The TRGTK's Patent MT System Description for NTCIR-10
Hao Xiong and Weihua Luo
Torangetek Inc.
Key Lab. of Intelligent Information Processing
{xionghao, luoweihua}@torangetek.com

Overall Architecture

System Bilingual Monolingual Patent Grant Data
Bilingual Training Corpus Monolingual Training Corpus
Segmentation, Tokenization Word Alignment Language Model Training

Step 1

System
Bilingual Monolingual Patent Grant Data
Segmentation, Tokenization Word Alignment Language Model Training

Step 2

Testing Documents Decoder Translation History Memory Translation Results

Document-level Decoding

S0: W0, W1, W2, ..., Wn
S1: W0, W1, W2, ..., Wn
S2: W0, W1, W2, ..., Wn...

Score(Trans_01) = LocalScore(Trans_01) + LM(Trans0, Trans1, Trans2 in S0) + LM(Trans0, Trans1, Trans2 in S1) + LM(Trans0, Trans1, Trans2 in S2)

Parallel Segmenter

Cn, C1, C2, ..., Cn
Cn, C1, C2, ..., Cn

CPU

Segmentation History Storage

Parallel Decoding

System
GPU
Time (minutes)
Sys1 500
0.3
Sys1 1
32
Sys2 500
0.5
Sys2 1
40

Data Usage

<table>
<thead>
<tr>
<th>System</th>
<th>Bilingual</th>
<th>Monolingual</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-E</td>
<td>1 Million</td>
<td>40 Million</td>
</tr>
<tr>
<td>J-E</td>
<td>3 Million</td>
<td>40 Million</td>
</tr>
<tr>
<td>E-J</td>
<td>3 Million</td>
<td>73 Million</td>
</tr>
</tbody>
</table>

Final Results

<table>
<thead>
<tr>
<th>System</th>
<th>C-E</th>
<th>J-E</th>
<th>E-J</th>
</tr>
</thead>
<tbody>
<tr>
<td>IE</td>
<td>34.63</td>
<td>26.99</td>
<td>32.21</td>
</tr>
<tr>
<td>ChE</td>
<td>33.46</td>
<td>26.34</td>
<td>31.4</td>
</tr>
<tr>
<td>ME</td>
<td>21.52</td>
<td>26.34</td>
<td></td>
</tr>
</tbody>
</table>

Parallel Segmenter

Sys1: use segmentation historical storage (SHS)
Sys2: does not use SHS
Corpus: 1 million training data + 10 million Chinese sentences generated SHS

Parallel Decoding

System
GPU
Time (minutes)
Parallel training 100
8 hours
Training 1
45 hours
Parallel decoding 100
800 words/s
Parallel decoding + Translation Memory 100
1000 words/s
Decoding 1
50 words/seconds

Term Recognition

C-Value

\[ C-value(a) = \left\{ \begin{array}{ll}
\log_2|a| \cdot f(a), & a \text{ is not nested,} \\
\log_2|a| \cdot f(a) - \frac{1}{\log_2 T(a)} \sum f(b) & \text{otherwise}
\end{array} \right. \]

Candidate Strings

<table>
<thead>
<tr>
<th>2 words</th>
<th>3 words</th>
<th>4 words</th>
<th>5 words</th>
<th>6 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>a+b</td>
<td>b+b</td>
<td>b+b</td>
<td>b+b</td>
<td>b+b</td>
</tr>
<tr>
<td>a+b</td>
<td>b+b</td>
<td>b+b</td>
<td>b+b</td>
<td>b+b</td>
</tr>
<tr>
<td>a+b</td>
<td>b+b</td>
<td>b+b</td>
<td>b+b</td>
<td>b+b</td>
</tr>
<tr>
<td>a+b</td>
<td>b+b</td>
<td>b+b</td>
<td>b+b</td>
<td>b+b</td>
</tr>
</tbody>
</table>

a is adjective, b is distinguish word, c is conjunction, d is adverb, n is noun, m is numbers, v is verb, u is particle, v|n is verb or noun.