Manipulating Input Data for Machine Translation

Liling Tan and Francis Bond
Universität des Saarlandes and Nanyang Technological University
alvations@gmail.com and bond@ieee.org

Introduction
- Data quality/quantity affects MT easily
- What happens if we try to remove, add or rearrange input data?

Approaches
- Context sensitive data selection
- Paraphrase data extension
- Lexicon addition
- Character level MT (JP-ZH)

Context Sensitive Data Selection
- Medical domain made up the majority of the ASPEC corpus (~30%)
- Build separate model for medical text
- Use generic model to decode the rest

Paraphrase Data Extension
- Generate EN paraphrases with ERG and ACE (Flickinger, 2000)
- Append EN paraphrases with original JP

EN Input: The particle sizes of the products decreased as the amount of seed increased.

Paraphrase: As the amount of seed increased, the particle sizes of the products decreased.

JP Input: 種の量が増加する程、生成物の粒子サイズ法は減少した。

Lexicon Addition
- Added JP-EN translation dictionary (JICT, 2004)

System Setup
- Moses phrase-based SMT (Koehn et al. 2007)
- GIZA++ IBM4 (Och and Ney, 2003)
- Bi-directional lex reordering (Koehn et al. 2005)
- KenLM 5grams (Heafield, 2011)
- Kneser-Ney smoothing (Kneser and Ney, 1995)
- MERT, truecasing

Character based MT
- Phrase-based SMT depends on tokens
- Diff segmenters produce diff tokens
- What if we use characters for MT? (Nakov and Tiedemann, 2012)

Input: これらカテゴリーに含まれる要素数を検討した
MeCab: これらカテゴリーに含まれる要素数を検討した
Juman: これらカテゴリーに含まれる要素数を検討した
KyTea: これらカテゴリーに含まれる要素数を検討した
Char: これらカテゴリーに含まれる要素数を検討した

Results (JP-EN)

JP-EN
- Paraphrases and lexicon have minor improvement from baseline
- Human evaluation shows implementation or config. errors

JP-ZH
- Character based MT achieves pretty good results for JP-ZH

Table 1: EN-JP Results

<table>
<thead>
<tr>
<th>BLEU</th>
<th>RIBES</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAIST</td>
<td>35.8</td>
<td>0.81</td>
</tr>
<tr>
<td>WEBLIO-EN</td>
<td>33.4</td>
<td>0.795</td>
</tr>
<tr>
<td>Organizer</td>
<td>32.1</td>
<td>0.760</td>
</tr>
<tr>
<td>Kyoto-U</td>
<td>31.7</td>
<td>0.771</td>
</tr>
<tr>
<td>SAS_MT</td>
<td>31.4</td>
<td>0.771</td>
</tr>
</tbody>
</table>

Paraphrase | 28.7 | 0.703 | - |
Baseline | 28.6 | 0.703 | 3.75 |
Lexicon | 28.1 | 0.693 | - |
Context | 27.1 | 0.697 | - |

Table 2: JP-EN Results

<table>
<thead>
<tr>
<th>BLEU</th>
<th>RIBES</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAIST</td>
<td>35.8</td>
<td>0.845</td>
</tr>
<tr>
<td>SAS_MT</td>
<td>37.0</td>
<td>0.833</td>
</tr>
<tr>
<td>ORGANIZER</td>
<td>36.6</td>
<td>0.825</td>
</tr>
<tr>
<td>Kyoto-U</td>
<td>34.8</td>
<td>0.802</td>
</tr>
<tr>
<td>Character</td>
<td>34.6</td>
<td>0.784</td>
</tr>
<tr>
<td>Baseline</td>
<td>33.4</td>
<td>0.771</td>
</tr>
<tr>
<td>EIWA</td>
<td>33.8</td>
<td>0.808</td>
</tr>
</tbody>
</table>

Table 3: ZH-JP Results

<table>
<thead>
<tr>
<th>BLEU</th>
<th>RIBES</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAIST</td>
<td>40.15</td>
<td>0.845</td>
</tr>
<tr>
<td>SAS_MT</td>
<td>37.0</td>
<td>0.833</td>
</tr>
<tr>
<td>ORGANIZER</td>
<td>36.6</td>
<td>0.825</td>
</tr>
<tr>
<td>Kyoto-U</td>
<td>34.8</td>
<td>0.802</td>
</tr>
<tr>
<td>Character</td>
<td>34.6</td>
<td>0.784</td>
</tr>
<tr>
<td>Baseline</td>
<td>33.4</td>
<td>0.771</td>
</tr>
<tr>
<td>EIWA</td>
<td>33.8</td>
<td>0.808</td>
</tr>
</tbody>
</table>

Table 4: ZH-JP Results

Conclusion
- Data manipulation seems to affect BLEU score minimally
- More experiments necessary to improve system before conclusive results

Acknowledgements
The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union’s Seventh Framework Programme FP7/2007-2013/under REA grant agreement n° 317471.