Using Word Posterior in Lattice Translation

Vicente Alabau

Institut Tecnològic d'Informàtica

e-mail: valabau@iti.upv.es

October 16, 2007
Index

- Motivation
- Word Posterior Probabilities
- Translation System
- Results
- Conclusions and Future Work
Motivation - Common approaches

• Serial approach:
  – + simple and fast - propagates errors from ASR

• Semi-coupled approach:
  – n-best: + simple - redundancy, time-consuming
  – lattice: + full searched space - time-consuming
  – confusion network: + simplified lattice, efficient - loss of grammar

• Integrated approach:
  – + theoretically promising - bad performance on non-simple corpora
Word Posterior Probabilities

• Motivation
  – One should maximize word posterior probabilities to minimize WER (Mangu00)
  – Confusion networks (Bertoldi05):
    * word posterior probabilities
    * lattice simplification

• Our approach
  – Word posterior probabilities over a lattice
  – Take advantage of techniques in confidence measures (Sanchis04)
Word Posterior Probabilities: Forward-Backward

- being \( w \) the hypothesized word, \( s \) the start node and \( e \) the end node:

\[
P([w, s, e] \mid \vec{x}_1^T) = \frac{1}{P(\vec{x}_1^T)} \sum_{f_1^J \in G : \exists [w', s', e'] : w' = w, s' = s, e' = e} P(f_1^J, \vec{x}_1^T)
\]  

(1)
Word Posterior Probabilities

- maximum of the frame time posterior probability (Wessel01)

\[
P_t(w \mid \vec{x}_1^T) = \sum_{t \in [s', e']} P([w, s', e'] \mid \vec{x}_1^T)
\]  

\[
P([w, s, e] \mid \vec{x}_1^T) = \max_{s \leq t \leq e} P_t(w \mid \vec{x}_1^T)
\]
Translation System

- Log-linear model:
  - Word posterior probabilities
  - GIATI:
    * Joint probability model
    * N-grams of bilingual pairs
    * 5-gram (w/o cutting off)
    * integrated lattice search
    * monotonous search
  - Output word penalty
  - Output language model (5-gram)
Translation System

- Reordering:
  - Serial, 1BEST approach
  - Monotonization of the output
  - Translate with moses from monotonized to regular word order
  - Models: reordering table and output language model
  - Monotonous search
Preprocess and postprocess

• Preprocess:
  – Case and punctuation were removed from training
  – Sentence splitting at sentence boundaries (.?!)
  – Lattice pruning

• Postprocess:
  – Punctuation and case restoration: IWSLT06 method using SRILM
  – Capitalization after punctuation marks
System architecture
## Corpus statistics

<table>
<thead>
<tr>
<th></th>
<th>Sentences</th>
<th>Running words</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train</strong></td>
<td>19971</td>
<td>172(k)</td>
<td>10,152 (10,152)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>189(k)</td>
<td>7,165 (7,165)</td>
</tr>
<tr>
<td><strong>Dev4</strong></td>
<td>489</td>
<td>4,831</td>
<td>224 (224)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6,848</td>
<td>208 (208)</td>
</tr>
<tr>
<td><strong>Dev5a</strong></td>
<td>500</td>
<td>5,607</td>
<td>296 (296)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7,491</td>
<td>264 (264)</td>
</tr>
<tr>
<td><strong>Dev5b</strong></td>
<td>996</td>
<td>8,487</td>
<td>591 (591)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11,968</td>
<td>611 (611)</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td>724</td>
<td>6,420</td>
<td>542 (542)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9,054</td>
<td>439 (439)</td>
</tr>
</tbody>
</table>
Effect of adding features to the baseline model

- Primary run: 16.13 BLEU

<table>
<thead>
<tr>
<th></th>
<th>dev4</th>
<th></th>
<th>dev5a</th>
<th></th>
<th>dev5b</th>
<th></th>
<th>test</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>NIST</td>
<td>BLEU</td>
<td>NIST</td>
<td>BLEU</td>
<td>NIST</td>
<td>BLEU</td>
<td>NIST</td>
</tr>
<tr>
<td>baseline</td>
<td>36.29</td>
<td>7.59</td>
<td>31.96</td>
<td>7.06</td>
<td>12.53</td>
<td>4.02</td>
<td>22.80</td>
<td>5.49</td>
</tr>
<tr>
<td>+WP</td>
<td>37.45</td>
<td>7.35</td>
<td>32.55</td>
<td>6.82</td>
<td>14.07</td>
<td>3.77</td>
<td>19.56</td>
<td>5.06</td>
</tr>
<tr>
<td>+OL</td>
<td>37.06</td>
<td>7.42</td>
<td>32.55</td>
<td>6.91</td>
<td>12.37</td>
<td>3.82</td>
<td>22.32</td>
<td>5.25</td>
</tr>
<tr>
<td>+WP+OL</td>
<td>38.19</td>
<td>7.20</td>
<td>32.67</td>
<td>6.66</td>
<td>13.44</td>
<td>4.20</td>
<td>21.83</td>
<td>5.57</td>
</tr>
<tr>
<td>+RM</td>
<td>37.53</td>
<td>7.95</td>
<td>32.74</td>
<td>7.41</td>
<td>13.94</td>
<td>4.30</td>
<td>23.92</td>
<td>5.79</td>
</tr>
<tr>
<td>+WP+OL+RM</td>
<td>38.98</td>
<td>7.81</td>
<td>32.86</td>
<td>7.18</td>
<td>14.34</td>
<td>4.37</td>
<td>23.22</td>
<td>5.86</td>
</tr>
</tbody>
</table>

- **WP**, output word insertion penalty
- **OL**, output language model
- **RM**, reordering model
Effect of adding dev corpus to the training corpus

- Primary run: 16.13 BLEU

<table>
<thead>
<tr>
<th></th>
<th>w/o dev</th>
<th></th>
<th>with dev</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>NIST</td>
<td>BLEU</td>
<td>NIST</td>
</tr>
<tr>
<td>baseline</td>
<td>22.80</td>
<td>5.49</td>
<td>31.29</td>
<td>6.66</td>
</tr>
<tr>
<td>+WP</td>
<td>22.09</td>
<td>5.56</td>
<td>12.16</td>
<td>2.97</td>
</tr>
<tr>
<td>+OL</td>
<td>22.79</td>
<td>5.52</td>
<td>30.83</td>
<td>6.64</td>
</tr>
<tr>
<td>+WP+OL</td>
<td>21.79</td>
<td>5.56</td>
<td>11.89</td>
<td>2.91</td>
</tr>
<tr>
<td>+RM</td>
<td>23.46</td>
<td>5.74</td>
<td>32.28</td>
<td>6.95</td>
</tr>
<tr>
<td>+WP+OL+RM</td>
<td>23.22</td>
<td>5.86</td>
<td>31.21</td>
<td>6.77</td>
</tr>
</tbody>
</table>

- **WP**, output word insertion penalty
- **OL**, output language model
- **RM**, reordering model
### Results for different input conditions

<table>
<thead>
<tr>
<th></th>
<th>dev4</th>
<th></th>
<th>dev5a</th>
<th></th>
<th>dev5b</th>
<th></th>
<th>test</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>NIST</td>
<td>BLEU</td>
<td>NIST</td>
<td>BLEU</td>
<td>NIST</td>
<td>BLEU</td>
<td>NIST</td>
</tr>
<tr>
<td>1BEST</td>
<td>33.53</td>
<td>6.92</td>
<td>26.97</td>
<td>6.12</td>
<td>13.21</td>
<td>4.19</td>
<td>21.50</td>
<td>5.56</td>
</tr>
<tr>
<td>LAT</td>
<td>33.69</td>
<td>6.95</td>
<td>27.24</td>
<td>6.14</td>
<td>13.35</td>
<td>4.16</td>
<td>18.71</td>
<td>5.22</td>
</tr>
<tr>
<td>GER</td>
<td>34.11</td>
<td>7.02</td>
<td>27.49</td>
<td>6.18</td>
<td>13.90</td>
<td>4.29</td>
<td>22.64</td>
<td>5.77</td>
</tr>
<tr>
<td>CLEAN</td>
<td>38.98</td>
<td>7.81</td>
<td>32.86</td>
<td>7.18</td>
<td>14.34</td>
<td>4.37</td>
<td>23.22</td>
<td>5.86</td>
</tr>
</tbody>
</table>

- **LAT**, lattice with word posterior probabilities
- **GER**, using the sentence from the lattice with less word error rate
Conclusions

- Word Posterior approach
  - Results not conclusive
  - Small differences between 1BEST and CLEAN scores
  - Some improvements were achieved
  - Needs work on pruning

- Adding devset to training matters
Future Work

- Comparison with n-best, confidence measures, lattice with acoustic scores
- Add additional state-of-the-art confidence features
- Add translation features
- Features based on multiple lattices
- Lattice reduction
Thank you for your attention!

Vicente Alabau

valabau@dsic.upv.es
References


