An EBMT System Based on Word Alignment

HOU Hongxu, DENG Dan, ZOU Gang, YU Hongkui, LIU Yang, XIONG Deyi and LIU Qun

hxhou@ict.ac.cn
Introduction

- EBMT
- Based on word alignment
- A 220K sentence pairs’ corpus
- 5 steps of translation
System Architecture

- **Corpus**
  - 220k sentence pairs
  - 460k words and phrases

- **Program**
  - Matching and searching
  - Fragment matching
  - Fragment assembling
  - Evaluation
  - Post processing
Corpus

- **Scale**
  - 220K sentence pairs, includes news, literal, dictionaries and dialogues
  - 460K words and phrases dictionary

- **POS tagged/tokenized**
  - ICTCLAS 2.0
  - Accuracy: 97.8%

- **Word aligned**
  - Based on large dictionary
  - Precision: 84.0%
  - Recall: 62.9%
Matching and Searching

- Searching the most similar sentences
- Measure of similarity
  - Weight of POS
  - Run length
Matching and Searching (cont.)

- An example

S: 能/v 给/p 我/at 药/n 和/cc 一/m 杯/q 水/n 吗/y ?/ww
(0.165) S₁: 能/v 给/p 我/at 些/q 药/n 吗/y ?/ww
(0.1547) S₂: 能/v 给/p 我/at 开/v 药/n 吗/y ?/ww
Matching and Searching (cont.)

• Searching
  - A index of words/POS
  - Remove high frequency words.
  - Remove named entities
  - Short sentences
Fragment matching

- Matching and mismatching fragment
- Word matched and POS matched
Fragment matching (cont.)

- An example

  S: 我/rr 的/ude1 脚盹/n 被/pbei 冲走/v 了/y

  S₁: 桥/n 被/pbei 冲走/v 了/y

  (2) (1)
Target fragment finding

- Based on word alignment
  - Target contents
    - Which words should be placed
  - Target positions
    - Which position should be placed
Target fragment finding

- An example

桥/n 被/pbei 冲走/v 了/y

The/DT bridge/NN was/VBD washed/VBN away/RB
Fragment assembling

- An example

S: 我/rr 的/ude1 脚蹼/n 被/pbei 冲走/v 了/y
S1: 桥/n 被/pbei 冲走/v 了/y
T1: The/DT bridge/NN was/VBD washed/VBN away/RB
T: My/PRP$ The/DT web/NN was/VBD was/VBD washed/VBN away/RB
Make choices

- Whether the non-aligned words should be appeared in the target sentence
- N-gram
Make choices (cont.)

• An example

(938.48) $T_1$: my the web was washed away
(858.60) $T_2$: my web was washed away
## Evaluation results

<table>
<thead>
<tr>
<th></th>
<th>Official results</th>
<th>Correct results</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>0.0798 (9)</td>
<td>0.2013 (8)</td>
</tr>
<tr>
<td>GTM</td>
<td>0.3862 (9)</td>
<td>0.6380 (5)</td>
</tr>
<tr>
<td>NIST</td>
<td>3.6443 (9)</td>
<td>6.4716 (5)</td>
</tr>
<tr>
<td>WER</td>
<td>0.8466 (9)</td>
<td>0.6275 (8)</td>
</tr>
<tr>
<td>PER</td>
<td>0.7650 (9)</td>
<td>0.5187 (7)</td>
</tr>
<tr>
<td>Fluency</td>
<td>2.7180 (7)</td>
<td></td>
</tr>
<tr>
<td>Adequacy</td>
<td>3.0820 (5)</td>
<td></td>
</tr>
</tbody>
</table>

* We have made a mistake in submitting result
Some cases

• Existed in corpus
• Good fragments of an example
• Replace some words
• Part of example
• Combine of phrases
• Small fragments
Further improvement

- Phrase recognition
- Word alignment
- Word cluster
- Sentence classify
- Corpus
Thank you