Exploiting Alignment Techniques in MaTrEx: the DCU MT System for IWSLT 2008

Yanjun Ma†, John Tinsley†, Hany Hassan†, Jinhua Dut, Andy Way‡
National Centre for Language Technology† and Centre for Next Generation Localisation‡
School of Computing,
Dublin City University, Dublin 9, Ireland
{jtinsley, yma, hhassan, jdu, away}@computing.dcu.ie

Proper training on large data is always preferential.
Syntax-enhanced word alignment leads to improvements.
Treebank phrase extraction improves inconsistently.
Word packing leads to drops in performance.
Smoothing and case/punctuation restoration techniques also effective.

MaTrEx (Machine Translation using Examples) is a hybrid system which can exploit EBMT, SMT and syntax-based techniques to build a combined translation model. MaTrEx is built following established Design Patterns and consists of a number of extensible and re-implementable modules. Some significant modules include:

- **Word Alignment Module**: outputs a set of word alignments given a parallel corpus;
- **Chunking Module**: outputs a set of chunks given an input corpus;
- **Chunk Alignment Module**: outputs aligned chunk pairs given source and target chunks from comparable corpora;
- **Decoder**: returns optimal translation given a set of aligned sentences, chunk/phrase and word pairs.

**New Alignment Techniques**

**Word Packing**

- **Candidate Extraction**

白葡萄酒：white wine
抱歉：excuse me
报警：call the police
fifteen：十五
here：在这里

- **Reliability Estimation**

**Bootstrapping Estimation**

**Syntax-enhanced Word Alignment**

- **Anchor Word Alignment**

- **Discriminative Syntax-Enhanced Word Alignment**

- **Search**

\[ \text{c: 我_1 打_2 网球_3 时_4 扭伤_5 的_6 到_7 } \]
\[ \text{e: I_1 twisted_2 it_3 playing_4 tennis_5 } \]

**Treebank-based Phrase Alignment**

- **Translation-based Punctuation Restoration**

**Majority voting techniques to restore the final punctuation mark**

**Official Results**

<table>
<thead>
<tr>
<th>System</th>
<th>Challenge Task</th>
<th>BTEC</th>
<th>Pivot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZH-EN</td>
<td>EN-ZH</td>
<td>AR-EN</td>
</tr>
<tr>
<td><strong>Baseline</strong></td>
<td>31.94</td>
<td>27.14</td>
<td>40.80</td>
</tr>
<tr>
<td><strong>Word Packing</strong></td>
<td>29.67</td>
<td>26.76</td>
<td>40.04</td>
</tr>
<tr>
<td><strong>Syntax-Enhanced</strong></td>
<td>34.52</td>
<td>29.31</td>
<td>42.43</td>
</tr>
<tr>
<td><strong>Treebank</strong></td>
<td>28.81</td>
<td>26.53</td>
<td>37.73</td>
</tr>
<tr>
<td><strong>OOV Smoothing</strong></td>
<td>32.59</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>All Smoothing</strong></td>
<td>23.95</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Data Combo</strong></td>
<td>36.40</td>
<td>30.86</td>
<td>46.30</td>
</tr>
</tbody>
</table>

**Discussion**

- Proper training on large data is always preferential.
- Syntax-enhanced word alignment leads to improvements.
- Treebank phrase extraction improves inconsistently.
- Word packing leads to drops in performance.
- Smoothing and case/punctuation restoration techniques also effective.