Toshiba MT System Description for the WAT2015 Workshop

Satoshi SONOH
Satoshi KINOSHITA

Knowledge Media Laboratory, Corporate Research & Development Center, Toshiba Corporation.
Motivations

• **Rule-Based Machine Translation (RBMT)**
  – We have been developed RBMT for more than 30 years.
  – Large technical dictionaries and translation rules

• **Pre-ordering SMT and Tree/Forest to String**
  – Effective solutions for Asian language translation (WAT2014)
  – But, *pre-ordering rules and parsers are needed*.

• **Our approach:**
  – Statistical Post Editing (SPE) (same as WAT2014)
    • Verify effectiveness in all tasks
  – System combination between SPE and SMT (new in WAT2015)
Statistical Post Editing (SPE)

Translating RBMT results to post-edited results.

1) We first translate source sentences by RBMT.

2) We train SPE model by translated corpus.

Input Sentence ➔ RBMT ➔ Translated Sentence ➔ Parallel Corpus (ASPEC / JPC) ➔ SPE Model ➔ SPE Result

本発明は以下の効果を有する。
Features of SPE

- **From RBMT’s standpoint**
  - **Correct mistranslations / Translate unknown words**
    - Phrase-level correction (domain adaptation)
  - **Improve fluency**
    - Use of more fluent expressions
    - Insertion of particles
  - Recover translation failure

- **From SMT’s standpoint**
  - **Pre-ordering by RBMT**
    - Reduction of NULL alignment (subject/particle)
    - Use of syntax information (polarity/aspect)
    - Enhancement of lexicon
SPE for Patent Translation

SPE shows:
- Better scores than PB-SMT in automatic evaluation
- Improvements of understandable level (>=C in acceptability)

Human evaluation for zh-ja

Adequacy

Acceptability

© 2015 Toshiba Corporation
System Combination

• How combine systems?
  – Selection based on SMT scores and/or other features.
  – Selection based on estimated score (Adequacy? Fluency? ...)
    • Need data to learn the relationship...

• Our approach in WAT2015:
  – Merge n-best candidates and rescore them.
  – We used **RNNLM** for reranking.
RNNLM reranking and Tuning

- **Reranking on the log-linear model**
  - Adding RNNLM score to default features of Moses.
  - RNNLM trained by rnnlm toolkit (Mikolov ’12).
    - 500,000 sentences for each language
    - # of hidden layer=500, # of class=50

- **Tuning**
  - Using tuned weights without RNNLM, we ran only 1 iteration.
    (to reduce tuning time)
Experimental Results

**BLEU for ASPEC**

<table>
<thead>
<tr>
<th></th>
<th>SMT</th>
<th>SPE</th>
<th>COMB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ja-en</td>
<td>17.41</td>
<td>22.65</td>
<td>23.00</td>
</tr>
<tr>
<td>en-ja</td>
<td>31.10</td>
<td>31.82</td>
<td></td>
</tr>
<tr>
<td>ja-zh</td>
<td>28.20</td>
<td>29.48</td>
<td>29.60</td>
</tr>
<tr>
<td>zh-ja</td>
<td>36.34</td>
<td>35.76</td>
<td>37.47</td>
</tr>
</tbody>
</table>

*SMT and SPE are 1-best results.

**BLEU for Patent**

<table>
<thead>
<tr>
<th></th>
<th>SMT</th>
<th>SPE</th>
<th>COMB</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPCzh-ja</td>
<td>38.77</td>
<td>39.01</td>
<td>40.23</td>
</tr>
<tr>
<td>JPCko-ja</td>
<td>70.17</td>
<td>68.47</td>
<td>70.4</td>
</tr>
</tbody>
</table>

+1.71

+1.93

+1.22
## Experimental Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RBMT</td>
<td>No</td>
<td>15.31</td>
<td>0.677</td>
<td>14.78</td>
<td>0.685</td>
<td>19.51</td>
<td>0.767</td>
<td>15.39</td>
<td>0.767</td>
</tr>
<tr>
<td>SMT</td>
<td>No</td>
<td>17.41</td>
<td>0.620</td>
<td>25.17</td>
<td>0.642</td>
<td>28.20</td>
<td><strong>0.810</strong></td>
<td>36.34</td>
<td><strong>0.810</strong></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>17.85</td>
<td>0.619</td>
<td>25.37</td>
<td>0.643</td>
<td>28.46</td>
<td>0.809</td>
<td>36.69</td>
<td>0.809</td>
</tr>
<tr>
<td>SPE</td>
<td>No</td>
<td>22.65</td>
<td>0.717</td>
<td>31.10</td>
<td>0.767</td>
<td>29.48</td>
<td>0.809</td>
<td>35.76</td>
<td>0.809</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>22.92 <strong>0.718</strong></td>
<td>31.73 <strong>0.770</strong></td>
<td>29.49 <strong>0.809</strong></td>
<td>36.06 <strong>0.809</strong></td>
<td><strong>23.00</strong></td>
<td><strong>0.716</strong></td>
<td><strong>31.82</strong></td>
<td><strong>0.770</strong></td>
</tr>
<tr>
<td>COMB</td>
<td>Yes</td>
<td><strong>23.00</strong></td>
<td><strong>0.716</strong></td>
<td><strong>31.82</strong></td>
<td><strong>0.770</strong></td>
<td><strong>29.60</strong></td>
<td><strong>0.810</strong></td>
<td><strong>37.47</strong></td>
<td><strong>0.810</strong></td>
</tr>
</tbody>
</table>

System Combination (COMB) achieved improvements of BLEU and RIBES score than SPE.

COMB is the best system except JPCko-ja task.
Which systems did the combination selected?

**ja-en/zh-ja**: about 80% translations come from SPE.

**ja-zh and JPCzh-ja**: COMB selected SPE and SMT, equivalently. (Because RBMT couldn’t translate well, % of SMT increased.)

“same” means that COMB results were included both SMT and SPE.
Toshiba MT system of WAT2015

- We additionally applied some pre/post processing.

**Technical Term Dictionaries**
Selecting RBMT dictionaries by devset.
+ JPO patent dictionary (2.2M words for JPCzh-ja)

**English Word Correction**
Edited-distance based correction.
continous -> continuous
behavior -> behavior
resolutin -> resolution

**KATAKANA Normalization**
Normalize to highly-frequent notations for “ー”.
スクリュ -> スクリュー
サーバー -> サーバ

**Post-translation**
Translate remaining unknown words by RBMT.
アルキメデス数 -> 阿基米德数
流入マハ수 -> 流入マッハ数
Official Results

- **SPE and SMT ranked in the top 3 HUMAN in ja-en/ja-zh/JPCzh-ja.**

<table>
<thead>
<tr>
<th>System</th>
<th>ja-en</th>
<th>en-ja</th>
<th>ja-zh</th>
<th>zh-ja</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>RIBES</td>
<td>HUMAN</td>
<td>BLEU</td>
</tr>
<tr>
<td>SPE</td>
<td>22.89</td>
<td>0.719</td>
<td>25.00</td>
<td>32.06</td>
</tr>
<tr>
<td>COMB</td>
<td>23.00</td>
<td>0.716</td>
<td>21.25</td>
<td>31.82</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System</th>
<th>JPCzh-ja</th>
<th>JPCko-ja</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>RIBES</td>
</tr>
<tr>
<td>SMT</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SPE</td>
<td>41.12</td>
<td>0.822</td>
</tr>
<tr>
<td>COMB</td>
<td>41.82</td>
<td>0.821</td>
</tr>
</tbody>
</table>

- **The correlation between BLEU/RIEBES and HUMAN is not clear in our system.**

![BLEU-HUMAN](image)

![RIBES-HUMAN](image)
Crowdsourcing Evaluation

• Analysis of JPCko-ja result (COMB vs Online A)
  – In in-house evaluation, COMB is better than Online A.

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>RIBES</th>
<th>Baseline</th>
<th>HUMAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMB</td>
<td>70.51</td>
<td>0.94</td>
<td>3.00</td>
<td>-</td>
</tr>
<tr>
<td>Online A</td>
<td>55.05</td>
<td>0.91</td>
<td>38.75</td>
<td>-10.75</td>
</tr>
</tbody>
</table>

– Effected by differences in number expressions !?
  
  SRC: 시스템(100) ⇒ Online A: システム(100)
  COMB(SMT): システム100
  ⇒ Equally evaluated in-house evaluation.

– Crowd-workers should be provided an evaluation guideline by which such a difference is considered.
Summary

• Toshiba MT system achieved a combination method between SMT and SPE by RNNLM reranking.

• Our system ranked the top 3 HUMAN score in ja-en/ja-zh/JPCzh-ja.

• We will aim for practical MT system by more effective combination systems (SMT, SPE, RBMT and more...)