I²R Chinese-English Translation System for IWSLT-2007

Boxing Chen, Jun Sun, Hongfei Jiang, Min Zhang, Ai Ti Aw

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

• Motivation
• Multi-pass approach
  – 1st pass: decoding
  – 2nd pass: regeneration
  – 3rd pass: rescoring
• Experiments and results
• Conclusion
Motivation

• A two-pass SMT system’s performance could be improved from two aspects:
  – Scoring models
  – N-best Hypotheses

• Rescoring focus on improving the scoring models

• We try to improve the N-best hypotheses through an additional pass: regeneration and system combination
Multi-pass Approach

- **1st Pass**
  - Decoding
  - Log-linear model
  - Multi decoders

- **2nd Pass**
  - n-gram expansion
  - System combination

- **3rd Pass**
  - Rescoring
  - Log-linear model
  - Additional features
1st Pass: Decoding

• 3 systems
  – Sys1: preprocessing setting 1 + Moses decoder
  – Sys2: preprocessing setting 2 + Moses decoder
  – Sys3: preprocessing setting 2 + STSG decoder
1st Pass: Syntax-based decoder

- STSG: Synchronous Tree Substitution Grammar
- A rule is a pair of elementary tree (PET) with alignment information.
  - PET is defined as a Triple $<\xi_s, \xi_t, A>$
    - $\xi_s$ and $\xi_t$ are source/target elementary tree
    - A is the alignments between leaf nodes of two elementary trees
- Two major benefits:
  - Possible to explicitly model the target syntax
  - Allow Multi-level global structure distortion
1\textsuperscript{st} Pass: STSG Modelling
2\textsuperscript{nd} pass: 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>
</tbody>
</table>

| 3-grams: | my book is, book is in, is in the, in the green, the green case, is inside the, the green basket ... |

<table>
<thead>
<tr>
<th>n-gram expansion</th>
<th>Partial Hyp:</th>
<th>my book is in</th>
</tr>
</thead>
<tbody>
<tr>
<td>n-gram:</td>
<td>______ is in the</td>
<td></td>
</tr>
<tr>
<td>New partial Hyp:</td>
<td>my book is in the</td>
<td></td>
</tr>
</tbody>
</table>

| New Hyp: | my book is in the green basket . |
2\textsuperscript{nd} Pass: System Combination

- System Combination
  - Hypotheses are simply joined
  - Duplicate hypotheses are removed
3rd Pass: Rescoring

- Rich additional feature functions (Chen et al., 2006)

**Moses Features:**
- Translation Model
- Reordering model
- Language Model
- Word penalty

**Translation confidence**

**Rescoring Features:**
1) Dir/Inv IBM model 1 and 3 score
2) CLA 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
Experiments: training data

• Task: Chinese-English Open data track
• Bilingual Training data: BTEC+HIT-corpus
  – Sys1 and Sys2:  
    • 400K sentence-pairs
    • 4.5M target words
  – Sys3:  
    • 90K sentence-pairs
    • 1.0M target words

• Additional target data: Tanaka corpus
  – 155K sentence-pairs, 1.4M target running words
Experiments: preprocessing

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

<table>
<thead>
<tr>
<th></th>
<th>Sys1</th>
<th>Sys2</th>
<th>Sys3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ch</td>
<td>en</td>
<td>ch</td>
</tr>
<tr>
<td>Tokenization</td>
<td>L x</td>
<td>I x</td>
<td>I x</td>
</tr>
<tr>
<td>Parsing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Txt-to-digit</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Lower-casing</td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>
Experiments: setting

- Two series of experiments:
  - DEV: dev1, TEST: dev2, dev3
  - DEV: dev4, TEST: dev5

- 6 types of MT outputs:
  - Sys1/2/3: 3 baselines
  - Resc1: rescoring on Sys1 N-best list
  - Resc2: rescoring on Sys1+Sys2 N-best lists
  - Comb: final translation output with n-gram expansion, system combination and rescoring incorporated
Results: Baseline

BLEU score of dev1/2/3 baseline

BLEU score of dev4/5 baseline
Results: Resc1/2 vs. Comb

Resc1/2:
Advantages: More features (include local feat. used in decoding)
Disadvantages: Less hypotheses

Comb:
Advantages: More hypotheses
Disadvantages: Less features (no local features)
Results: Resc1/2 vs. Comb

BLEU score of dev1/2/3

<table>
<thead>
<tr>
<th>BLEU (%)</th>
<th>Resc1</th>
<th>Resc2</th>
<th>Comb</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev1</td>
<td>55</td>
<td>57</td>
<td>59</td>
</tr>
<tr>
<td>dev2</td>
<td>57</td>
<td>59</td>
<td>61</td>
</tr>
<tr>
<td>dev3</td>
<td>59</td>
<td>61</td>
<td>63</td>
</tr>
</tbody>
</table>

BLEU score of dev4/5

<table>
<thead>
<tr>
<th>BLEU (%)</th>
<th>Resc1</th>
<th>Resc2</th>
<th>Comb</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev4</td>
<td>32</td>
<td>34</td>
<td>36</td>
</tr>
<tr>
<td>dev5</td>
<td>30</td>
<td>32</td>
<td>34</td>
</tr>
</tbody>
</table>
Results: Analysis

• Average length and relative improvements on BLEU (Resc2 vs. Comb)

<table>
<thead>
<tr>
<th></th>
<th>Dev1</th>
<th>Dev2</th>
<th>Dev3</th>
<th>Dev4</th>
<th>Dev5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>6.7</td>
<td>7.0</td>
<td>7.5</td>
<td>12.1</td>
<td>12.6</td>
</tr>
<tr>
<td>Δ</td>
<td>-1.5</td>
<td>-0.2</td>
<td>0.8</td>
<td>6.3</td>
<td>5.3</td>
</tr>
</tbody>
</table>

• Number of new generated hypotheses in Comb (about 500 sentences for each dev set.)

<table>
<thead>
<tr>
<th></th>
<th>Dev1</th>
<th>Dev2</th>
<th>Dev3</th>
<th>Dev4</th>
<th>Dev5</th>
</tr>
</thead>
<tbody>
<tr>
<td>#new hypo</td>
<td>29</td>
<td>18</td>
<td>12</td>
<td>59</td>
<td>74</td>
</tr>
</tbody>
</table>

• n-gram expansion benefits longer sentences more than short sentences. Because it permits long distance word movements through a low-order LM (e.g. a bi-gram LM).
Results: test set

- Test set are more similar to dev1 than other dev sets:
  - average length 6.5 (test) vs. 6.7(dev1)
- On dev1: “Resc2” produces better BLEU score than “Comb”

<table>
<thead>
<tr>
<th></th>
<th>Official submission</th>
<th>Only BTEC data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU(%)</td>
<td>Rank</td>
</tr>
<tr>
<td>Run1 (Resc2)</td>
<td>40.77</td>
<td>1</td>
</tr>
<tr>
<td>Run2 (Comb)</td>
<td>39.42</td>
<td>2</td>
</tr>
</tbody>
</table>
Conclusion

• Multi-pass system
  – Multi-decoder to produce N-best lists
  – n-gram expansion to generate new hypotheses
  – Rich additional feature functions to do rescoring

• Rescoring gives significant improvements

• n-gram expansion and system combination give consistent improvement on longer sentences
Thanks for your attention!
Any questions?