Statistical Machine Translation without Long Parallel Sentences for Training Data

○ Jin'ichi Murakami, Masato Tokuhisa, Satoru Ikehara
  (Tottori University, Japan)
The Strategy of Our Statistic Machine Translation

1) Long Phrase Tables (Adequacy)

Adequacy ~ translation model $P(E/C)$
long phrase tables $=$ achieve high accuracy
20 words

English to German
Word position change is very small.
→ short phrase table

Chinese to English
Some word are moved from their original position.
→ long phrase tables.
The Strategy of Our Statistic Machine Translation

2) 4-gram Model (Fluency)

- Fluency ~ language model \( P(E) \)

  Not use higher N-gram model.
  (the reliability for each parameter becomes low)

  normal 4-gram model
The Strategy of Our Statistic Machine Translation

3) Remove Long Parallel Sentences

Long Parallel Sentences $\rightarrow$ Wrong Phrase Table
$\rightarrow$ Low Blue Score

Much Parallel Sentences $\rightarrow$ High Translation?
The Strategy of Our Statistic Machine Translation

4) Standard Tools

GIZA++.2003-09-30.tar.gz

moses.2007-05-29.tgz

training-release-1.3.tgz(train-phrase-model.perl)

(Made only some small tools to build a temporal corpus.)
C 1 在 门 厅 下 面 。 我 这 就 给 您 拿 一 些 。如 果 您 还 有 什 么 需 要 的 请 告 诉 我 。

C 2 不 用 担 心 那 个 。我 要 买 它 你 不 需 要 把 它 包 起 来 。

C 3 你 可 以 改 改 吗 ？

C 4 红 绿 灯 是 红 的 。

C 5 我 们 想 要 一 张 靠 窗 户 的 桌 子 。

E 1 It's just down the hall. I'll bring you some now. If there is anything else you need, just let me know.

E 2 No worry about that. I'll take it and you need not wrap it up.

E 3 Do you do alterations?

E 4 The light was red.

E 5 We want to have a table near the window.
E1 it's just down the hall i'll bring you some now if there is anything else you need just let me know
E2 no worry about that i'll take it and you need not wrap it up
E3 do you do alterations
E4 the light was red
E5 we want to have a table near the window

C1 在 门厅 下面 。 我 这 就 给 您 拿 一些 。 如果 您 还 有 什么 需要 的 请 告诉 我 。
C2 不用 担心 那个 。我 要 买 它 你 不 需要 把 它 包 起来 。
C3 你 可以 改改 吗 ？
C4 红绿 灯 是 红 的 。
C5 我们 想 要 张 靠 窗户 的 桌子 。

Challenge-EC training-data
(Small case,Punctuation procedure)
Long Phrase Table

train-phrase-model.perl (training-release-1.3.tgz)

Long phrase table:
   Max-phrase-length: 20  (default 7)

Other parameters :defaults value.
Example of Phrase Tables (BTEC-CE)

一 个 日语 导游  ||||
    a Japanese speaking guide  ||||
    0.5 0.00339841 0.333333 0.00723042 2.718
一 个 日语 导游 吗 ？  ||||
    a Japanese speaking guide ？  ||||
    1 0.000771748 0.5 0.00668676 2.718
一 个 时钟 收音机 谢谢  ||||
    a clock radio , please  ||||
    1 0.000602041 1 0.0325873 2.718
一 个 明天 十点 开始 的  ||||
    a tee-off time for ten tomorrow  ||||
    1 0.000547434 1 3.02033e-05 2.718
一 个 明治 神殿 的 护身符 可以 预 知  ||||
    A charm from Meiji shrine , a written oracle key holder  ||||
    1 3.76995e-05 1 7.77965e-08 2.718
4-gram language model

Best language model for IWSLT2007)

Stanford Research Institute Language Model (SRILM) toolkit

smoothing parameter : " -ukndiscount -interpolate".

19972 parallel sentences
1-gram, 8346 lines
2-gram, 49685 lines
3-gram, 17241 lines
4-gram, 14651 lines
It's just down the hall. I'll bring you some now. If there is anything else you need, just let me know.
在门厅下面。我这就给您拿一些。如果有需要的请告诉我。

I twisted it playing tennis. It felt Okay after the game but then it started turning black-and-blue. Is it serious?
我打网球时扭伤的。刚打完后觉得没什么可是现在它开始变得青一块紫一块的。它严重吗？

I'm looking for a nice, quiet grill-type restaurant. Would you point them out on this map?
我在找一家好点的安静的烧烤类型的餐馆。你能在这张地图上指出它们吗？

The pleasure is all mine, Mr. Green. I've heard a lot about you from Mr. Smith.
我真高兴格林先生。我从史密斯先生那儿听到很多有关你的情况。

From two hours before the departure. And please come to the counter at least thirty minutes before flight time.
从起飞前两个小时开始办理。请至少在班机起飞前三十分钟来到柜台。
Decoder: Moses (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
                (Cross Entropy)
weight-w      -1
distortion-limit  -1
                (The position of the verb changed significantly)
```
## Results of Challenge-EC

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>bleu</th>
<th>nist</th>
<th>wer</th>
<th>per</th>
<th>gtm</th>
<th>meteor</th>
<th>ter</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(case+punc)</strong></td>
<td><strong>primary</strong></td>
<td>ASR.1</td>
<td>0.35</td>
<td>5.67</td>
<td>0.54</td>
<td>0.46</td>
<td>0.86</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CRR</td>
<td>0.4</td>
<td>6.19</td>
<td>0.49</td>
<td>0.4</td>
<td>0.85</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td><strong>contrast</strong></td>
<td>ASR.1</td>
<td>0.36</td>
<td>5.92</td>
<td>0.54</td>
<td>0.45</td>
<td>0.86</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CRR</td>
<td>0.4</td>
<td>6.48</td>
<td>0.48</td>
<td>0.38</td>
<td>0.86</td>
<td>0.62</td>
</tr>
<tr>
<td><strong>(no-case +no-punc)</strong></td>
<td><strong>primary</strong></td>
<td>ASR.1</td>
<td>0.33</td>
<td>5.62</td>
<td>0.59</td>
<td>0.49</td>
<td>0.84</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CRR</td>
<td>0.38</td>
<td>6.17</td>
<td>0.53</td>
<td>0.42</td>
<td>0.83</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td><strong>contrast</strong></td>
<td>ASR.1</td>
<td>0.33</td>
<td>5.9</td>
<td>0.59</td>
<td>0.48</td>
<td>0.84</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CRR</td>
<td>0.39</td>
<td>6.47</td>
<td>0.52</td>
<td>0.41</td>
<td>0.84</td>
<td>0.6</td>
</tr>
</tbody>
</table>

**Primary : Standard moses:**
- 19972 English-Chinese parallel sentences

**Contrast:** Remove Long Parallel Sentences (> 96 char)
- 19387 English-Chinese parallel sentences
## Results of Challenge-CE

<table>
<thead>
<tr>
<th></th>
<th>bleu</th>
<th>nist</th>
<th>wer</th>
<th>per</th>
<th>gtm</th>
<th>meteor</th>
<th>ter</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>primary</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(case+punc)</td>
<td>ASR.1</td>
<td>0.23</td>
<td>4.38</td>
<td>0.65</td>
<td>0.58</td>
<td>0.55</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>CRR</td>
<td>0.27</td>
<td>4.7</td>
<td>0.62</td>
<td>0.55</td>
<td>0.58</td>
<td>0.5</td>
</tr>
<tr>
<td>(no-case +no-punc)</td>
<td>ASR.1</td>
<td>0.21</td>
<td>4.15</td>
<td>0.67</td>
<td>0.6</td>
<td>0.53</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>CRR</td>
<td>0.26</td>
<td>4.56</td>
<td>0.64</td>
<td>0.57</td>
<td>0.56</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Primary: Standard moses

19972 Chinese-English parallel sentences

Contrast: Remove Long Parallel Sentences (> 48 char)

19327 Chinese-English parallel sentences
## Results of BTEC-CE

<table>
<thead>
<tr>
<th></th>
<th>bleu</th>
<th>nist</th>
<th>wer</th>
<th>per</th>
<th>gtm</th>
<th>meteoi</th>
<th>ter</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(case+punc)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>primary</td>
<td>0.29</td>
<td>6.03</td>
<td>0.62</td>
<td>0.54</td>
<td>0.6</td>
<td>0.49</td>
<td>54.18</td>
</tr>
<tr>
<td>CRR</td>
<td>0.33</td>
<td>6.42</td>
<td>0.59</td>
<td>0.5</td>
<td>0.64</td>
<td>0.52</td>
<td>50.45</td>
</tr>
<tr>
<td>contrast</td>
<td>0.28</td>
<td>5.89</td>
<td>0.64</td>
<td>0.55</td>
<td>0.59</td>
<td>0.48</td>
<td>55.46</td>
</tr>
<tr>
<td><strong>(no-case no-punc)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>primary</td>
<td>0.31</td>
<td>6.89</td>
<td>0.61</td>
<td>0.52</td>
<td>0.62</td>
<td>0.52</td>
<td>53.65</td>
</tr>
<tr>
<td>CRR</td>
<td>0.35</td>
<td>7.43</td>
<td>0.57</td>
<td>0.47</td>
<td>0.67</td>
<td>0.55</td>
<td>49.15</td>
</tr>
<tr>
<td>contrast</td>
<td>0.29</td>
<td>6.72</td>
<td>0.63</td>
<td>0.53</td>
<td>0.62</td>
<td>0.51</td>
<td>55.48</td>
</tr>
</tbody>
</table>

Primary: Standard moses
19972 Chinese-English parallel sentences
Contrast: Remove Long Parallel Sentences (> 48 char)
19327 Chinese-English parallel sentences
Results of IWSLT08

Remove Long Parallel Corpus

→ Effective for Challenge-EC
→ Not effective for Challenge-CE & BTEC-CE

(Effective for Many Cases)
01. Do you or the Yasaka-shrine and yasaka-shrine 和 吗？
02. Could you tell me where the Osaka Chinese out?
03. Do or I can find the Yasaka-shrine out?
04. Where is the Yasaka-shrine yasaka-shrine 在 哪？
05. Well they're chips please out.
06. Which way to Kiyomizu-temple kiyomizu-temple 怎么 走 ？
07. How can I get there out?
08. How long does it take to Kiyomizu-temple kiyomizu-temple 要 花 多长时间 ？
09. Here comes our boss now out.
10. So far people seem quite receptive to talking to the machine of oh it's quite difficult ah because of the position of the microphone out.
01. Do you know where the Yasaka Shrine is?
   - 你 知道 Yasaka Shrine 在 哪 ？

02. Could you tell me where the Yasaka Shrine is?
   - 你 能 告诉 我 在 哪 Yasaka Shrine 吗 ？

03. Do you know where I can find the Yasaka Shrine?
   - 你 知道 我 在 哪 能 找到 Yasaka Shrine 。

04. Where is the Yasaka Shrine?
   - Yasaka Shrine 在 哪 ？

05. World Heritage please.
   - 世界 Heritage 。

06. Which way to Kiyomizu Temple?
   - Kiyomizu 寺庙 怎么 走 ？

07. How can I get there?
   - 我 怎么 去 ？

08. How long does it take to Kiyomizu Temple?
   - Kiyomizu 寺庙 要 花 多长 时间 ？

09. Here comes the bus now.
   - 现在 巴士 来 了 。

10. So far people seem quite receptive talking the machine although it's quite difficult because of the position of a microphone.
   - 很 像 是 很 难 说话 远 receptive 虽然 因为 这个 机器 。position 的 microphone 的 人 。

Outputs for Challenge-EC CRR
01 in 旅游 费用 包括 晚餐 吗
  out Does it include dinner, the charge.
02 in 我 是 歌手
  out I 是 歌手.
03 in 怎样 才 能 到 那里 去 呢
  out How about going to get there
04 in 请 叫 服务生 来
  out 服务生 call, please.
05 in 这个 怎么样
  out How about this
06 in 只要 按 这里 就 行
  out Just press right here.
07 in 我 买了 成了
  out I bought into.
08 in 请 再 拿 一 条 毛巾 来
  out Bring a towel, please.
09 in 这个 菜 没有 点
  out The food hasn't p.
10 in 死 于 日本 的 营销 人员 来 说 关系 把 越 大 越 好 不过 关系网 或者 通过 别人 介 绍 而 开设 的 工作 肯定 会 有 一 个 良好 的 开端
  out I must be fine, but there is a large, the better leave as in Japan 来 说 关系 or by the 开设 job 开端 recommend 关系网 someone else 死 marketing 人员
01in 旅游 费用 包括 晚餐 吗

outDoes it include dinner, the charge.

02in 我 是 歌手

outI'm a singer.

03in 怎样 才 能 到 那里 去 呢

outHow about going to get there

04in 请 叫 服务生 来

outCall the waiter, please.

05in 这个 怎么样

outHow about this

06in 只 按 这里 就 行 了

outJust press here just

07in 我 买了 成药

outI'll take medicine.

08in 请 再 拿 一条 毛巾 来

outBring a towel, please.

09in 这个 菜 没有 点

outThe food hasn't been ordered.

10in 对于 日本 的 营销 人员 来说 关系网 越大 越好 通过 关系网 或者 通过 别人 介绍 而 开始 的 工作 肯定 会 有一个 良好的 开端

outAnd for Japan by the must be fine, recommend a large, the better leave as relationship network or by the job someone else's relationship network marketing personnel to say

Outputs for BTEC-CE CRR
outpostcard, please.

Could you tell me the post office Where

Could you tell me the post office Where

Excuse me, but I'd like to buy a stamp

Where is the counter

I want to postcard

Wait a minute, uh, please.

How do I get to the post office

How do I get to the nearest post office

Do you have any commemorative stamps
Please write postcard

Could you tell me the post office where

Could you tell me the post office where

Excuse me, but I'd like to buy a stamp

Where is the counter

I want to postcard

Wait a minute please.

How do I get to the post office

How do I get to the nearest post office

Do you have any commemorative stamps

Outputs for Challenge-CE CRR
Consideration: Remove Unknown Words

### Results of BTEC-CE

<table>
<thead>
<tr>
<th></th>
<th>bleu</th>
<th>nist</th>
<th>wer</th>
<th>per</th>
<th>gtm</th>
<th>meteor</th>
</tr>
</thead>
<tbody>
<tr>
<td>(case+punc)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>primary CRR</td>
<td>0.34</td>
<td>6</td>
<td>0.58</td>
<td>0.5</td>
<td>0.66</td>
<td>0.54</td>
</tr>
<tr>
<td>contrast CRR</td>
<td>0.33</td>
<td>5.93</td>
<td>0.59</td>
<td>0.51</td>
<td>0.65</td>
<td>0.53</td>
</tr>
<tr>
<td>(no-case +no-punc)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>primary CRR</td>
<td>0.37</td>
<td>7.15</td>
<td>0.55</td>
<td>0.46</td>
<td>0.69</td>
<td>0.57</td>
</tr>
<tr>
<td>contrast CRR</td>
<td>0.36</td>
<td>7.08</td>
<td>0.57</td>
<td>0.47</td>
<td>0.68</td>
<td>0.56</td>
</tr>
</tbody>
</table>

### Results of Challenge-EC

<table>
<thead>
<tr>
<th></th>
<th>bleu</th>
<th>nist</th>
<th>wer</th>
<th>per</th>
<th>gtm</th>
<th>meteor</th>
</tr>
</thead>
<tbody>
<tr>
<td>(case+punc)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>primary CRR</td>
<td>0.41</td>
<td>6.13</td>
<td>0.48</td>
<td>0.39</td>
<td>0.87</td>
<td>0.6</td>
</tr>
<tr>
<td>contrast CRR</td>
<td>0.41</td>
<td>6.48</td>
<td>0.47</td>
<td>0.38</td>
<td>0.87</td>
<td>0.62</td>
</tr>
<tr>
<td>(no-case +no-punc)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>primary CRR</td>
<td>0.4</td>
<td>6.04</td>
<td>0.52</td>
<td>0.42</td>
<td>0.85</td>
<td>0.58</td>
</tr>
<tr>
<td>contrast CRR</td>
<td>0.39</td>
<td>6.41</td>
<td>0.51</td>
<td>0.41</td>
<td>0.86</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Primary: Standard moses  
Contrast: Remove Long Parallel Sentences
Future study

- Optimize parameters
- Unknown word procedure
- More large database
- Not used parallel sentence
  (If output likelihood is high, use as parallel sentence)
Conclusions

- Remove Long Parallel Sentences
- Long phrase table
- Standard tools
- Statistical Example Based Translation

- Good results for Change-EC
- 0.4047 BLEU score for IWSLT08