Ngram-based Vs Phrase-based SMT

Josep M. Crego
Marta R. Costa-jussà
José B. Mariño
José A. R. Fonollosa
Outline

- Introduction
  - Previous work

- Ngram-based and Phrase-based TM
  - Tuple/Phrase extraction & modeling

- SMT system
  - Additional feature models & decoding

- Comparison
  - Accuracy & efficiency

- Conclusions & Future
Introduction

- Previous work...
  - Spanish-English
  - BM + TM (noisy channel)
  - monotone decoding
  - small data corpus (verbmobil)

J.M. Crego, A. de Gispert, J.B. Mariño
*Finite-state-based and Phrase-based SMT*
ICSLP’04 (Jeju, Korea)

comparable results !!
Introduction

- Current SMT systems...
  - context within the TM
  - log-linear combination of feature models
  - reordering capabilities

\[
\tilde{t} = \max_{i_i'} \left\{ \sum_m \lambda_m h_m(s, t) \right\}
\]
Outline

- Introduction
  - Previous work

- Ngram-based and Phrase-based TM
  - Tuple/Phrase extraction & modeling

- SMT system
  - Additional features & decoding

- Comparison
  - Accuracy & efficiency

- Conclusions & Future
Ngram-based and Phrase-based TM

\[ h_{BM}(s, t) = \log \prod_{k=1}^{K} p(u_k | \ldots, u_{k-2}, u_{k-1}) \quad \text{where} \quad u_i = \tilde{s}_i \# \tilde{t}_i \]

\[ h_{BM}(s, t) = \log \prod_{k=1}^{K} p(\tilde{s}_k | \tilde{t}_k) \quad \text{where} \quad p(\tilde{s}_k | \tilde{t}_k) = \frac{N(\tilde{s}_k | \tilde{t}_k)}{N(\tilde{t}_k)} \]
Ngram-based TM

- Regular tuples
  - NULL where is the nearest train station?
  - ¿dónde está la estación de tren más cercana?

- Unfolded tuples
  - NULL where is the train station nearest?
  - ¿dónde está la estación de tren más cercana?

 NULLs solved using IBM1 lexicon probabilities

loss of information !!
sparseness !!
src words reordering !!
decoding needs reordering !!
Outline

- Introduction
  - Previous work

- Ngram-based and Phrase-based TM
  - Tuple/Phrase extraction & modeling

- SMT system
  - Additional features & decoding

- Comparison
  - Accuracy & efficiency

- Conclusions & Future
SMT System

Additional feature models

- Target language model

\[
h_{TM}(s, t) = \log \prod_{n=1}^{l} p(t_n \mid \ldots, t_{n-2}, t_{n-1})
\]

- Word penalty (Compensate longer translations)

\[
h_{WP}(s, t) = I
\]

- Distortion model (distance based)

\[
h_{RM}(s, t) = d_k
\]
SMT System

Additional feature models

- 2 Lexicon models (forward and backward IBM model 1)

\[
h_{IBM1}(s,t) = \log \frac{1}{(I'+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(t_i | s_j)
\]

- Phrase penalty

\[
h_{PP}(s,t) = K \quad \text{number of phrases}
\]

- Posterior phrase model

\[
h_{FR}(s,t) = \log \frac{N(\tilde{t}, \tilde{s})}{N(\tilde{s})}
\]
Decoding (MARIE)

- Beam search with pruning
- Hypotheses are built from left to right
- Stored in several groups containing ordered lists (according to accumulated cost)
- Expanding each group:
  - Histogram pruning: $B$
  - Threshold pruning: $T$
- Expanding each list:
  - Histogram pruning: $b$
  - Threshold pruning: $t$
- Recombination
- Reordering constraints:
  - Max distance: $m$
  - Max reorderings: $j$

SIMPLEX algorithm to tune model weights!!
Outline

- Introduction
  - Previous work

- Ngram-based and Phrase-based TM
  - Tuple/Phrase extraction & modeling

- SMT system
  - Additional features & decoding

- Comparison
  - Accuracy & efficiency

- Conclusions & Future
## Comparison

### Databases

- **EPPS es-en**
  - LARGE data task
  - NO reordering needs

- **BTEC zh-en**
  - SMALL data task
  - Reordering needs

<table>
<thead>
<tr>
<th></th>
<th>EPPS</th>
<th>BTEC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Sents</strong></td>
<td><strong>Sents</strong></td>
</tr>
<tr>
<td>train</td>
<td>1.2M</td>
<td>20k</td>
</tr>
<tr>
<td></td>
<td>34.8M</td>
<td>182.9k</td>
</tr>
<tr>
<td></td>
<td>169k</td>
<td>8.1k</td>
</tr>
<tr>
<td>dev</td>
<td>504</td>
<td>506</td>
</tr>
<tr>
<td></td>
<td>15.4k</td>
<td>3.5k</td>
</tr>
<tr>
<td></td>
<td>2.8k</td>
<td>870</td>
</tr>
<tr>
<td>test</td>
<td>840</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>22.7k</td>
<td>3.7k</td>
</tr>
<tr>
<td></td>
<td>4k</td>
<td>893</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Comparison

Translation units

<table>
<thead>
<tr>
<th>BTEC</th>
<th>Vocabulary</th>
<th>Total</th>
<th>Intersection</th>
</tr>
</thead>
<tbody>
<tr>
<td>phrases</td>
<td>124.4k</td>
<td>281.8k</td>
<td>-</td>
</tr>
<tr>
<td>tuples</td>
<td>31.2k</td>
<td>110.6k</td>
<td>22.8k</td>
</tr>
<tr>
<td>tuples’</td>
<td>298.6k</td>
<td>-</td>
<td>62.3k</td>
</tr>
</tbody>
</table>

Pruned out those exceeding size 3
# Comparison

- **Translation Model N-grams (train and test)**

<table>
<thead>
<tr>
<th></th>
<th>1gr</th>
<th>2gr</th>
<th>3gr</th>
<th>4gr</th>
<th>1gr</th>
<th>2gr</th>
<th>3gr</th>
<th>3gr</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BTEC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>EPPS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PB</strong></td>
<td>59.61</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>7,017.89</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>NB</strong></td>
<td>8,999</td>
<td>23.33</td>
<td>3.42</td>
<td>1.99</td>
<td>335,299</td>
<td>1,426,58</td>
<td>767,82</td>
<td>-</td>
</tr>
<tr>
<td><strong>BTEC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>EPPS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PB</strong></td>
<td>2.51</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>15.61</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>NB</strong></td>
<td>1.65</td>
<td>1.24</td>
<td>28</td>
<td>89</td>
<td>2,988</td>
<td>8.49</td>
<td>9.33</td>
<td>3</td>
</tr>
</tbody>
</table>

*Pruned out those exceeding size 4+7*
Comparison

- Accuracy and efficiency results

<table>
<thead>
<tr>
<th>BTEC</th>
<th>mWER</th>
<th>BLEU</th>
<th>TIME (sec)</th>
<th>SIZE (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>baseline</strong></td>
<td>49.68</td>
<td>35.41</td>
<td>17</td>
<td>1.2</td>
</tr>
<tr>
<td><strong>Baseline+IBM1</strong></td>
<td>48.42</td>
<td>35.75</td>
<td>21</td>
<td>1.4</td>
</tr>
<tr>
<td><strong>Baseline+IBM1+reord</strong></td>
<td>45.30</td>
<td>41.66</td>
<td>225</td>
<td>1.6</td>
</tr>
<tr>
<td><strong>Baseline</strong></td>
<td>50.02</td>
<td>36.32</td>
<td>23</td>
<td>2.4</td>
</tr>
<tr>
<td><strong>Baseline+FR</strong></td>
<td>49.57</td>
<td>37.02</td>
<td>28</td>
<td>2.8</td>
</tr>
<tr>
<td><strong>Baseline+FR+IBM1+reord</strong></td>
<td>48.60</td>
<td>39.65</td>
<td>438</td>
<td>3.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EPPS</th>
<th>mWER</th>
<th>BLEU</th>
<th>TIME (sec)</th>
<th>SIZE (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>baseline</strong></td>
<td>39.61</td>
<td>48.49</td>
<td>641</td>
<td>580</td>
</tr>
<tr>
<td><strong>Baseline+IBM1</strong></td>
<td>34.86</td>
<td>54.38</td>
<td>801</td>
<td>600</td>
</tr>
<tr>
<td><strong>baseline</strong></td>
<td>39.35</td>
<td>48.84</td>
<td>900</td>
<td>1,180</td>
</tr>
<tr>
<td><strong>Baseline+FR+IBM1</strong></td>
<td>35.10</td>
<td>54.19</td>
<td>1084</td>
<td>1,640</td>
</tr>
</tbody>
</table>

Pentium IV 3.06GHz 4Gb RAM

\[ \text{TIME}_{\text{NB}} \sim \text{TIME}_{\text{PB}} \cdot 0.7 \]

\[ \text{SIZE}_{\text{NB}} \sim \text{SIZE}_{\text{PB}} \cdot 0.5 \]

BLEU and mWER used in IWSLT’04
Conclusions & Future

- Comparison of state-of-the-art SMT systems
- Main differences found on translation modeling of context
- Fair comparison:
  - Corpora
  - Word-to-word alignment
  - Decoder
  - Additional models
Conclusions & Future

- About results
  - Similar translation accuracy
  - How translation units are extracted and scored.
  - Which additional models are used.
  - Opposite accuracy results in IWSLT’05 !!!
  - NB improves PB in efficiency (memory size and computation time)
  - Smaller vocabulary of translation units
Conclusions & Future

- Further Research:
  - ASR/SMT coupling
  - Rescoring
  - Unfolded units refinement
  - Reordered search (modeling)
...thanks

marie decoder is free available at:
(binary and source code)

http://gps-tsc.upc.es/veu/soft/soft/marie