Dissecting Human Pre-Editing toward Better Use of Off-the-Shelf Machine Translation Systems

Rei Miyata (Nagoya University; miyata@nuee.nagoya-u.ac.jp)  Atsushi Fujita (NICT)

1. Objective

1) Investigate the capability of the pre-editing strategy
   • Design a human-in-the-loop protocol to collect pre-edit instances
   • Collect pre-edit instances in Japanese-to-English translation tasks on 4 datasets
2) Provide an overview of possible edit operations
   • Create a typology of edit operations

2. Protocol for Collecting Pre-Edit Instances

Human editors incrementally edit source texts (STs) relying on their introspection, so that improved MT quality is achieved. (Miyata et al. 2015)

New features:
• Record ST after every minimal edit is performed
• Allow editors to resume editing from any given past versions of ST

3. Pilot Run

• 4 datasets: hospital conversation (hosp), municipal information (muni), 2 news articles (bccwj: Japanese-origin & reuters: English-origin)
• MT system: TexTra (freely-available, state-of-the-art phrase-based SMT)
• A Japanese editor with ample experience in evaluating MT quality

Statistical results of the collected data: More than 85% of STs achieved satisfactory quality

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Mode</th>
<th>Avg. num. of tokens in original ST (s.d.)</th>
<th>Num. of versions of ST</th>
<th>Num. of units</th>
</tr>
</thead>
<tbody>
<tr>
<td>hosp</td>
<td>spoken</td>
<td>12.1 (4.5)</td>
<td>1199</td>
<td>12.0</td>
</tr>
<tr>
<td>muni</td>
<td>written</td>
<td>21.3 (10.0)</td>
<td>2119</td>
<td>21.2</td>
</tr>
<tr>
<td>bccwj</td>
<td>written</td>
<td>26.9 (16.0)</td>
<td>3823</td>
<td>38.2</td>
</tr>
<tr>
<td>reuters</td>
<td>written</td>
<td>34.8 (12.6)</td>
<td>5546</td>
<td>55.5</td>
</tr>
</tbody>
</table>

3.1 Example of Best ST with satisfactory MT quality

<table>
<thead>
<tr>
<th>ST</th>
<th>Original</th>
<th>MT output</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST</td>
<td>Original</td>
<td>MT output</td>
</tr>
<tr>
<td>hosp</td>
<td>2.73</td>
<td>2.93 **</td>
</tr>
<tr>
<td>muni</td>
<td>2.84</td>
<td>2.89</td>
</tr>
<tr>
<td>bccwj</td>
<td>2.39</td>
<td>2.75 **</td>
</tr>
<tr>
<td>reuters</td>
<td>2.61</td>
<td>2.77</td>
</tr>
</tbody>
</table>

4. MTranslatability for Different Languages

Human evaluation of MT quality (Chinese & Korean)

<table>
<thead>
<tr>
<th>ST: Original vs. Best</th>
<th>Org</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>hosp</td>
<td>2.73</td>
<td>2.93 **</td>
</tr>
<tr>
<td>muni</td>
<td>2.84</td>
<td>2.89</td>
</tr>
<tr>
<td>bccwj</td>
<td>2.39</td>
<td>2.75 **</td>
</tr>
<tr>
<td>reuters</td>
<td>2.61</td>
<td>2.77</td>
</tr>
</tbody>
</table>

5. Typology of Edit Operations

• Best path of 10 randomly sampled units
  = 979 instances of primitive edit operations
• 53 types of edit operations in 7 major categories:
  S (Structure), O (Content word), F (Functional word),
  T (Terminology), O (Orthography), I (Information),
  E (Edit that causes/resolves error in ST)

5.1 Frequent types

C01: Alternative lexical choice
  一度 → 一回 (once)
  習得する → 学ぶ (learn)

S05: Phrase reordering
  S07: Insertion/deletion of punctuation

Domain specific types (e.g. news domain)
  S15: Use/disuse of clause-ending noun
  S20: Use/disuse of nominal/verbal suffix

6. Conclusion and Future Work

1) Collected 12,287 pre-edit instances
   • High capability of the off-the-shelf MT system
   • Improved Chinese and Korean translatability
2) Built a typology of a wide range of edit operations

Develop an automatic pre-editor
• Controlled language formulation by assessing the effectiveness of each type of edit operation
• Statistical model based on our collection of pre-edit instances