Sparse Features for Reordering
(Final Report)

Barry Haddow
Lexi Birch
Han Dan
(Lane Schwartz)

14th September, 2013
Lexicalised Reordering Models in PBMT

Max-Likelihood [Tillmann, 2004], [Koehn et al., 2005, 2007]
- Count orientations in training data
- Maybe smooth
- “Standard Model”

Maxent e.g. [Zens and Ney, 2006]
- Build a classifier to predict M, S or D
- Use lexical features, part-of-speech etc.
Lexicalised Reordering Models in PBMT

- Count orientations in training data
- Maybe smooth
- “Standard Model”

Maxent e.g. [Zens and Ney, 2006]
- Build a classifier to predict M, S or D
- Use lexical features, part-of-speech etc.

Add small number of features (e.g. 6) to translation system
We now have methods to train (i.e. tune) translation systems with thousands or even millions of features.

Normally predicates on hypotheses.

Replace maxent model (indirect objective) ...

... with sparse feature model (direct objective, e.g. BLEU).

Cherry showed improvements on large zh-en and ar-en models

→ Over a maxent baseline

→ In addition to Tillman-style model

→ Lattice batch MIRA better than k-best batch MIRA
Sparse Features for Reordering – Example

**Template**

- `src.left \times orientation`:
  - `sl.sie_M`, `sl.stimmen_D`, `sl.gegen_S`  
- `src.right \times orientation`:
  - `sr.wurden_M`, `sr.stimmen_D`, `sr.sie_S`  
- `tgt.left \times orientation`:
  - `tl.they_M`, `tl.vote_D`, `tl.against_S`  
- `tgt.right \times orientation`:
  - `tr.would_M`, `tr.vote_D`, `tr.you_S`

**Features**

- `Sie würden gegen sie stimmen`  
  - They would vote against you
Lexicalised Reordering in Hiero Models

- Zens&Ney Reordering (maxent)
  [Huck et al, 2012]
- Tillman Reordering (max-like)
  [Huck et al, 2013]
- Latter shown to perform better – both beat baseline
Suppose we apply a rule:

\[ X \rightarrow a \ X_1 \ b \ X_2 \mid X_1 \ z \ X_2 \ y \]

With \( X_1 \) covering \( c \) and \( X_2 \) covering \( de \)

→ Add features indicating that words \( c \) are monotone with respect to \( de \)

Or if the rule is:

\[ X \rightarrow a \ X_1 \ b \ X_2 \mid X_2 \ p \ X_1 \]

→ Add features indicating that words \( c \) are swapped with respect to \( de \)
Suppose we apply a rule:

\[ X \rightarrow a \ X_1 \ b \ X_2 \ | \ X_1 \ z \ X_2 \ y \]

With \( X_1 \) covering \( c \) and \( X_2 \) covering \( de \)

\[ \rightarrow \] Add features indicating that words \( c \) are monotone with respect to \( de \)

Or if the rule is:

\[ X \rightarrow a \ X_1 \ b \ X_2 \ | \ X_2 \ p \ X_1 \]

\[ \rightarrow \] Add features indicating that words \( c \) are swapped with respect to \( de \)
Sparse Reordering in Hiero Models – Example

\[ X \rightarrow acd X_1 b X_2 \mid qrX_2 p X_1 \]

\[
\begin{array}{cc}
| & | & | & | & |
\hline
q & r & X_2 & p & X_1 \\
\hline
\end{array}
\]

- src\_left\_d\_mono
- src\_right\_e\_mono
- src\_left\_f\_swap
- src\_right\_b\_swap
- src\_left\_b\_swap
- src\_right\_g\_swap

- \( X_1 \) covers ef, \( X_2 \) covers g
Results

- Small de-en model (news commentary ≈ 130k sentences)
- Use src.left variant, and top 100 words.

<table>
<thead>
<tr>
<th>Model</th>
<th>Tune</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>26.83</td>
<td>27.71</td>
</tr>
<tr>
<td>Sparse Reorder</td>
<td>27.15</td>
<td>27.90</td>
</tr>
</tbody>
</table>

(Baseline trained 25 iterations, Sparse reordering 10)
<table>
<thead>
<tr>
<th>Feature</th>
<th>Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>und_swap</td>
<td>-0.159</td>
</tr>
<tr>
<td>,_swap</td>
<td>-0.111</td>
</tr>
<tr>
<td>ein_mono</td>
<td>-0.046</td>
</tr>
<tr>
<td>sind_mono</td>
<td>-0.045</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>sich_mono</td>
<td>0.039</td>
</tr>
<tr>
<td>,_mono</td>
<td>0.057</td>
</tr>
<tr>
<td>OTHER_mono</td>
<td>0.062</td>
</tr>
<tr>
<td>die_mono</td>
<td>0.100</td>
</tr>
</tbody>
</table>