Machine translation for patent documents combining rule-based translation and statistical post-editing

Terumasa EHARA
Yamanashi Eiwa College
Agenda

• Motivation
• System architecture
• Translation model training
• Experiments
• Conclusion
Motivation

• Rule-based MT technique can use accumulated knowledge from long history of MT researches.

• Statistical MT technique can use large power of computer hardware and database.

• Combining the two techniques can make MT more accurate.
System architecture

Japanese document

Rule-based MT from J to E (RBMT)

Grammar Dictionaries

Translation model (TM)

TM training part

Unexamined Japanese patent applications

U.S. patent grant data

Machine translated English document

Statistical post editing (SPE)

Sentence aligned

Language model (LM)

LM training part

Post edited English document

Data selection / Translation from J to E

The 8th NTCIR Workshop
System architecture (2)

- Tools that are used in our system
  - RBMT part: commercial based MT system specialized to patent translation
  - SPE processor: Moses (2007.05.29)
  - LM training tool: Srilm (ver.1.5.5)
  - TM training tool: Giza++ (v1.0.1)
Translation model training

- Extracting only matched data to the input, we make the translation model training data from the total parallel data.

- Matching algorithm is:
  \[
  \text{sim} = \frac{2 \times \#(T \cap S)}{\#(T) + \#(S)}
  \]
Experiments

• Intrinsic evaluation of MT from J to E
  ▶ JE parallel data (NTCIR-8 PAT-MT data)
    old data: 1,798,571 sentence pairs
    new data: 1,387,713 sentence pairs
  ▶ Test data: 1,251 Japanese sentences
  ▶ Development data: 2,000 JE sentence pairs
LM and TM training data

- LM training data: all of English part of new data (1,387,713 sentences)
- TM training data: extracted from old and new data matched to the test data (152,072 sentence pairs)
## Results

<table>
<thead>
<tr>
<th></th>
<th>BLEU (modified)</th>
<th>NIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBMT output</td>
<td>0.1907</td>
<td>6.1466</td>
</tr>
<tr>
<td>SPE output</td>
<td>0.3444</td>
<td>7.7538</td>
</tr>
</tbody>
</table>
Translation example

• Test sentence #2
このような構成になる弾性糸のクランプカッター装置1において、弾性糸SYを把持して切断する動作手順を、図1A、図1B、図1Cに示してある。

• Reference translation
In the clump cutter apparatus 1 of the elastic yarn configured as described above, the operational procedure to hold and cut the elastic yarn SY is shown in FIG. 1A, FIG. 1B, and FIG. 1C.
Translation example (2)

- Keyword list
  - 構成 (configured),
  - 弾性 (elastic),
  - 糸 (yarn),
  - クランプ (clump),
  - カッター (cutter),
  - 装置 (apparatus),
  - 把持 (hold),
  - 切断 (cut),
  - 動作 (operational),
  - 手順 (procedure),
  - 図 (FIG),
  - 示し (is shown)
Translation example (3)

- Extracted TM training data (top 3)
  - In FIG. 2, a cutter 46 is attached to the cutting drum 38.
  - Firstly, with reference to FIG. 1, a whole constitution of a clamping apparatus will be explained hereinafter.
  - Next, the operation of a device constructed in this way is explained using the flowcharts shown in FIGS. 12, 13, 14, and 15.
Translation example (4)

- Output of RBMT part
  In clamp cutter device 1 of the elastic yarn which becomes such composition, the procedure of operation of grasping and cutting elastic yarn SY is shown in Drawing 1A, Drawing 1B, and Drawing 1C.

- Output of SPE part
  The clamp cutter device 1 of the elastic yarn structure, the procedure of the operation of holding and cutting the elastic yarn SY is shown in FIG. 1A, FIG. 1B and FIG. 1C.
Conclusion

• Adding statistical post-editing part to rule-based machine translation, we can improve BLEU score from 0.1907 to 0.3444.

• Mean PER (position-independent word error rate) value for the RBMT outputs compared to the SPE outputs is 0.280.
Remaining issues

• To improve the parsing accuracy in the RBMT part.

• Syntactically collapsed outputs from the RBMT part can't be recovered by the SPE part.