Context-Dependent Alignment Models

Jamie Brunning, Adria de Gispert, Bill Byrne
Cambridge University
Context Dependent Translation in Alignment

- Basic models rely on context-independent word-to-word translation -- Models 1, 2, ..., HMMs and variants
- Not easy to introduce translation dependency on context
  - need to control model size and computational complexity
- **New:** Context clustering via decision trees
  - Each source word gets its own tree (many, many trees...)
  - Questions are asked about the source word contexts
  - Clustering is performed during iterative EM training
  - EM auxiliary function is used as the clustering purity function
    - identical to HTK acoustic clustering

---

Context-dependent alignment models for statistical machine translation.

Context-dependent alignment models: Clustering by Part of Speech tags

- POS taggers were run on both sides of the parallel texts
- Created question sets for POS contexts
  - questions covered 5-word contexts
  - Grew four sets of trees: two for Ar/Zh->En and two for En->Ar/Zh

Most frequent root node questions for Ar-En

<table>
<thead>
<tr>
<th>English Question</th>
<th>Count</th>
<th>Arabic Questions</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is_Next_Preposition</td>
<td>1523</td>
<td>Is_Prev_Preposition</td>
<td>1110</td>
</tr>
<tr>
<td>Is_Prev_Determiner</td>
<td>1444</td>
<td>Is_Next_Preposition</td>
<td>993</td>
</tr>
<tr>
<td>Is_Prev_Preposition</td>
<td>1209</td>
<td>Is_Prev_Noun</td>
<td>981</td>
</tr>
<tr>
<td>Is_Prev_Adjective</td>
<td>864</td>
<td>Is_Prev_Coordinating_Conj</td>
<td>627</td>
</tr>
</tbody>
</table>

- 26% and 40% of Arabic and English words have CD translations
- Question combinations get fairly complex further down the tree
CD models improve both alignment and translation quality

- CD Model 1: Large reductions in Alignment Error Rate
- CD HMMs: Improvements in alignment and in BLEU

### BLEU scores on development training and test sets

<table>
<thead>
<tr>
<th></th>
<th>Ar - En</th>
<th>Zh - En</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alignment</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CI-HMM</td>
<td>49.4</td>
<td>28.5</td>
</tr>
<tr>
<td>CD-HMM</td>
<td>49.7</td>
<td>29.0</td>
</tr>
<tr>
<td><strong>mt02-05test</strong></td>
<td>46.3</td>
<td>26.9</td>
</tr>
<tr>
<td><strong>MT08-nw</strong></td>
<td>46.9</td>
<td>27.7</td>
</tr>
<tr>
<td><strong>test-nw</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MT08-nw</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Using decision trees for context dependent translation models scales well to larger data sets:
  - Parallelized implementation using MTTK alignment tools
  - Good gains on the full AGILE training sets (0.5 -- 1.0 BLEU)
Hierarchical Phrase-based Translation with Weighted Finite State Transducers

Adria de Gispert, Bill Byrne
Cambridge University
Gonzalo Iglesias, Eduardo Banga
University of Vigo
HiFST: WFST implementation of Hiero

- WFST formulation of Hiero-style translation
  - Alternative implementation to Cube Pruning (CPH)
  - Unioning/concatenating/pruning/etc. is done with standard WFST ops
  - Based on Google OpenFST toolkit
- Key idea: work with lattices containing many hypotheses and avoid working with individual translation hypotheses
  - vastly larger hypothesis spaces
- Source language sentence is parsed using source-side of the Hiero translation rules
  - CYK grid is built up and applicable rules are kept in each cell
- A translation lattice is built for every cell, based on its rules
  - lattice arcs have translated text or pointers to other lattices
  - these intermediate lattices can be minimized, determinized, etc.
  - lattice pointers are only expanded when translation is complete or if intermediate pruning is required
    - fast, and controls memory usage

Hierarchical phrase-based translation with weighted finite state transducers.
Delayed Translation

lattices with translated text and pointers to lower lattices produced by hierarchical rules

CYK grid

pointers to lattices at lower cells

lattices with translated text
Translation improves with denser search spaces

• More efficient search:
  – 48% reduction in search errors for ZH->EN translation
  – AR->EN translation runs without pruning

• Direct generation of translation lattices

• Richer search space improves subsequent rescoring
  – HiFST is comparable to CPH in first pass [ (a) vs (c) ]
  – HiFST produces richer/better hypotheses for rescoring
    • increased gains from 5Gram LM + Minimum Bayes Risk rescoring [ (b) vs (d) ]

<table>
<thead>
<tr>
<th></th>
<th>Decoder</th>
<th>ZH-&gt;EN test-nw</th>
<th>ZH-&gt;EN NIST MT08</th>
<th>AR-&gt;EN mt02-05-test</th>
<th>AR-&gt;EN NIST MT08</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BLEU / TER</td>
<td>BLEU / TER</td>
<td>BLEU / TER</td>
<td>BLEU / TER</td>
</tr>
<tr>
<td>(a)</td>
<td>CPH</td>
<td>32.2 / 59.9</td>
<td>27.2 / 60.2</td>
<td>51.5 / 42.2</td>
<td>42.5 / 48.6</td>
</tr>
<tr>
<td>(b)</td>
<td>CPH +5G+MBR</td>
<td>32.7 / 59.4</td>
<td>28.1 / 59.3</td>
<td>52.6 / 41.4</td>
<td>43.4 / 48.1</td>
</tr>
<tr>
<td>(c)</td>
<td>HiFST</td>
<td>32.2 / 60.0</td>
<td>27.1 / 60.5</td>
<td>51.6 / 42.1</td>
<td>42.4 / 48.7</td>
</tr>
<tr>
<td>(d)</td>
<td>HiFST+5G+MBR</td>
<td>33.4 / 58.5</td>
<td>28.9 / 58.9</td>
<td>53.3 / 40.9</td>
<td>44.0 / 48.0</td>
</tr>
</tbody>
</table>