“Learning Hierarchical Translation Structure with Linguistic Annotations”

Markos Mylonakis and Khalil Sima‘an
ILLC, University of Amsterdam

Presented by Greg Hanneman
11-734: Advanced MT Seminar
March 21, 2012
Setup

• Hierarchical MT systems:

Constrained to linguistic syntax
- GHKM
- tree-to-string
- tree transducers

vs.

Soft syntactic constraints
- Hiero
- SAMT
- Marton and Resnik

This work
Hierarchical Reordering SCFG

- All rules have one of the following forms:

  **Monotonic translation**
  \[
  A \rightarrow [B \ C]::[B \ C] \\
  A^L \rightarrow [B \ C]::[B \ C] \\
  A^R \rightarrow [B \ C]::[B \ C]
  \]

  **Reordered translation**
  \[
  A \rightarrow [B^L \ C^R]::[C^R \ B^L] \\
  A^L \rightarrow [B^L \ C^R]::[C^R \ B^L] \\
  A^R \rightarrow [B^L \ C^R]::[C^R \ B^L]
  \]

  **Phrase pair emission**
  \[
  A \rightarrow [A_p]::[A_p] \\
  A^L \rightarrow [A_p]::[A_p] \\
  A^R \rightarrow [A_p]::[A_p]
  \]

  **Phrase pair generation**
  \[
  A_p \rightarrow [x]::[y] \\
  A_p^L \rightarrow [x]::[y] \\
  A_p^R \rightarrow [x]::[y]
  \]
Extracting Rules

• Word-align parallel corpus

• Parse source side (Charniak)

• Construct all possible labels for consistently aligned spans (SAMT-style)

• Extract (minimal?) rules
Extracting Rules

which is the problem
Extracting Rules

\[
\begin{align*}
\text{NP}^R & \rightarrow [\text{DT} \ \text{NN}]::[\text{DT} \ \text{NN}] \\
\text{VP}/\text{VBZ}^R & \rightarrow [\text{X} \ \text{X}]::[\text{X} \ \text{X}] \\
\text{NP}^R & \rightarrow [\text{NP}_p^R]::[\text{NP}_p^R] \\
\text{NP}_p^R & \rightarrow [\text{the problem}]::[\text{das Problem}]
\end{align*}
\]
Extracting Rules

VP → [VBZ\textsuperscript{L} NP\textsuperscript{R}]::
[NP\textsuperscript{R} VBZ\textsuperscript{L}]

X → [VBZ\textsuperscript{L} X\textsuperscript{R}]::[X\textsuperscript{R} VBZ\textsuperscript{L}]

VBZ+NP → [VBZ+NP\textsubscript{p}]::
[VBZ+NP\textsubscript{p}]

…

<table>
<thead>
<tr>
<th>der das Problem</th>
<th>which is the problem</th>
</tr>
</thead>
</table>

\begin{itemize}
  \item which
  \item is
  \item the
  \item problem
\end{itemize}

\begin{align*}
  X & \rightarrow [\text{WHNP} + \text{VP}]::[\text{WHNP+VP} + \text{NP}] \\
  X & \rightarrow [\text{WHNP} + \text{VBZ} + \text{NP}]::[\text{WHNP+VBZ} + \text{NP}]
\end{align*}
Scoring Rules

• Phrase pair rules $A_P \rightarrow [x]::[y]$
  – $P(x, y \mid A_P)$ estimated from label charts
  – Standard $P(x \mid y)$ and $P(y \mid x)$
  – Standard $P_{\text{lex}}(x \mid y)$ and $P_{\text{lex}}(y \mid x)$
  – Word penalty $f(|y|)$
Scoring Rules

- Hierarchical rules
  - Estimate $P(\text{RHS} \mid \text{LHS})$, but not using MLE

Cross-Validating EM

Maximize likelihood of this portion of the training data... ... using the model that comes from the rest of the data

- Reordering count
So That’s Zillions of Rules...

- Rules appearing in only one partition of training data get ignored
- Rules below a minimum expected count in CV-EM get removed
- Decoder restricted to label chart of an input sentence
- Decoder cells have separate bins for each nonterminal
Scoring Derivations

• Probability of a derivation =
  Language model probability ×
  Product of scores for each phrase pair ×
  Product of scores for each hierarchical rule

  Feature weights trained using MERT

• Probability of a joint output =
  Sum over all derivations that produce it

  No Viterbi approximation?
Experiments

• Trained on Europarl / news data
  – Very small: 200k or 400k sentence pairs!
  – En to French, German, Dutch, and Chinese

• WMT 2007 (news) test set

• Modified KN language model
  – Very small: 1 million sentences
  – Trigram

• Compare with Joshua/Hiero baseline
### Characteristic Results

- 400k training sentences, BLEU

<table>
<thead>
<tr>
<th>Language</th>
<th>Joshua</th>
<th>LTS</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>29.58</td>
<td>29.83</td>
<td>+0.25</td>
</tr>
<tr>
<td>German</td>
<td>18.86</td>
<td>19.49</td>
<td>+0.63</td>
</tr>
<tr>
<td>Dutch</td>
<td>22.25</td>
<td>22.92</td>
<td>+0.67</td>
</tr>
<tr>
<td>Chinese</td>
<td>23.24</td>
<td>25.16</td>
<td>+1.92</td>
</tr>
</tbody>
</table>
Discussion: Contributions

- Viable syntactic grammar that’s some amalgam ITG, SAMT, or Hiero
- More explicit modeling of reordering behavior by category type / context
- Scoring that uses held-out data to go beyond “count and normalize”
- Follow-up experiment: Non-X labels help
Discussion: “How You Say It”

- Syntax-based MT has “inadequate constraints”? Doesn’t desire to soften indicate restrictive constraints?
- SCFGs have “weak independence assumptions”? Doesn’t passing ordering info weaken a strong assumption?
- Highest-probability rule will “always” win? 
  - Conditional probabilities?
  - Reordering-based features?
  - Language model?