Phrase-Based Statistical Machine Translation with Pivot Languages

N. Bertoldi, †M. Barbaiani, M. Federico, R. Cattoni

FBK, Trento - Italy
† Rovira i Virgili University, Tarragona - Spain

October 21st, 2008
Pivot Translation

- **Assumptions:**
  - no parallel data between source language $\mathcal{F}$ and target language $\mathcal{E}$
  - two independent parallel corpora $(\mathcal{F}, \mathcal{G}_F)$ and $(\mathcal{G}_E, \mathcal{E})$
  - two full-fledged MT systems $\mathcal{F} \rightarrow \mathcal{G}$ and $\mathcal{G} \rightarrow \mathcal{E}$

- **Problem:** how to perform translation from $\mathcal{F}$ to $\mathcal{E}$?

- **Approach 1:** Bridging at translation time
  
  source text $\mathcal{F} \rightarrow \mathcal{G}$ pivot text $\mathcal{G} \rightarrow \mathcal{E}$ target text $\mathcal{f} \rightarrow g \rightarrow e$

- **Approach 2:** Bridging at training time
  
  synthetic training data generated by translating with system
  
  $$(\mathcal{F}, \bar{\mathcal{E}}_F) \quad G_F \text{ of } (\mathcal{F}, \mathcal{G}_F) \quad \mathcal{G} \rightarrow \mathcal{E}$$
  $$\bar{\mathcal{F}}_E, \mathcal{E} \quad G_E \text{ of } (\mathcal{G}_E, \mathcal{E}) \quad \mathcal{G} \rightarrow \mathcal{F}$$
Pivot Task description

- BTEC domain data
- Pivot Task of IWSLT 2008: Chinese-English-Spanish

- training data: CE1, CE2, ES1, and CS1 (19K sentences)
- disjoint condition: CE2 and ES1
- overlap condition: CE1 and ES1
- direct condition: CS1

- dev set: 506 Chinese sentences with 7 refs in English and Spanish
- test set: 1K sentences with 1 reference extracted from CES1
Statistical Machine Translation

source text $f$  \rightarrow  target text $e$

• alignment-based parametric model

$$p(e \mid f) = \sum_ap(e,a \mid f) = \sum_ap_{\theta_{FE}}(e,a \mid f)$$

• parameter estimation:

$$\hat{\theta}_{FE} = \arg\max_{\theta_{FE}} \prod p_{\theta_{FE}}(e_i \mid f_i) \quad \text{given } \{(f_i, e_i)\}$$

• search criterion:

$$f \rightarrow \hat{e} \approx \arg\max_{e} \max_{a} \max p_{\theta_{FE}}(e,a \mid f)$$
Direct baseline system

- open-source MT toolkit **Moses**
- statistical **log-linear** model with 8 features
- weight optimization by means of a **minimum error training** procedure

- **phrase-based** translation model:
  - direct and inverted frequency-based and lexical-based probabilities
  - phrase pairs extracted from symmetrized word alignments (GIZA++)
- 5-gram word-based LM exploiting Improved Kneser-Ney smoothing (IRSTLM)
- standard negative-exponential distortion model
- word and phrase penalties
Bridging at translation time

\[ p(e \mid f) = \sum_g p(e, g \mid f) = \sum_g p(g \mid f) \ p(e \mid g) \]

\[ = \sum_g \sum_b p_{\theta_{FG}}(g, b \mid f) \ \sum_a p_{\theta_{GE}}(e, a \mid g) \]

\[ f \rightarrow \hat{e} \ \approx \ \arg \max_{e,g} \max_{a,b} \ p_{\hat{\theta}_{FG}}(g, b \mid f)p_{\hat{\theta}_{GE}}(e, a \mid g) \]

- two full-fledged systems trained on corpora \((F,G_F)\) and \((G_E,E)\)
- search including the pivot language increases complexity
Coupling with Unconstrained Alignments

- **sentence-level** coupling
- requires performing search over two alignments
- search can be decoupled over a subset of hypotheses:
  - N-best list (or word lattices) of source-to-pivot translations

\[
\begin{align*}
\text{since} & \quad \text{the new administration took office this year} \\
\text{desde que la nueva administracion tomo posesion de su cargo este año} \\
\end{align*}
\]
Coupling with Unconstrained Alignments

Diagram showing the process of coupling with unconstrained alignments.
### Coupling with Unconstrained Alignments

<table>
<thead>
<tr>
<th>$n,m$</th>
<th>rescoring features</th>
<th>dev</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>25.13</td>
<td>16.44</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>25.28</td>
<td>16.60</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>26.65</td>
<td>17.59</td>
</tr>
<tr>
<td>20</td>
<td>16</td>
<td>27.18</td>
<td>17.03</td>
</tr>
<tr>
<td>50</td>
<td>16</td>
<td>27.78</td>
<td>16.96</td>
</tr>
<tr>
<td>100</td>
<td>16</td>
<td>27.89</td>
<td>17.64</td>
</tr>
</tbody>
</table>

- 16 feature scores $> 2$ global scores
- 100x100-best gives best performance on dev set

- time expensive: $(N + 1)$ translation + rescoring
Coupling with Constrained Alignments

- **phrase-level** coupling
- share segmentation on the pivot language and use just one re-ordering
- needs one distortion model that directly models source to target
- needs only one target language model
Coupling with Constrained Alignments

- needs to modify decoder, or
- compose phrase table before decoding

\[ PT(F, E) = PT(F, G) \otimes PT(G, E) \]
\[ = \{ (\tilde{f}, \tilde{e}) | \exists \tilde{g} \text{ s.t. } (\tilde{f}, \tilde{g}) \in PT(F, G_F) \land \exists (\tilde{g}, \tilde{e}) \in PT(G_E, E) \} \]

\[ \phi(\tilde{f}, \tilde{e}) = \left\{ \begin{array}{l}
\sum_{\tilde{g}} \phi(\tilde{f}, \tilde{g}) \phi(\tilde{g}, \tilde{e}) \quad \text{integration} \\
\max_{\tilde{g}} \phi(\tilde{f}, \tilde{g}) \phi(\tilde{g}, \tilde{e}) \quad \text{maximization}
\end{array} \right. \]
Coupling with Unconstrained Alignments
### Coupling with Unconstrained Alignments

<table>
<thead>
<tr>
<th></th>
<th>CE2</th>
<th>CE1</th>
<th>ES1</th>
<th>CE2</th>
<th>CE1</th>
<th>ES1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>src phr</strong></td>
<td>76K</td>
<td>128K</td>
<td>277K</td>
<td>21K</td>
<td>94K</td>
<td></td>
</tr>
<tr>
<td><strong>trg phr</strong></td>
<td>82K</td>
<td>134K</td>
<td>284K</td>
<td>32K</td>
<td>108K</td>
<td></td>
</tr>
<tr>
<td><strong>phr pairs</strong></td>
<td>133K</td>
<td>185K</td>
<td>333K</td>
<td>592K</td>
<td>696K</td>
<td></td>
</tr>
<tr>
<td><strong>avg trans</strong></td>
<td>1.8</td>
<td>1.4</td>
<td>1.2</td>
<td>28.2</td>
<td>7.4</td>
<td></td>
</tr>
<tr>
<td><strong>common</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>59K</td>
<td>143K</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>disjoint</th>
<th>overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>integration</strong></td>
<td>16.65</td>
<td>23.50</td>
</tr>
<tr>
<td><strong>maximization</strong></td>
<td>15.88</td>
<td>22.82</td>
</tr>
</tbody>
</table>

- limited intersection among ⟨g phrases in the disjoint condition:
  - only 27% of Chinese phrases are bridged into Spanish through English
  - only 11% of Spanish are reached through English
- ambiguity increases (esp. for short phrases)
- integration > maximization

- overlap data would be very useful
Bridging at Training Time

- Standard training criterion for (IBM) alignment models

\[ \theta_{FE}^* = \arg \max_{\theta_{FE}} \prod_i p_{\theta_{FE}}(f_i \mid e_i) \quad \text{given } \{(f_i, e_i)\} \]

- Goal: estimate parameters of a "direct" F-E system without a (F,E) corpus

- Assumption: a parallel corpus \{\{(f_i, g_i)\}\}, a full-fledged G-E system \(p_{\hat{\theta}_{GE}}\)

- Solution: \(p(f \mid g)\) above can be replaced with the marginal distribution:

\[ p(f \mid g) = \sum_e p(f \mid e) p_{\hat{\theta}_{GE}}(e \mid g) \]

\[ \hat{\theta}_{FE} = \arg \max_{\theta_{FE}} \sum_{e_i} p_{\theta_{FE}}(f_i \mid e_i) p_{\hat{\theta}_{GE}}(e_i \mid g_i) \]

assuming independence between \(e\) and \(f\) given \(g\).
Approximate ML Estimates

- **Approximation 1**: limit the support of $p_{\hat{\theta}_{GE}}(e \mid g)$ to the best translation
  - basically, we generate a synthetic parallel corpus $(F, \bar{E}_F)$

- **Approximation 2**: limit support over the N-best translations
  - requires MLE of IBM models work with two hidden variables
  - still to be developed

We only experimented the first method, called Viterbi approximation
Random Sampling Method

**Idea**: Generate parallel data by sampling translations from an SMT system

- **Ingredients**: corpus \((F, G)\) and SMT system \(G \rightarrow E\)
- For each example \((f_i, g_i)\) in the training corpus \((F, G)\) generate a random sample of \(m\) translations \(e_{ij}\) of \(g_i\) according to \(p_{\hat{\theta}_{GE}}(e \mid g)\).
- Then build a translation system from \((F, E) = \{(f_i, e_{ij})\}, j = 1, \ldots, m\), by maximizing:

\[
\hat{\theta}_{FE} = \arg\max_{\theta_{FE}} \prod_{i,j} P_{\theta_{FE}}(f_i \mid e_{ij})
\]

- **Implementation**: sample with replacement from the \(n\)-best list of translations \(e\) from \(g_i\) according to \(p_{\hat{\theta}_{GE}}(e \mid g_i)\).
- This approach is indeed more sound than just taking the list of \(n\)-best!
Random Sampling Method
Random Sampling Method

- random sampling with replacement 10 times from a 10-best list of translation
- normalization of Moses scores
- more importance to more reliable translations

Bertoldi et al. Phraes-Based SMT with Pivot Languages October 21st, 2008
Random Sampling Method

<table>
<thead>
<tr>
<th>Method</th>
<th>( n,m )</th>
<th>( \text{Im} )</th>
<th>dev</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viterbi Training</td>
<td>1</td>
<td>S1</td>
<td>22.05</td>
<td>14.56</td>
</tr>
<tr>
<td>Viterbi Training</td>
<td>1</td>
<td>( \tilde{S}2 )</td>
<td>23.58</td>
<td>15.38</td>
</tr>
<tr>
<td>Viterbi Training</td>
<td>1</td>
<td>S1+( \tilde{S}2 )</td>
<td>24.57</td>
<td>16.13</td>
</tr>
<tr>
<td>N-best Training</td>
<td>100</td>
<td>S1+( \tilde{S}2 )</td>
<td>26.04</td>
<td>17.03</td>
</tr>
<tr>
<td>Random Sampling</td>
<td>100</td>
<td>S1+( \tilde{S}2 )</td>
<td>26.02</td>
<td>17.68</td>
</tr>
</tbody>
</table>

- \( \text{LM}(E_1 \cup \tilde{E}_2) > \text{LM}(\tilde{E}_2) > \text{LM}(E_1) \)
- \( \text{N-best Training} > \text{Viterbi Training} \)
- \( \text{N-best Training} \approx \text{Random Sampling} \)
- 21% relative improvement wrt Viterbi-S1 (15% wrt Viterbi-\( \tilde{S}2 \))
Experimental Results

<table>
<thead>
<tr>
<th>Method</th>
<th>CES task</th>
<th>disjoint</th>
<th>overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td></td>
<td>–</td>
<td>23.67</td>
</tr>
<tr>
<td>Cascade 1-best</td>
<td></td>
<td>16.44</td>
<td>24.04</td>
</tr>
<tr>
<td>Cascade N-best</td>
<td></td>
<td>17.64</td>
<td>25.16</td>
</tr>
<tr>
<td>Phrase Table Product</td>
<td></td>
<td>16.65</td>
<td>23.50</td>
</tr>
<tr>
<td>Random Sampling</td>
<td></td>
<td>17.68</td>
<td>25.19</td>
</tr>
</tbody>
</table>

- Cascade 1-best ≈ Phrase Table Product
- Random Sampling ≈ Cascade N-best > Direct
Summary

• approaches to pivot translation task
• mathematical foundation
• experimental comparison

• random sampling approach is the most appealing:
  – quality and efficiency

• unsupervised technique to generate new parallel data
  – suitable to domain adaptation
  – suitable for multi-language pivot translation
Thank you!