I²R Multi-Pass Machine Translation System for IWSLT-2008

Boxing Chen, Deyi Xiong, Min Zhang, Aiti Aw and Haizhou Li

Department of HLT
Institute for Infocomm Research (I²R), Singapore
Tasks

• Chinese-to-English
  – BTEC task
  – Challenge task

• Chinese-to-English-to-Spanish PIVOT task
  – Joint effort with UPC-TALP
  – Will be reported by the co-worker of UPC
Outline

• Multi-pass MT System
  – System Architecture
  – 1\textsuperscript{st} pass: preprocessing
  – 2\textsuperscript{nd} pass: decoding
  – 3\textsuperscript{rd} pass: rescoring
  – 4\textsuperscript{th} pass: system combination

• Experiments and results

• Conclusion
Preprocessing

1st pass: preprocessing
Preprocessing

- Preprocessing
  - Tools: LDC-SEG (L), ICTCLAS (I)

<table>
<thead>
<tr>
<th></th>
<th>Preprocessing 1</th>
<th>Preprocessing 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ch</td>
<td>en</td>
</tr>
<tr>
<td>Tokenization</td>
<td>L</td>
<td>x</td>
</tr>
<tr>
<td>Txt-to-digit</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Lower-casing</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>
Decoding

2\textsuperscript{nd} Pass: decoding
Decoding

- **Preprocessing Setting**
  - Moses1: preprocessing 1
  - All other 4 systems: preprocessing 2

- **Moses**: (open source)
  - Phrase-based system

- **JosHUA**: (open source)
  - Hierarchical phrase-based MT system

- **Tranyu**: (in-home)
  - BTG-based system
Tranyu

- Adapting BTG to phrasal translation
- CKY-style decoder
- Reordering models
  - MaxEnt-based
  - Features:
    - Boundary words (Xiong et al. 2006)
    - Linguistic annotations (Xiong et al. 2008)
- Systems
  - Tranyu 1: boundary words based reordering model
  - Tranyu 2: boundary words based reordering + linguistically annotated reordering model
Rescoring

3rd Pass: rescoring
### Rescoring

- Rich additional feature functions

#### Moses Features:
- Translation Model
- Reordering model
- Language Model
- Word penalty
- Translation confidence

#### Rescoring Features:
1. Dir/Inv IBM model 1 and 3 score
2. Word-based association score
3. lexicalized word/block reordering probabilities
4. 6-gram target LM
5. 8-gram target word-class based LM
6. source and target length ratio
7. question feature
8. frequency of n-grams in the N-best
9. n-gram post-probabilities
10. sentence length post-probabilities
System Combination

4th Pass: System combination
System Combination: Simple Cascading

Simple Cascading: rescoring on two Moses generated N-best lists
System Combination: N-gram expansion
System Combination: N-gram expansion

- \( n \)-gram expansion generates new hypotheses
  - Collect all the \( n \)-grams from the original N-best
  - Continuously expand the partial hypothesis through the \( n \)-grams.

<table>
<thead>
<tr>
<th>Reference:</th>
<th>my book is in the green basket .</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original entry:</td>
<td>my book is in the green case .</td>
</tr>
<tr>
<td></td>
<td>my book is inside the green basket .</td>
</tr>
<tr>
<td>3-grams:</td>
<td>my book is, book is in, is in the, in the green, the green case, is inside the, the green basket ...</td>
</tr>
<tr>
<td>( n )-gram expansion</td>
<td>Partial Hyp: my book is in ( n )-gram: _ _ _ is in the New partial Hyp: my book is in the _ _ _</td>
</tr>
<tr>
<td>New Hyp:</td>
<td>my book is in the green basket .</td>
</tr>
</tbody>
</table>
System Combination: Weighted Voting

Weighted Voting:
Sentence-level System combination
Experiments

• Training data
  – Bilingual Training data:
    • BTEC supplied data (20K sentences)
    • HIT-corpus (132K sentences)
    • Olympic-corpus (54K sentences)
    • PKU-corpus (200K sentences)
    • Total: 399K sentence-pairs, 5.2M target words
  – Dev data: 6K sentences
  – Additional target data: Tanaka corpus
    • 155K sentence-pairs, 1.4M target running words
## Effect of additional data

- **BLEU% score on dev sets**

<table>
<thead>
<tr>
<th></th>
<th>CSTAR03 (BTEC)</th>
<th>DEV08 (Challenge)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplied data</td>
<td>40.96</td>
<td>36.12</td>
</tr>
<tr>
<td>+dev. data</td>
<td>45.76</td>
<td>42.29</td>
</tr>
<tr>
<td>+addi. Data</td>
<td>50.98</td>
<td>44.92</td>
</tr>
<tr>
<td>All data</td>
<td>52.28</td>
<td>46.45</td>
</tr>
</tbody>
</table>
Results (Base & Resc): BTEC task

- Performances of five systems are very similar:
- Baseline: JosHUa > Moses2 > Tranyu2 > Moses1 > Tranyu1
- Rescoring: Moses2 = JosHUa > Tranyu2 > Moses1 > Tranyu1
- Rescoring improved about 1.6-2 BLEU-score for all systems
Results (Base & Resc): Challenge Task

- Performances of five systems are also similar:
- Baseline: JosHUa > Tranyu2 > Tranyu1 > Moses2 > Moses1
- Rescoring: Tranyu2 > JosHUa > Moses2 > Moses1 > Tranyu1
- Rescoring improved about 2-3 BLEU-score for all systems
Results: 1st stage System Combination

Simple Cascading

Advantages: More features (include local feat. used in decoding)

Disadvantages: Less distinct hypotheses

N-gram expansion

Advantages: More distinct hypotheses

Disadvantages: Less features (no local features)
Results: SC vs NE

- N-gram expansion (NE) outperformed simple cascading (SC)
- Compared with the best single system:
  - Simple cascading obtained about 0.9/1.3 BLEU-score
  - N-gram expansion obtained about 1.3/2.3 BLEU-score
Results: weighted voting

- Simple voting: weights of all systems are set to 1.
- Compared with the results of n-gram expansion:
  - Simple voting obtained about 0.8/0.6 BLEU-score
  - Weighted voting obtained about 1.0/1.3 BLEU-score
**Official score**

- Chinese-to-English BTEC and Challenge tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>ASR BLEU (%)</th>
<th>NIST</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTEC task</td>
<td>43.57</td>
<td>6.87</td>
<td>0.6017</td>
</tr>
<tr>
<td>CRR</td>
<td>49.26</td>
<td>7.65</td>
<td>0.6446</td>
</tr>
<tr>
<td>Challenge task</td>
<td>39.38</td>
<td>5.96</td>
<td>0.6142</td>
</tr>
<tr>
<td>CRR</td>
<td>46.89</td>
<td>6.66</td>
<td>0.6560</td>
</tr>
</tbody>
</table>
Conclusion and Discussion

• Conclusion
  – Multi-decoder to produce N-best lists
  – Rich additional feature functions to do rescoring
  – $n$-gram expansion to generate new hypotheses
  – Two-stage system combination

• Comments
  – As one reviewer pointed out: *the 5 systems are all phrase-based system, so the N-best lists are quite similar. It could not provide enough space for system combination. This may be the primary reason that its evaluation result is not too outstanding.*
Thanks for your attention!

Any questions?