Statistical Machine Translation adding Pattern-based Machine Translation in Chinese-English Translation

○Jin'ichi Murakami, Masato Tokuhisa, Satoru Ikehara (Tottori University, Japan)
Phrase Based Statistical Machine Translation

Problem

a) Small Database
   Unknown word

b) N-gram
   Local language information
Proposed Method:
Two-stage machine translation

First stage: Rule-based MT
   a) Few unknown words
   b) Include grammatical information
   c) Low levels of fluency and naturalness

Second stage: Normal SMT
   a) Revise the outputs of the first stage
   b) Improve the naturalness and fluency

Chinese-English: SYSTRAN + Moses
Training

Parallel Corpus

Chinese
(你可以改改吗？)

Rule based MT

ENGLISH
(You may change ?)

training-phrase-model.perl

Phrase Table
(ENGLISH ||| English)
(You may ||| Do you ||| 0.3)

English
(Do you do alterations ?)

ngram-count-lm

N-gram
(English)
Decoding

Chinese (红绿灯是红的。)

Rule-Based MT (Chinese → ENGLISH)

ENGLISH (The traffic light is red.)

SMT (ENGLISH → English)

English (The light was red.)
不用担心那个。
我要买它你不需要把它包起来。

English: No worry about that.
I'll take it and you need not wrap it up.

SYSTARN: Does not need to worry that.
I must buy its you not to need to wrap it.

你 可 以 改 改 吗 ？
Do you do alterations？
You may change？

红绿灯是红的。
The light was red.
The traffic light is red.
Examples of phrase-tables
(ENGLISH - English BTEC-CE)

Extremely appropriate . ||| It fits very well .|||
  1 0.0037774 1 0.000165701
Extremely appropriate ||| It fits very well|||
  1 0.00394828 1 0.000167943
Extremely attractive . ||| It is very beautiful .|||
  1 0.00468009 0.5 0.000167226
Extremely attractive . ||| Very beautiful .|||
  1 0.121764 0.5 0.0529012
Extremely attractive ||| It is very beautiful|||
  1 0.00489181 0.5 0.00169488
Extremely attractive ||| Very beautiful|||
  1 0.127273 0.5 0.053617
want to go to eat meal . ||| like to have dinner .|||
  1 4.70488e-06 0.5 0.00340606
want to go to eat meal . ||| want to go to the restaurant .|||
  1 1.02487e-05 0.5 4.7193e-06
want to go to eat meal ||| like to have dinner|||
  0.333333 4.91772e-06 0.5 0.00345215
want to go to eat meal ||| want to go to the restaurant|||
  1 1.07123e-05 0.5 4.78316e-06
want to go to eat ||| like to have|||
  0.0222222 3.18012e-05 1 0.0191019
5-gram Language Model
SRILM " -ukndiscount -interpolate"

Decoder
Moses (No parameter optimize)
moses.ini
   ttable-limit 40 0
   weight-d 0.1
   weight-l 1.0
   weight-t 0.5 0.0 0.5 0.1 0.0
   weight-w -1
   distortion-limit (-1 or 6)

Standard SMT: distortion-limit -1
Proposed Methods (2-stage): distortion-limit 6
01 Chinese  
Proposed  
Moses  
SYSTRAN  

02 Chinese  
Proposed  
Moses  
SYSTRAN  

03 Chinese  
Proposed  
Moses  
SYSTRAN  

04 Chinese  
Proposed  
Moses  
SYSTRAN  

<table>
<thead>
<tr>
<th>Chinese</th>
<th>我 想 和 史密斯 先生 谈话。</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>I'd like to converse to Mr. Smith.</td>
</tr>
<tr>
<td>Moses</td>
<td>I'd like to talk to Mr. Smith，谈话。</td>
</tr>
<tr>
<td>SYSTRAN</td>
<td>I want to converse with Mr. Smith.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chinese</th>
<th>有 套餐 的 菜单 吗？</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>Do you have a menu set meal？</td>
</tr>
<tr>
<td>Moses</td>
<td>Do you have set meals menu？</td>
</tr>
<tr>
<td>SYSTRAN</td>
<td>Has the prix fixe the menu？</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chinese</th>
<th>日本 的 十 几 岁 青少年 很 喜欢 玩 电子 游戏机。</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>The game a few years old young people like an electronic games mechanical well。</td>
</tr>
<tr>
<td>Moses</td>
<td>How many Japanese ten years old 青少年 electronic 游戏机 really like fun。</td>
</tr>
<tr>
<td>SYSTRAN</td>
<td>Japan's several year old young people like playing the electronic mechanical games very much。</td>
</tr>
</tbody>
</table>
### Results of Experiments

<table>
<thead>
<tr>
<th>TASK</th>
<th>BTEC_CE</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>case+punc</td>
<td>bleu</td>
<td>meteor</td>
<td>f1</td>
<td>prec</td>
<td>recl</td>
<td>wer</td>
<td>per</td>
<td>ter</td>
<td>gtm</td>
<td>nist</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.3151</td>
<td>0.6169</td>
<td>0.6569</td>
<td>0.6465</td>
<td>0.6676</td>
<td>0.5590</td>
<td>0.4760</td>
<td>48.0710</td>
<td>0.6478</td>
<td>6.3834</td>
</tr>
<tr>
<td>Moses</td>
<td>0.3311</td>
<td>0.6109</td>
<td>0.6610</td>
<td>0.6758</td>
<td>0.6468</td>
<td>0.5377</td>
<td>0.4567</td>
<td>44.8140</td>
<td>0.6423</td>
<td>6.1511</td>
</tr>
<tr>
<td>Systran</td>
<td>0.1070</td>
<td>0.4697</td>
<td>0.5619</td>
<td>0.5671</td>
<td>0.5567</td>
<td>0.7017</td>
<td>0.6182</td>
<td>60.0070</td>
<td>0.4863</td>
<td>3.9727</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TASK</th>
<th>CT_CE</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>case+punc</td>
<td>bleu</td>
<td>meteor</td>
<td>f1</td>
<td>prec</td>
<td>recl</td>
<td>wer</td>
<td>per</td>
<td>ter</td>
<td>gtm</td>
<td>nist</td>
</tr>
<tr>
<td>Proposed.CRR</td>
<td>0.2797</td>
<td>0.5971</td>
<td>0.6306</td>
<td>0.6092</td>
<td>0.6536</td>
<td>0.6590</td>
<td>0.5099</td>
<td>61.3850</td>
<td>0.6592</td>
<td>5.5309</td>
</tr>
<tr>
<td>Moses.CRR</td>
<td>0.2706</td>
<td>0.5881</td>
<td>0.6189</td>
<td>0.5945</td>
<td>0.6453</td>
<td>0.6712</td>
<td>0.5113</td>
<td>62.4990</td>
<td>0.6533</td>
<td>5.4633</td>
</tr>
<tr>
<td>Systran.CRR</td>
<td>0.0642</td>
<td>0.3953</td>
<td>0.4928</td>
<td>0.5051</td>
<td>0.4811</td>
<td>0.8046</td>
<td>0.6823</td>
<td>74.9560</td>
<td>0.4312</td>
<td>3.2979</td>
</tr>
<tr>
<td>Proposed.ASR.1</td>
<td>0.2482</td>
<td>0.5489</td>
<td>0.5910</td>
<td>0.5773</td>
<td>0.6053</td>
<td>0.6943</td>
<td>0.5456</td>
<td>64.8360</td>
<td>0.6136</td>
<td>5.0705</td>
</tr>
<tr>
<td>Moses.ASR.1</td>
<td>0.2650</td>
<td>0.5610</td>
<td>0.6000</td>
<td>0.5876</td>
<td>0.6128</td>
<td>0.6647</td>
<td>0.5220</td>
<td>62.0140</td>
<td>0.6307</td>
<td>5.2804</td>
</tr>
<tr>
<td>Systran.ASR.1</td>
<td>0.0602</td>
<td>0.3654</td>
<td>0.4644</td>
<td>0.4822</td>
<td>0.4479</td>
<td>0.8148</td>
<td>0.7018</td>
<td>76.1960</td>
<td>0.4009</td>
<td>2.9995</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TASK</th>
<th>CT_EC</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>case+punc</td>
<td>bleu</td>
<td>meteor</td>
<td>f1</td>
<td>prec</td>
<td>recl</td>
<td>wer</td>
<td>per</td>
<td>ter</td>
<td>gtm</td>
<td>nist</td>
</tr>
<tr>
<td>Proposed.CRR</td>
<td>0.2759</td>
<td>0.5328</td>
<td>0.5500</td>
<td>0.5150</td>
<td>0.5900</td>
<td>0.7421</td>
<td>0.5382</td>
<td>68.6970</td>
<td>0.6914</td>
<td>5.3888</td>
</tr>
<tr>
<td>Moses.CRR</td>
<td>0.3391</td>
<td>0.5744</td>
<td>0.6204</td>
<td>0.6430</td>
<td>0.5994</td>
<td>0.5942</td>
<td>0.4356</td>
<td>52.3780</td>
<td>0.6930</td>
<td>6.1764</td>
</tr>
<tr>
<td>Systran.CRR</td>
<td>0.2300</td>
<td>0.5063</td>
<td>0.5596</td>
<td>0.5599</td>
<td>0.5594</td>
<td>0.6993</td>
<td>0.4987</td>
<td>63.2230</td>
<td>0.6304</td>
<td>5.4766</td>
</tr>
<tr>
<td>Proposed.ASR.1</td>
<td>0.2214</td>
<td>0.4417</td>
<td>0.4516</td>
<td>0.4100</td>
<td>0.5025</td>
<td>0.8518</td>
<td>0.6447</td>
<td>80.8210</td>
<td>0.6399</td>
<td>4.5091</td>
</tr>
<tr>
<td>Moses.ASR.1</td>
<td>0.2853</td>
<td>0.5134</td>
<td>0.5604</td>
<td>0.5784</td>
<td>0.5435</td>
<td>0.6609</td>
<td>0.4986</td>
<td>59.2510</td>
<td>0.6331</td>
<td>5.4212</td>
</tr>
<tr>
<td>Systran.ASR.1</td>
<td>0.1902</td>
<td>0.4483</td>
<td>0.4986</td>
<td>0.4948</td>
<td>0.5025</td>
<td>0.7627</td>
<td>0.5683</td>
<td>70.5120</td>
<td>0.5689</td>
<td>4.6699</td>
</tr>
</tbody>
</table>
Discussion

<no native speakers>

Unknown Words
  Proposed method:
    very few unknown words

Grammatical Correctness
  Proposed method:
    more grammatically correct sentences.

However, the BLEU score was not so good?
Conclusion

Our System:
   Two-stage machine translation system.
      First stage : Rule-based machine translation
      Second stage : SMT

Object:
   Fewer unknown words &
   Fewer ungrammatical sentences.

Results:
   Not so good

Future:
   a) Optimize parameters & reordering model
   b) SYSTRAN ?
# Results of Parameter Tunings

<table>
<thead>
<tr>
<th>TASK</th>
<th>BTEC_CE</th>
</tr>
</thead>
<tbody>
<tr>
<td>case+punc</td>
<td>bleu</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.3351</td>
</tr>
<tr>
<td>Moses</td>
<td>0.3423</td>
</tr>
<tr>
<td>Systran</td>
<td>0.1070</td>
</tr>
</tbody>
</table>