The CUED HiFST System for WMT10
J. Pino, G. Iglesias, A. de Gispert, G. W. Blackwood, J. Brunning, W. J. Byrne
Department of Engineering, University of Cambridge

SUMMARY
✦ Cambridge University Engineering Department submission to WMT10.
✦ Translation tasks: FR→EN and SP→EN shared translation tasks.
✦ Decoder: HiFST (Iglesias et al., NAACL’09), a hierarchical phrase-based decoder implemented using WFSTs.
✦ Shallow hierarchical grammar that requires no pruning in search.
✦ Investigate the use of context-dependent alignment models in the FR-EN system.
✦ Additional experiment on multi-source translation: lattice minimum Bayes-risk decoding is an effective framework for multi-source translation, leading to large gains in BLEU score.

TRANSLATION PIPELINE
✦ Preprocessing
   HTML tags, tokenize, lowercase.
✦ Alignment: MTTK (Deng and Byrne, EMNLP’05)
   Word-to-phrase HMM

P(K=6j=5, j=6)

<table>
<thead>
<tr>
<th>The</th>
<th>Le</th>
</tr>
</thead>
<tbody>
<tr>
<td>φ=1</td>
<td>φ=1</td>
</tr>
<tr>
<td>φ=2</td>
<td>φ=1</td>
</tr>
</tbody>
</table>

Context dependent HMM (Brunning et al., NAACL’09)

\[ p(j'_i, a_i|c_i, c_i') = \prod_{j=1}^{f} \alpha(a_i, a_{i-1}, f) \beta(f, c_i, \mu(c_i)) \]

\( c_i; \text{context defined as part-of-speech tags of } c_{i-1}, c_{i-1} \)

\( \mu; \text{clustering of context with decision trees} \)


2. Grammar Extraction
   ✦ Viterbi union
   ✦ Rule extraction: standard phrase and rule extraction, rule filtering by pattern (Iglesias et al., EACL’09).
   ✦ Shallow-1 grammar (Iglesias et al., EACL’09):

      Hierarchical rule non terminals must be rewritten with phrase-based rules (degree of rule nesting is 1). Previously shown to work well for the SP-EN language pair.

3. Hierarchical Lattice-Based Decoding
   ✦ Build target language lattice by traversing the CYK grid built on the source.
   ✦ Compose target language translation lattice with language model acceptor.
   ✦ No pruning in search, full exploration of the search space: lattice + shallow-1 grammar.


5g Rescoring: build stupid-backoff 5-gram language models (Brants et al., EMNLP’07) on all available monolingual data.

Lattice MBR (LMBR) Rescoring: maximize conditional expected gain under linearised sentence-level BLEU score (Tromble et al., EMNL’08; Blackwood et al., ACL’10). Also used for hypothesis combination.

Postprocessing: uppercase, detokenize, join apostrophes in French.

EXPERIMENTS AND RESULTS

<table>
<thead>
<tr>
<th>Task</th>
<th>Configuration</th>
<th>newstest2008</th>
<th>newstest2009</th>
<th>newstest2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR → EN</td>
<td>HiFST (A)</td>
<td>23.4</td>
<td>26.4</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>HiFST (B)</td>
<td>24.0</td>
<td>27.3</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>HiFST (CD)</td>
<td>24.2</td>
<td>27.6</td>
<td>28.0</td>
</tr>
<tr>
<td></td>
<td>HiFST (C)</td>
<td>24.7</td>
<td>28.4</td>
<td>28.5</td>
</tr>
<tr>
<td></td>
<td>+5g+LMBR</td>
<td>25.3</td>
<td>29.1</td>
<td>29.3</td>
</tr>
<tr>
<td></td>
<td>LMBR (CD)(C)</td>
<td>25.6</td>
<td>29.3</td>
<td>29.6</td>
</tr>
<tr>
<td>EN → FR</td>
<td>HiFST (A)</td>
<td>22.5</td>
<td>24.2</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>HiFST (B)</td>
<td>23.4</td>
<td>24.8</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>HiFST (CD)</td>
<td>23.3</td>
<td>24.8</td>
<td>26.7</td>
</tr>
<tr>
<td></td>
<td>+5g+LMBR</td>
<td>23.7</td>
<td>25.3</td>
<td>27.1</td>
</tr>
<tr>
<td></td>
<td>LMBR (CD)(C)</td>
<td>24.2</td>
<td>26.1</td>
<td>28.2</td>
</tr>
<tr>
<td>SP → EN</td>
<td>HiFST (A)</td>
<td>23.9</td>
<td>25.8</td>
<td>27.8</td>
</tr>
<tr>
<td></td>
<td>+5g+LMBR</td>
<td>24.7</td>
<td>25.5</td>
<td>29.1</td>
</tr>
</tbody>
</table>

MULTI-SOURCE TRANSLATION
✦ LMBR decoding (de Gispert et al., CL’10) in FR-EN union SP-EN lattices

Linear interpolation:

\[ p(u|E, λ) = λ_u p(u|E, λ_u) + (1 - λ_u) p(u|E, μ_u) \]

Very large gains with respect to best individual systems in combination

CONCLUSIONS
✦ More parallel data helps for FR-EN
✦ Context dependent models help in the FR→EN direction
✦ Rescoring methods give consistent gains

FEATURES
✦ Lattice decoding (de Gispert et al., CL’10) in FR-EN union SP-EN lattices

Linear interpolation:

\[ p(u|E, λ) = λ_u p(u|E, λ_u) + (1 - λ_u) p(u|E, μ_u) \]

Very large gains with respect to best individual systems in combination

CONCLUSIONS
✦ More parallel data helps for FR-EN
✦ Context dependent models help in the FR→EN direction
✦ Rescoring methods give consistent gains