Low-Resource Machine Translation Using \textsc{MaTrEx}: The DCU Machine Translation System for IWSLT 2009

Yanjun Ma, Tsuyoshi Okita, Özlem Çetinoğlu, Jinhua Du, Andy Way, Dublin City University, CNGL/School of Computing
IWSLT T Rationale

- IWSLT pursues research aspects: No additional resources other than corpora provided.

  ... certain gains in performance were triggered by better suited language resources (engineering aspects) or by improvements in the underlying decoding algorithms and statistical models (research aspects). (IWSLT organizer)
MaTrEx: Low-Resource Machine Translation

- MaTrEx for Low-Resource MT
  - Word Lattice
    - Rational: We have space to investigate various segmentation in Chinese and Turkish.
  - Noise Reduction
    - Rational: There would be various paraphrases, multiword expressions, non-literal translations included in bitext.
- Multiple System Combination
- Case and Punctuation Restoration
MaTrEx: Low-Resource Machine Translation

- MaTrEx for Low-Resource MT
  - Word Lattice
    - Rational: We have space to investigate various segmentation in Chinese and Turkish.
  - Noise Reduction
    - Rational: There would be various paraphrases, multiword expressions, non-literal translations included in bitext.
- Multiple System Combination
- Case and Punctuation Restoration

- MaTrEx participated in 2006/7/8/9, Turkish first time
IWSLT 2009 Corpora

- **BTEC** task (Basic Travel Expression Corpus) and **CHALLENGE** task (which uses Spoken Language Databases corpus).
  - BTEC task: Chinese-English and Turkish-English
  - CHALLENGE task: Chinese-English and English-Chinese
IWSLT 2009 Corpora

- **BTEC** task (Basic Travel Expression Corpus) and **CHALLENGE** task (which uses Spoken Language Databases corpus).
  - **BTEC** task: Chinese-English and Turkish-English
  - **CHALLENGE** task: Chinese-English and English-Chinese

<table>
<thead>
<tr>
<th>Task</th>
<th>Train Set</th>
<th>Dev Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>BT-TR-EN</td>
<td>27,972</td>
<td>506 (×16)</td>
<td>469</td>
</tr>
<tr>
<td>BT-ZH-EN</td>
<td>47,098</td>
<td>507 (×16)</td>
<td>469</td>
</tr>
<tr>
<td>CH-ZH-EN</td>
<td>75,231</td>
<td>489 (×7)</td>
<td>405</td>
</tr>
<tr>
<td>CH-EN-ZH</td>
<td>39,228</td>
<td>210 (×4)</td>
<td>393</td>
</tr>
</tbody>
</table>

Table: Parallel corpus size of IWSLT 2009 (Only our participated tasks)
Table Of Contents

1. MaTrEx
2. Four Techniques Investigated
3. Experiments
4. Conclusions
Speech recognition: first determine the best word segmentation and perform decoding (the acoustic signal underdetermines the choice of source word sequence).

\[
\hat{v}_1^K = \arg \max_{v_1^K, K} \{ P(v_1^K | f_1^I) \}, \quad \hat{e}_1^J = \arg \max_{e_1^J, J} \{ P(e_1^J | \hat{v}_1^K) \}
\]
Word Lattice

- **Speech recognition**: first determine the best word segmentation and perform decoding (the acoustic signal underdetermines the choice of source word sequence).

\[
\hat{v}_1^K = \arg \max_{v_1^K, K} \{ P(v_1^K | f_1^I) \}, \quad \hat{e}_1^J = \arg \max_{e_1^J, J} \{ P(e_1^J | \hat{v}_1^K) \}
\]

- **Word lattice-based approach** in SMT: to allow the MT decoder to consider all possibilities for \( f \) by encoding the alternatives compactly as a word lattice. [Xu et al., 2005][Bertoldi et al., 2007][Dyer et al., 2008][Ma and Way, EACL2009].

\[
\hat{e}_1^J = \arg \max_{e_1^J, J} \{ \max_{v_1^K, K} P(e_1^J, v_1^K | f_1^I) \} = \arg \max_{e_1^J, J} \{ \max_{v_1^K, K} P(e_1^J) P(v_1^K | e_1^J, f_1^I) \} \approx \arg \max_{e_1^J, J} \{ \max_{v_1^K, K} p(e_1^J) p(v_1^K | f_1^I) p(v_1^K | e_1^J) \}
\]
Word Lattice: Generation (Chinese)

Chinese (word boundaries are not orthographically marked)
在门厅下面。我这就给您拿一些。
(zai men ting xia mian. wo zhe jiu gei nin na yi xie)
Chinese (word boundaries are not orthographically marked)
在门厅下面。我这就给您拿一些。
(zai men ting xia mian. wo zhe jiu gei nin na yi xie)

1. Manual segmentation
在_门厅_下面_。_我_这_就_给_您_拿_一些_。
Chinese (word boundaries are not orthographically marked)
在门厅下面。我这就给您拿一些。
(zai men ting xia mian. wo zhe jiu gei nin na yi xie)

1. Manual segmentation
   在_门厅_下面_。_我_这_就_给_您_拿_一些_。

2. LDC segmentation
   在_门_厅_下面_。_我_这_就_给_您_拿_一些_。
Word Lattice: Generation (Chinese)

Chinese (word boundaries are not orthographically marked)
在门厅下面。我这就给您拿一些。
(zai men ting xia mian. wo zhe jiu gei nin na yi xie)

1. Manual segmentation
   在_门厅_下面_。_我_这_就_给_您_拿_一些_。

2. LDC segmentation
   在_门厅_下面_。_我_这_就_给_您_拿_一些_。

3. Character-based segmentation
   在_门厅_下面_。_我_这_就_给_您_拿_一些_。
Word Lattice: Generation (Turkish)

Turkish (rich morphology language)

Bu mevsimin en yeni rengi ne?

1. lowercased original data
   - each word is a segment
     bu mevsimin en yeni rengi ne?
Word Lattice: Generation (Turkish)

**Turkish (rich morphology language)**

*Bu mevsimin en yeni rengi ne?*

1. lowercased original data
   - each word is a segment
     *bu mevsimin en yeni rengi ne?*

2. morphologically analyzed [Oflazer, 94] and disambiguated [Sak, 07], and reduced analysis, i.e., only informative morphemes are kept [Oflazer].
   - each analysis is a segment
     *bu+Det mevsim+Noun+Gen en+Adverb yeni+Adj renk+Noun+P3sg ne+Adverb ?*
Word Lattice: Generation (Turkish)

Turkish (rich morphology language)

\textbf{Bu mevsimin en yeni rengi ne?}

1. lowercased original data
   - each word is a segment
     \textit{bu mevsimin en yeni rengi ne ?}

2. morphologically analyzed [Oflazer, 94] and disambiguated [Sak, 07], and reduced analysis, i.e., only informative morphemes are kept [Oflazer].
   - each analysis is a segment
     \textit{bu Det mevsim Noun Gen en Adverb yeni Adj renk Noun P3sg ne Adverb ?}
   - each morpheme is a segment
     \textit{bu Det mevsim Noun Gen en Adverb yeni Adj renk Noun P3sg ne Adverb ?}
Word Lattice: An Example

Figure: An example of a word lattice for a Chinese sentence

- Arc: segmented words.
- Numbers at arc: transition probabilities (1, 1/3, 1/2, and so forth).
Noise Reduction in MT

- Noise: statistical property
  - Noise reduction for phrase alignment [Tomeh et al., 2009]
- Outlier: dependent on underlying machine learning algorithm
  - Noise reduction for word alignment [Okita, ACL09SRW]
- Noise: defined by similarity measure (In sentence alignment, the removal of some particular sentence does not matter the quality in later stages)
  - Noise reduction for sentence alignment [Utiyama and Isahara, 2003]
(Training Phase) We let our MT systems learn by training set.

<table>
<thead>
<tr>
<th>Original Text</th>
<th>MT Systems</th>
<th>Noisy Channel</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>c’ est la vie .</td>
<td>MT Systems</td>
<td>Noisy Channel</td>
<td>that is life .</td>
</tr>
<tr>
<td>je t’ aime .</td>
<td></td>
<td></td>
<td>i love you .</td>
</tr>
<tr>
<td>elle est petite .</td>
<td></td>
<td></td>
<td>she is small .</td>
</tr>
</tbody>
</table>
(Training Phase) We let our MT systems learn by training set.

\[
\begin{array}{ccc}
\text{c' est la vie .} & \text{MT Systems} & \text{that is life .} \\
\text{je t' aime .} & \Rightarrow \text{Noisy Channel} & \Rightarrow \text{i love you .} \\
\text{elle est petite .} & \Rightarrow \text{MT Systems} & \Rightarrow \text{??} \\
\end{array}
\]

(Testing Phase) We can expect if we translate our training set our MT systems learn most of them in good faith (considering a bit about generalisation error).

\[
\begin{array}{ccc}
\text{c' est la vie .} & \text{above trained} & \text{that is life .} \\
\text{je t' aime .} & \Rightarrow \text{MT systems} & \Rightarrow \text{i love you .} \\
\text{elle est petite .} & \Rightarrow \text{??} & \Rightarrow \text{she is small .} \\
\end{array}
\]
Noise Reduction [Okita, ACL09SRW]

(Training Phase) We let our MT systems learn by training set.

<table>
<thead>
<tr>
<th>c’est la vie</th>
<th>MT Systems</th>
<th>that is life</th>
</tr>
</thead>
<tbody>
<tr>
<td>je t’aime</td>
<td>⇒ Noisy Channel</td>
<td>i love you</td>
</tr>
<tr>
<td>elle est petite</td>
<td>⇒</td>
<td>she is small</td>
</tr>
</tbody>
</table>

(Training Phase) We train our multiclass classifier by training set.

<table>
<thead>
<tr>
<th>c’est la vie</th>
<th>multiclass classifier</th>
<th>blue</th>
</tr>
</thead>
<tbody>
<tr>
<td>je t’aime</td>
<td>⇒</td>
<td>red</td>
</tr>
<tr>
<td>elle est petite</td>
<td>⇒</td>
<td>purple</td>
</tr>
</tbody>
</table>

(Test Phase) We can expect if we translate our training set our MT systems learn most of them in good faith (considering a bit about generalisation error).

<table>
<thead>
<tr>
<th>c’est la vie</th>
<th>above trained</th>
<th>that is life</th>
</tr>
</thead>
<tbody>
<tr>
<td>je t’aime</td>
<td>⇒ MT systems</td>
<td>i love you</td>
</tr>
<tr>
<td>elle est petite</td>
<td>⇒</td>
<td>she is small</td>
</tr>
</tbody>
</table>
(Training Phase) We let our MT systems learn by training set.

- *c'est la vie.* → MT Systems → *that is life.*
- *je t'aime.* → *Noisy Channel* → *i love you.*
- *elle est petite.* → *she is small.*

(Training Phase) We train our multiclass classifier by training set.

- *c'est la vie.* → multiclass classifier → *blue*
- *je t'aime.* → *red*
- *elle est petite.* → *purple*

(Training Phase) We expect that multiclass classifier outputs similar color in our training set.

- *c'est la vie.* → multiclass classifier → *blue*
- *je t'aime.* → *red*
- *elle est petite.* → *purple*
Noise Reduction

► (Training Phase) 总共 是 多少？ (zong gong shi duo shao) → what does that come to ?

<table>
<thead>
<tr>
<th>TOTAL</th>
<th>(0)</th>
<th>(0)</th>
<th>0.142857</th>
<th>0.0543478</th>
<th>0.125</th>
<th>0.0862069</th>
<th>2.718</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL</td>
<td>(0)</td>
<td>(0)</td>
<td>0.233578</td>
<td>0.0543478</td>
<td>0.125</td>
<td>0.0862069</td>
<td>2.718</td>
</tr>
<tr>
<td>TOTAL</td>
<td>(0)</td>
<td>(0)</td>
<td>0.275862</td>
<td>0.0883644</td>
<td>0.00298954</td>
<td>0.000933415</td>
<td>2.718</td>
</tr>
<tr>
<td>TOTAL</td>
<td>(0)</td>
<td>(0)</td>
<td>0.0480072</td>
<td>0.109269</td>
<td>0.254808</td>
<td>0.157088</td>
<td>2.718</td>
</tr>
<tr>
<td>TOTAL</td>
<td>(0)</td>
<td>(0)</td>
<td>0.447633</td>
<td>0.620852</td>
<td>0.931172</td>
<td>0.967281</td>
<td>2.718</td>
</tr>
</tbody>
</table>

cause word alignment problem.

► (Test Phrase) 总共 是 多少？ → what ’s the total ?
Why is this noise reduction for word alignment?

- ‘word alignment + phrase extraction heuristics’ is a compromise to solve a phrase alignment task [Marcu and Wong, 2002],
- By definition, a word alignment task will not capture the NtoM mapping objects such as paraphrases, multi-word expressions and non-literal translations.

(Heuristics in outlier detection literature): If we collect ’good points’, we may be able to avoid outliers [Forsyth and Ponce, 2003].
Noise Reduction: Algorithm and Results

Algorithm 1 Good Points Algorithm

Step 1: Train word-based SMT, and translate all the sentences to get n-best lists.
Algorithm 2 Good Points Algorithm

Step 1: Train word-based SMT, and translate all the sentences to get n-best lists.

Step 2: Obtain the sentence-based cumulative $X$-gram ($X \in \{1, \cdots, 4\}$) score $S_{WB,X}$. 
Algorithm 3 Good Points Algorithm

**Step 1**: Train word-based SMT, and translate all the sentences to get n-best lists.

**Step 2**: Obtain the sentence-based cumulative $X$-gram ($X \in \{1, \cdots, 4\}$) score $S_{WB,X}$.

**Step 3**: Train PB-SMT, and translate all training sentences to get n-best lists.
Algorithm 4 Good Points Algorithm

**Step 1**: Train word-based SMT, and translate all the sentences to get n-best lists.

**Step 2**: Obtain the sentence-based cumulative $X$-gram ($X \in \{1, \cdots, 4\}$) score $S_{WB,X}$.

**Step 3**: Train PB-SMT, and translate all training sentences to get n-best lists.

**Step 4**: Obtain the sentence-based cumulative $X$-gram ($X \in \{1, \cdots, 4\}$) score $S_{PB,X}$. 
Algorithm 5 Good Points Algorithm

Step 1: Train word-based SMT, and translate all the sentences to get n-best lists.

Step 2: Obtain the sentence-based cumulative $X$-gram ($X \in \{1, \cdots, 4\}$) score $S_{WB,X}$.

Step 3: Train PB-SMT, and translate all training sentences to get n-best lists.

Step 4: Obtain the sentence-based cumulative $X$-gram ($X \in \{1, \cdots, 4\}$) score $S_{PB,X}$.

Step 5: Remove sentence pairs where $S_{WB,2} = 0$ and $S_{PB,2} = 0$. 
Algorithm 6 Good Points Algorithm

Step 1: Train word-based SMT, and translate all the sentences to get n-best lists.
Step 2: Obtain the sentence-based cumulative $X$-gram ($X \in \{1, \cdots, 4\}$) score $S_{WB,X}$.
Step 3: Train PB-SMT, and translate all training sentences to get n-best lists.
Step 4: Obtain the sentence-based cumulative $X$-gram ($X \in \{1, \cdots, 4\}$) score $S_{PB,X}$.
Step 5: Remove sentence pairs where $S_{WB,2} = 0$ and $S_{PB,2} = 0$.
Step 6: The remaining sentence pairs after removal in Step 5 are used to train the final PB-SMT systems.
Noise Reduction: Example of Detected Outliers

<table>
<thead>
<tr>
<th>总共 是 多少？</th>
<th>what does that come to？</th>
</tr>
</thead>
<tbody>
<tr>
<td>服务台 的 号码 是 多少？</td>
<td>what number should i dial for information？</td>
</tr>
<tr>
<td>它 在 星期几 开？</td>
<td>what days of the week does it take place？</td>
</tr>
<tr>
<td>这 是 钥匙。</td>
<td>the keys go here 。</td>
</tr>
<tr>
<td>一点 过 五 分。</td>
<td>it ’s five after one 。</td>
</tr>
</tbody>
</table>

Table: Outliers for BTEC Chinese–English task by Good Point algorithm.
System Combination

- Minimum Bayes-Risk-Confusion Network (MBR-CN) framework [Kumar and Byrne, 2004][Du et al., WMT2008] (Work very well in our recent MT-eval campaigns).

\[
\hat{e}_i = \arg \min_{e_i} \sum_{j=1}^{N} \{1 - \text{BLEU}(e_j, e_i)\}
\]

- Confusion Network:
  - (backbone) output of MBR decoder, (other elements) other hypotheses are aligned by TER.
  - Features: 1) word posterior probability, 2) trigram and 4-gram target language model, 3) word length penalty, and 4) NULL word length penalty.
  - MERT is used to tune the weights of CN.
Case and Punctuation Restoration (1)

- Translation-based approach [Hassan et al., 07] (best system for Arabic-EN human evaluation)
  - Treating case / punctuation restoration as a translation task
    - source: lower-cased sentences
    - target: true-cased sentences (case restoration), text with punctuation (punctuation restoration)
Case and Punctuation Restoration (2)

- Punctuation restoration
  - Combination of translation-based approach and LM-based approach (by majority voting); If no solution can be found using this approach, we choose the first hypothesis proposed by the LM-based method).

- Case restoration
  - Translation-based approach.
Table Of Contents

1. MaTrEx
2. Four Techniques Investigated
3. Experiments
4. Conclusions
Experimental Setup

- **Baseline System**: Standard log-linear PB-SMT system
  - word alignment by Giza++,
  - phrase extraction heuristics,
  - MERT (optimised by Bleu),
  - 5-gram language model with Kneser-Ney smoothing by SRILM, and
  - Moses [Koehn et al., 07] for decoding.

- **System Combination**
  - Joshua (Hierarchical Phrase-Based system) [Li et al., 09],
  - SAMT (Syntax-Based SMT system) [Zollmann et al., 06].

- **Additional Tools**
  - LDC segmenter (Additional Chinese segmentation for word lattice),
  - Berkeley parser (required for Syntax-Based SMT systems),
### Notation

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDF</td>
<td>grow-diag-final</td>
</tr>
<tr>
<td>INT</td>
<td>intersection</td>
</tr>
<tr>
<td>DS-GDF</td>
<td>noise reduction after grow-diag-final</td>
</tr>
<tr>
<td>Lattice</td>
<td>word lattice</td>
</tr>
<tr>
<td>HPB</td>
<td>hierarchical MT (joshua)</td>
</tr>
<tr>
<td>SAMT</td>
<td>syntax-based MT (SAMT)</td>
</tr>
</tbody>
</table>
### BTEC Chinese–English translation

<table>
<thead>
<tr>
<th></th>
<th>PB-SMT</th>
<th></th>
<th>Lattice</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GDF</td>
<td>INT</td>
<td>DS</td>
<td>GDF</td>
<td>INT</td>
<td>HPB</td>
<td>SAMT</td>
</tr>
<tr>
<td>c/p</td>
<td>.3903</td>
<td>.3856</td>
<td>.3733</td>
<td>.4002</td>
<td>.3672</td>
<td>.3783</td>
<td>.3612</td>
</tr>
<tr>
<td>n c/p</td>
<td>.3808</td>
<td>.3717</td>
<td>.3617</td>
<td>.3811</td>
<td>.3463</td>
<td>.3614</td>
<td>.3466</td>
</tr>
<tr>
<td>OOV</td>
<td>139</td>
<td>90</td>
<td>191</td>
<td>40</td>
<td>6</td>
<td>139</td>
<td>141</td>
</tr>
</tbody>
</table>

**Table:** Performance of single systems and multiple system combination for BTEC Chinese–English translation (BLEU)

- sys combo 5 % increase than GDF.
- OOV
## BTEC Turkish–English translation

<table>
<thead>
<tr>
<th></th>
<th>PB-SMT</th>
<th>Lattice</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GDF</td>
<td>INT</td>
<td>DS</td>
<td>GDF</td>
<td>INT</td>
<td>HPB</td>
<td>SAMT</td>
</tr>
<tr>
<td>c/p</td>
<td>.4831</td>
<td>.4656</td>
<td>.4591</td>
<td>.5233</td>
<td>.5247</td>
<td>.4711</td>
<td>.4708</td>
</tr>
<tr>
<td>n c/p</td>
<td>.4590</td>
<td>.4394</td>
<td>.4390</td>
<td>.5008</td>
<td>.5065</td>
<td>.4455</td>
<td>.4516</td>
</tr>
<tr>
<td>OOV</td>
<td>106</td>
<td>61</td>
<td>106</td>
<td>21</td>
<td>11</td>
<td>88</td>
<td>80</td>
</tr>
</tbody>
</table>

**Table:** Performance of single systems and multiple system combination for BTEC Turkish–English translation (BLEU)

- sys combo 7% increase.
### Table: Performance of single systems and multiple system combination for CHALLENGE Chinese–English translation (BLEU)

<table>
<thead>
<tr>
<th></th>
<th>PB-SMT</th>
<th>Lattice</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GDF</td>
<td>INT</td>
<td>DS</td>
<td>GDF</td>
<td>INT</td>
<td>HPB</td>
<td>SAMT</td>
<td>Combo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>crr c/p</td>
<td>.3169</td>
<td>.3278</td>
<td>.3143</td>
<td>.3436</td>
<td>.3335</td>
<td>.3148</td>
<td>.2978</td>
<td>.3689</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n c/p</td>
<td>.3109</td>
<td>.3262</td>
<td>.3088</td>
<td>.3371</td>
<td>.3310</td>
<td>.3057</td>
<td>.2906</td>
<td>.3673</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OOV</td>
<td>197</td>
<td>76</td>
<td>188</td>
<td>21</td>
<td>0</td>
<td>191</td>
<td>197</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>asr c/p</td>
<td>.2918</td>
<td>.2915</td>
<td>.2913</td>
<td>.2724</td>
<td>.2958</td>
<td>.2869</td>
<td>.2700</td>
<td>.3161</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n c/p</td>
<td>.2789</td>
<td>.2825</td>
<td>.2752</td>
<td>.2660</td>
<td>.2861</td>
<td>.2744</td>
<td>.2536</td>
<td>.3064</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OOV</td>
<td>158</td>
<td>96</td>
<td>153</td>
<td>5</td>
<td>5</td>
<td>157</td>
<td>154</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CHALLENGE Chinese–English translation
### Table: Performance of single systems and multiple system combination for BTEC English–Chinese translation (BLEU)

<table>
<thead>
<tr>
<th></th>
<th>GDF</th>
<th>INT</th>
<th>DS</th>
<th>HPB</th>
<th>SAMT</th>
<th>Combo</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>crr c/p</strong></td>
<td>.3531</td>
<td>.3833</td>
<td>.3547</td>
<td>.3797</td>
<td>.3563</td>
<td>.3725</td>
</tr>
<tr>
<td><strong>n c/p</strong></td>
<td>.3555</td>
<td>.3885</td>
<td>.3570</td>
<td>.3832</td>
<td>.3613</td>
<td>.3757</td>
</tr>
<tr>
<td><strong>OOV</strong></td>
<td>99</td>
<td>32</td>
<td>91</td>
<td>102</td>
<td>101</td>
<td>38</td>
</tr>
<tr>
<td><strong>asr c/p</strong></td>
<td>.2970</td>
<td>.3264</td>
<td>.3138</td>
<td>.3332</td>
<td>.3088</td>
<td>.3273</td>
</tr>
<tr>
<td><strong>n c/p</strong></td>
<td>.2987</td>
<td>.3315</td>
<td>.3154</td>
<td>.3372</td>
<td>.3110</td>
<td>.3306</td>
</tr>
<tr>
<td><strong>OOV</strong></td>
<td>129</td>
<td>64</td>
<td>141</td>
<td>112</td>
<td>120</td>
<td>40</td>
</tr>
</tbody>
</table>

- Sys combo decreases. This problem was investigated further [Du et al., ICASSP submitted].
Translation Example: Notation

1. PB
2. PB-INT
3. HIERO
4. SAMT
5. LATTICE
6. LATTICE-INT
7. DS-GDF
8. COMBO
Translation Example (1)

我想订的房间住两天是十月二十七日和二十八日还有空房吗？

1. I'd like to reserve a room for two nights, on October 二十七日 and 二十八日. Do you have any available rooms?

2. I’d like to reserve a room for two days, October 二十七日 and 二十八日 available?

3. I’d like to reserve a room for two nights, on October 二十七日 and 二十八日. Do you have any available rooms?

4. I’d like to reserve a room for two days, it’s October 二十七日 and 二十八日. Do you have any available rooms?

5. I’d like to reserve a room for the twenty-seventh and twenty-eighth two Japanese yen available?

6. I’d like to reserve a room for two for the twenty-seventh and twenty-eighth of Japanese yen available?

7. I’d like to reserve a room for two days is on October 二十七日 and 二十八日. Do you have any rooms available?

8. I’d like to reserve a room for two for the twenty-seventh and do you have any available rooms?
Translation Example (2)

1. My name is Naoko Suzuki.
2. My name is Naoko Suzuki.
3. My name is Naoko Suzuki.
4. My name is Naoko Suzuki.
5. My name is Naoko Suzuki.
6. My name is Naoko Suzuki.
7. My name is Naoko Suzuki.
8. My name is Naoko Suzuki.
嗯 我 还 一 个 问 题 要 问 。

1. I’d also like to know a problem.
2. Well, I have a question.
3. Ask me, I’d like a question.
4. Yes, and I’d like to ask a problem.
5. I have a problem.
6. I’d like to ask you a question.
7. I’d also like to know a problem.
8. I’d like to have a problem.
TranslationExample (4)

1. Can I use a Master card to pay.
2. I use a Master Card.
3. I use a Master card to pay.
4. I want to pay by MasterCard, please.
5. I use a Master Card.
6. I use a Master Card.
7. I want to pay by MasterCard, please.
8. I want to use a Master Card.

我 用 万事达信用卡 付款 。

1. Can I use a Master card to pay.
2. I use a Master Card.
3. I use a Master card to pay.
4. I want to pay by MasterCard, please.
5. I use a Master Card.
6. I use a Master Card.
7. I want to pay by MasterCard, please.
8. I want to use a Master Card.
Translation Example (5)

是的 没错。

1. Yes, 没错.
2. Yes, 没错.
3. Yes, 没错.
4. Yes, 没错.
5. Yes, that’s correct.
6. Yes, that’s right.
7. Yes 没错.
8. Yes, right.
Translation Example (6)

我 拿 北纬 太 贵 了。

1. I have 北纬 too expensive.
2. I 北纬 too expensive.
3. I have 北纬 too expensive.
4. I have 北纬 too expensive.
5. I have north Montreal too expensive.
6. I Montreal north too expensive.
7. I'd like 北纬 too expensive.
8. I have north too expensive.
System Combination (Problem)

Figure: Performance of the systems on development set and test set
Why Noise Reduction Did Not Work? (1)

- (At first sight) Too much removal of sentences, **10-15 %**, caused the problem (Our experiences for European language pairs, **3-5 %** worked well).

- Close look at parallel corpus:
  - There are indeed a lot of duplicate pairs of sentences (this might cause the similar effect of noise reduction algorithm; removal vs duplication).

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Train Set</th>
<th>Pure Train Set</th>
<th>Redundancies</th>
</tr>
</thead>
<tbody>
<tr>
<td>BT-TR-EN</td>
<td>27,972</td>
<td>26,970</td>
<td>3.0 %</td>
</tr>
<tr>
<td>BT-ZH-EN</td>
<td>47,098</td>
<td>43,657</td>
<td>12.2 %</td>
</tr>
<tr>
<td>CH-ZH-EN</td>
<td>75,231</td>
<td>69,680</td>
<td>4.0 %</td>
</tr>
<tr>
<td>CH-EN-ZH</td>
<td>39,228</td>
<td>38,227</td>
<td>12.0 %</td>
</tr>
</tbody>
</table>

**Table:** Redundancies in Parallel corpus
Sentence duplication algorithm [Okita, CLUKI09].

- motivated by statistics, make the tails of a probability distribution heavier.
- We tuned parameter by trial and error.

**Algorithm 7** Sentence Duplication Algorithm

1. Conditioned on a sentence length pair \((l_e, l_f)\), we count the numbers of them. We calculate the ratio \(r_{i,j}\) of this number over the number of all sentences.
2. If this ratio \(r_{i,j}\) is under the threshold \(X\), we duplicate \(N\) times.
Why Noise Reduction Did Not Work? (3)

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>train set</th>
<th>pure train set</th>
<th>noise reduction</th>
<th>removal</th>
</tr>
</thead>
<tbody>
<tr>
<td>BT-TREN</td>
<td>.4831</td>
<td>.4478</td>
<td>.4611</td>
<td>7.1 %</td>
</tr>
<tr>
<td>BT-ZHEN</td>
<td>.3903</td>
<td>.3750</td>
<td>.3741</td>
<td>10.4 %</td>
</tr>
<tr>
<td>CH-ENZH</td>
<td>.3169</td>
<td>.2847</td>
<td>.3011</td>
<td>10.6 %</td>
</tr>
<tr>
<td>CH-ZHEN</td>
<td>.3531</td>
<td>.3154</td>
<td>–</td>
<td>9.5 %</td>
</tr>
</tbody>
</table>

Table: BLEU score of original / non-redundant train set / noise reduced for non-redundant train set (PB-SMT by GDF setting).

- After applied such algorithm, noise reduction won’t work.
Table Of Contents

1. MaTrEx
2. Four Techniques Investigated
3. Experiments
4. Conclusions
Conclusions

We focus on low-resource scenario by MaTrEx: 4 new techniques.

For the CHALLENGE Chinese–English translation task, our system achieved the top Bleu score among other systems.

Word lattice

- best single system for ZN–EN and TR–EN.
- We show greater benefit for TR–EN (morphologically rich languages).

Noise reduction

- Under 3-12 percents of duplication, our noise reduction may not work (= If it’s intentional, IWSLT organizer has more effective algorithm than ours).

System combination techniques

- For ZN–EN and TR–EN, the best performance is achieved.
- Only for EN–ZH translation, slightly inferior performance.
Thank you.

Acknowledgement:

- This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation at Dublin City University.
- Irish Centre for High-End Computing.
- Kemal Oflazer for providing us the output of morphological analyser.
Reference


Reference

