
</tr>
</tbody>
</table>
Results: HTER scores for each editor
Results: Editing time for each editor

This translation is not too bad
Results: Number of MT selections differing from the most common
Results: Number of PE versions differing from the most common

<table>
<thead>
<tr>
<th>Editor</th>
<th>Number of sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>FI01</td>
<td>0</td>
</tr>
<tr>
<td>FI02</td>
<td>5</td>
</tr>
<tr>
<td>FI04</td>
<td>10</td>
</tr>
<tr>
<td>FI05</td>
<td>15</td>
</tr>
<tr>
<td>FI06</td>
<td>20</td>
</tr>
<tr>
<td>FI08</td>
<td>25</td>
</tr>
<tr>
<td>FI09</td>
<td>30</td>
</tr>
<tr>
<td>FI10</td>
<td>35</td>
</tr>
<tr>
<td>FI11</td>
<td></td>
</tr>
<tr>
<td>FI13</td>
<td></td>
</tr>
<tr>
<td>FI14</td>
<td></td>
</tr>
</tbody>
</table>

M. Koponen (University of Helsinki)
Observations about accepted/rejected suggestions and PE differences

- Most rejected suggestions contain multiple errors affecting both meaning and language.
- Differences in selections and PE versions generally involved:
  - Synonymous words and expressions
  - Word order
  - Punctuation
  - Alternate spellings, alternate suffix forms
  - Editor errors
- Some particularly strong preferences were observed:
  - Correct MT edited to include a preferred expression
  - Incorrect MT with a preferred expression selected over correct MT
- Some sentences attracted a particularly large number of PE versions
  - Often idiomatic expressions
Example: Editor preferences

ST  Does his daughter want to go to the most popular museum?

MT1  Tahtooko hänen tyttärensä mennä suositumpaan museoon?
     ’Does his daughter want to go to the most popular museum?’

MT2  Haluaako tyttärensä Siirry suosituimmista museo?
     ’Does daughter (of?) want Move (sg2 imperative) from most popular (plural) museum?’

MT3  Ei hänen tyttärensä haluaa lähteä suosituin museo?
     ’No his daughter want to leave most popular museum?’

This translation is not too bad
Example: Editor preferences

ST  Does his daughter want to go to the most popular museum?

MT1  Tahtooko hänä tyttärenä mennä suositumpaan museoon?
  ’Does his daughter want to go to the most popular museum?’

MT2  Haluaako tyttärenä Siirry suosituimmista museo?
  ’Does daughter (of?) want Move (sg2 imperative) from most popular
  (plural) museum?’

MT3  Ei hänä tyttärenä haluaa lähteä suosituin museo?
  ’No his daughter want to leave most popular museum?’

8 out of 11 select MT1 as the best suggestion – 6 accept as-is.
3 out of 11 prefer to edit MT2, although it contains multiple errors.
Example: Differing PE versions

ST  This apple is not too bad.
FI01  Tämä omena ei ole hullumpi.
FI02  Tämä omena ei ole hassumpi.
FI04  Tämä omena ei ole hullumpi.
FI05  Tämä omena ei ole niin pahaa.
FI06  Tämä omena ei ole niin huono.
FI08  Tämä omena ei ole hullumpaa.
FI09  Tämä omena ei ole kummempaa.
FI10  Tämä omena ei ole ihan huono.
FI11  Tämä omena ei ole liian pahaa.
FI13  Tämä omena ei ole kovin huono.
FI14  Tämä omena on ihan hyvä.
Conclusions

- For most source sentences, all editors or all but one select the same MT suggestion.

- For most source sentences, there is one or two PE versions.
  - This result is likely affected by the nature of the controlled language sentences and good MT quality (large number of segments accepted without editing).
Conclusions

- For most source sentences, all editors or all but one select the same MT suggestion.
- For most source sentences, there is one or two PE versions.
  - This result is likely affected by the nature of the controlled language sentences and good MT quality (large number of segments accepted without editing).
- Different editors show differing editing patterns and preferences.
- Preferences for some words or expressions appeared particularly strong.
Thank you! Kiitos!

maarit.koponen@helsinki.fi