The NiCT-ATR Speech Translation System for IWSLT 2006 Evaluation

Ruiqiang Zhang, Hirofumi Yamamoto, Michael Paul, Hideo Okuma, Keiji Yasuda, Yves Lepage, Etienne Denoual, Daichi Mochihashi, Andrew Finch, Eiichiro Sumita
Main techniques

• Multiple engines: TATR, HPATR3, EM
• Language model adaptation: topic-based
• Subword-based unknown word (UW) translation
• Pre-processing:
  - Word segmentation
  - Sentence splitting
• Post-processing:
  - Higher-order LM for rescoring N-best translation results
  - Punctuation, Capitalization
Multi-engine SMT (Paul, 2005)

• TATR: phrase-based SMT
  - Multiple features:
    • Pr(e|f), Pr(f|e), Lex(e|f), Lex(f|e), LM
• HPATR3: SMT based on syntax-transfer
• EM: exact match based on translation memory
• Selector: multiple set of translation models(TM) and LMs.
Language model adaptation: clustering bilingual training data

1. The number of clusters (topics) is predetermined before clustering.
2. Initialization: assign a class for each pair randomly
3. Reassign the class for each pair
4. Repeat 3 until no entropy reduction is found.
Performance of LM adaptation

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.222</td>
<td>6.68</td>
<td>0.681</td>
<td>0.517</td>
</tr>
<tr>
<td>Adapted</td>
<td>0.236(+1.4)</td>
<td>6.81</td>
<td>0.672</td>
<td>0.505</td>
</tr>
</tbody>
</table>

IWSLT J-E translation

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.212</td>
<td>6.78</td>
<td>0.709</td>
<td>0.513</td>
</tr>
<tr>
<td>Adapted</td>
<td>0.228(+1.6)</td>
<td>6.96</td>
<td>0.698</td>
<td>0.508</td>
</tr>
</tbody>
</table>

IWSLT C-E translation
Achilles – Chinese word segmentation (Zhang, 2006)

1. Dictionary based N-gram word segmentation
   黄 英 春 住 在 北京市

2. Subword-based IOB tagging
   黄/B 英/I 春/I 住/O 在/O 北京/B 市/I

• Combine the two by confidence measure
   黄/B/0.7 英/I/0.7 春/I/0.7 住/O/1.0 在/O/0.9 北京/B/1.0 市/I/0.9

the best balanced point between in vocabulary rate
and out-of-vocabulary rate
Evaluation of Achilles

• The second Sighan bakeoff CWS evaluation
  - State-of-the-art CWS F-SCOREs (higher than the best-known)

• A higher translation performance (+1.1% BLEU) than the LDC default (using NIST 2005 data).
  - A plausible word segmenter.

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST†</th>
<th>WER</th>
<th>PER</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDC default</td>
<td>0.226</td>
<td>7.62</td>
<td>0.895</td>
<td>0.642</td>
<td>0.528</td>
</tr>
<tr>
<td>OURs</td>
<td>0.237 (+1.1)</td>
<td>7.93</td>
<td>0.867</td>
<td>0.614</td>
<td>0.525</td>
</tr>
</tbody>
</table>
Sentence splitting

• Shorter sentences are easier for translation.
• ASR output does not include punctuation.
• Sentence splitting following automatic punctuation:

1. We added punctuation using SRI tools
2. Sentence is split at the place of the added punctuation
3. Sentence segments are translated individually, without using punctuation.
4. Translations results are re-linked in the same order as the input.
Performance of sentence splitting

- Sentence splitting improved CE and JE: BLEU scores (+1)

<table>
<thead>
<tr>
<th></th>
<th>CE spont.</th>
<th>CE read</th>
<th>CE CRR</th>
<th>JE read</th>
<th>JE CRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o splitting</td>
<td>0.1551</td>
<td>0.1756</td>
<td>0.2051</td>
<td>0.1817</td>
<td>0.2023</td>
</tr>
<tr>
<td>w/splitting</td>
<td>0.1591</td>
<td>0.1775</td>
<td>0.206</td>
<td>0.1899</td>
<td>0.2122</td>
</tr>
</tbody>
</table>
Performance of sentence splitting

- Sentence splitting does not work for AE and IE translation: BLEU scores

<table>
<thead>
<tr>
<th></th>
<th>AE read</th>
<th>AE CRR</th>
<th>IE read</th>
<th>IE CRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o splitting</td>
<td>0.2122</td>
<td>0.2365</td>
<td>0.2991</td>
<td>0.3774</td>
</tr>
<tr>
<td>w/splitting</td>
<td>0.2117</td>
<td>0.2384</td>
<td>0.2989</td>
<td>0.3763</td>
</tr>
</tbody>
</table>
Unknown word (UW) translation – subword approach for CE translation

- Unknown words (out of phrase-translation table) cannot be translated by the phrase-based translation engines.
- For Chinese, longer UW split into translatable, shorter subwords. Training a subword translation model by
  - Same training approach as for full word translation model but
  - Use a subword lexicon (size=5000) for word segmentation

- **Effective** for UWs such as named entity, Chinese name, foreign words, and numbers. Translation obtained by combination of the meanings of subwords
- **Useless** if UW’s meaning is unrelated to subwords.
- 0.5% BLEU improvement
Punctuation and Capitalization

- **Punctuation tool**: Use SRI hidden-ngram
- **Capitalization tool**: an in-house CRF-based tagger
  - Capitalization as part-of-speech tagging
    1. AL (All Lowercase)
    2. AU (all uppercase)
    3. IU (initial uppercase)
    4. MX (mix case)
  - Label example:
    "McAdam is the CEO of a British company"
    "mcadam/MX  is/AL the/AL ceo/AU  of/AL a/AL british/IU  company/AL"
System overview

Input

Segmentation
Sentence Split

MT1:
TATR

MT2:
HPATR3

MT3:
EM

Selector
TM*LM

Output (OPEN)

Capitalization
Punctuation

Output (CSTAR)

Train $n$ multiple pairs of TM and LM

parallel text corpus
Tools and Resources

• Morphological analysis
  – Achilles (CWS), BAMA (Arabic), SRI (LM, punctuation)

• Training translation model
  – GIZA++, Pharaoh

• Language data resources:
  – BTEC, LDC

<table>
<thead>
<tr>
<th></th>
<th>Source</th>
<th>English</th>
<th>Source</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>OPEN</td>
<td>39,953(37,559)</td>
<td>39,953(39,633)</td>
<td>342,362</td>
</tr>
<tr>
<td></td>
<td>CSTAR</td>
<td>678,748(399,527)</td>
<td>716,280(358,681)</td>
<td>4,606,373</td>
</tr>
<tr>
<td>JE</td>
<td>OPEN</td>
<td>39,953(37,173)</td>
<td>39,953(39,633)</td>
<td>398,498</td>
</tr>
<tr>
<td></td>
<td>CSTAR</td>
<td>691,711(490,499)</td>
<td>651,558(444,859)</td>
<td>6,795,833</td>
</tr>
<tr>
<td>AE</td>
<td>OPEN</td>
<td>19,972(19,777)</td>
<td>19,972(19,880)</td>
<td>154,279</td>
</tr>
<tr>
<td>IE</td>
<td>OPEN</td>
<td>19,972(19,641)</td>
<td>19,972(19,880)</td>
<td>171,764</td>
</tr>
</tbody>
</table>
Evaluation tracks

• **Open track**
  - ASR spontaneous: CE
  - ASR read translation: CE, JE, AE, IE
  - Correct recognition result: CE, JE, AE, IE

• **CSTAR track**
  - ASR spontaneous: CE
  - ASR read translation: CE, JE
  - Correct recognition result: CE, JE
Translation results: Single engine vs. multiple engine

Table: BLEU score in the CSTAR track (with punctuation and case sensitive)

<table>
<thead>
<tr>
<th></th>
<th>CE spontaneous</th>
<th>CE read</th>
<th>JE read</th>
</tr>
</thead>
<tbody>
<tr>
<td>SELECTOR</td>
<td>0.2008</td>
<td>0.2155</td>
<td>0.2487</td>
</tr>
<tr>
<td>TATR</td>
<td>0.2002</td>
<td>0.2189</td>
<td>0.2463</td>
</tr>
<tr>
<td>HPATR3</td>
<td>0.2177</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Remark: Mixed results made by SELECTOR. Phrase-based TATR gives the best results.
# Overall translation results: BLEU scores

<table>
<thead>
<tr>
<th></th>
<th>OPEN w/ca. punct.</th>
<th>OPEN w/o ca. punct.</th>
<th>CSTAR w/ca. punct.</th>
<th>CSTAR w/o ca. punct.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE spont.</td>
<td>0.1591</td>
<td>0.1615</td>
<td>0.2008</td>
<td>0.2039</td>
</tr>
<tr>
<td>CE read</td>
<td>0.1775</td>
<td>0.1772</td>
<td>0.2155</td>
<td>0.2214</td>
</tr>
<tr>
<td>CE CRR</td>
<td>0.206</td>
<td>0.2123</td>
<td>0.2654</td>
<td>0.2751</td>
</tr>
<tr>
<td>JE read</td>
<td>0.1899</td>
<td>0.1832</td>
<td>0.2487</td>
<td>0.2466</td>
</tr>
<tr>
<td>JE CRR</td>
<td>0.2122</td>
<td>0.2077</td>
<td>0.2861</td>
<td>0.2867</td>
</tr>
<tr>
<td>AE read</td>
<td>0.2117</td>
<td>0.2164</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AE CRR</td>
<td>0.2365</td>
<td>0.2463</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE read</td>
<td>0.2989</td>
<td>0.3194</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE CRR</td>
<td>0.3763</td>
<td>0.412</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Result analysis

- BLEU scores: CRR (+4) > read speech (+1.5) > spont. speech
- BLEU scores: IE (+10) > AE (+2) > JE (+2) > CE
- For AE and IE, without ca. and punct. > with ca. and punct.
- For CE and JE, without ca. and punct = with ca. and punct.
- Capitalization and punctuation is important for AE and IE.
Conclusions

• We used some pre-processing approaches.
  - Word segmentation
  - Language model adaptation
  - Sentence splitting
  - Subword-based translation

• Phrase-based translation engine (TATR)