International Workshop on Spoken Language Translation
Kyoto, Japan
September 30 - October 1, 2004

Alignment Templates: the RWTH SMT System

Oliver Bender, Richard Zens, Evgeny Matusov, and Hermann Ney

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik VI
Computer Science Department
RWTH Aachen University
D-52056 Aachen
Content

1. overview: statistical machine translation
2. loglinear models
3. alignment templates
4. feature functions
5. minimum error training
6. $n$-best lists and rescoring
7. experimental results
8. summary
Related work


Overview: Statistical Machine Translation

- source string $f^J_1 = f_1 ... f_j ... f_J$ to be translated into a target string $e^I_1 = e_1 ... e_i ... e_I$.

- classical source-channel approach:

  $$
  \hat{e}^I_1 = \arg\max_{e^I_1} \{Pr(e^I_1|f^J_1)\}
  $$
  $$
  = \arg\max_{e^I_1} \{Pr(e^I_1) \cdot Pr(f^J_1|e^I_1)\}
  $$

- $Pr(f^J_1|e^I_1)$: translation model
  (usually can be further decomposed into alignment and lexicon model)

- $Pr(e^I_1)$: language model
Loglinear models

• alternative: direct modeling of the posterior probability $Pr(e_1^I|f_1^J)$

• use a loglinear model (Och and Ney 2002):

$$Pr(e_1^I|f_1^J) = p_{\lambda}^M(e_1^I|f_1^J) = \frac{\exp \left[ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right]}{\sum_{e_1'} \exp \left[ \sum_{m=1}^{M} \lambda_m h_m(e_1', f_1^J) \right]}$$

• decision rule:

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\}$$

• advantages:
  
  – easy integration of additional models/feature functions $h_m$
  
  – minimum error training of model scaling factors $\lambda_m$
Alignment Templates

- primary translation model: alignment templates
- describes the alignment between sequences of source and target words
- automatically trained word classes are used instead of words for better generalization
- translation model incorporates:
  - phrase alignment probability
  - probability to apply an alignment template
  - phrase translation probability
- alignment templates extracted automatically from automatic word alignments
Alignment Templates: Example

- alignment $A$ is a mapping from source sentence positions to target sentence positions $a_1 \ldots a_J$, $a_j \in \{0, \ldots, I\}$.
- alignment may contain connections $a_j = 0$ with the ‘empty’ word $e_0$
- alignments are created automatically with GIZA++ using IBM-1, HMM, and IBM-4 models
Alignment Combination Heuristics

- word alignments $A_1$ and $A_2$ are trained in source-to-target and target-to-source direction, respectively.
- such alignments contain many-to-one mappings in one direction only.
- alignment combination depends on the particular language pair.
- best translation results achieved:
  - Chinese-English: using alignments which only allow many-to-one mappings of English words.
    * extend intersection $A_1 \cap A_2$ by additional points.
    * add a new point if either a horizontal or a vertical direct neighbor point exists.
Base Models Used in Search

- alignment templates
- single-word translation model $p(e|f)$
- word-based trigram language model
- class-based five-gram language model
- word penalty model
- phrase penalty model
- penalty for alignment template reorderings
Minimum Error Training

- optimize the model scaling factors $\lambda_i^M$
- training criterion: minimal number of errors on a development corpus
- optimization with respect to a certain automatic translation score
  $(100 - \text{NIST}, 1 - \text{BLEU}, \text{WER})$
- use the downhill simplex optimization algorithm
- translate the whole development corpus in each iteration of the algorithm
- algorithm converges after about 200 iterations
Search

• search characteristics:
  – reordering: within alignment templates: fixed in training
  – reordering of alignment templates: unconstrained or ITG (Japanese-English)
  – search organization along target string positions
  – beam search to handle the huge search space

• generation of $n$-best lists:
  – during search, generate word graphs
  – using the $A^*$ search algorithm,
    compute $n$-best lists from the word graphs
Additional $n$-best List Features

- (inverse) IBM-1 lexicon model $p(f|e)$ (as trained with GIZA++)
  + captures lexical co-occurrences, helpful for translation adequacy
- deletion model
  + penalizes too short translation hypotheses
- high-order $n$-gram language models ($n = 4, 5, \ldots, 9$)
  + enrich the system with knowledge about longer target language phrases
Deletion Model

- the produced translations are often shorter than the reference translations
- longer hypotheses are to be favored
- deletion model feature (Och et al. 2004): for a given threshold $\alpha$:
  - count the number of source words, for which the IBM-1 translation probability given any of the target words in the hypothesis is below $\alpha$.
  - use several features with different values of $\alpha$ (0.1, 0.01, etc.)
- threshold $\alpha$ tuned on a development corpus
Experimental results

- IWSLT 2004 Evaluation
- rescoring improvements
Evaluation Methodology

- subjective evaluation as specified by the IWSLT 2004 consortium
  - translation fluency: from 1 (“incomprehensible”) to 5 (“flawless English”)
  - translation adequacy: how much information from a gold standard translation is contained in the hypothesis, from 1 (“none”) to 5 (“all”)
- objective evaluation: different automatic metrics computed using multiple references
  - Word Error Rate (mWER)
  - Position-Independent Word Error Rate (mPER)
  - BLEU score
  - NIST score
  - GTM score
## BTEC Chinese-English Supplied Corpus Statistics

<table>
<thead>
<tr>
<th></th>
<th>Chinese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>train sentences</strong></td>
<td>20,000</td>
<td></td>
</tr>
<tr>
<td>words</td>
<td>182,904</td>
<td>160,523</td>
</tr>
<tr>
<td>singletons</td>
<td>3,525</td>
<td>2,948</td>
</tr>
<tr>
<td>vocabulary</td>
<td>7,643</td>
<td>6,982</td>
</tr>
<tr>
<td><strong>dev sentences</strong></td>
<td>506</td>
<td></td>
</tr>
<tr>
<td>words</td>
<td>3,515</td>
<td>3,595</td>
</tr>
<tr>
<td><strong>test sentences</strong></td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>words</td>
<td>3,794</td>
<td>–</td>
</tr>
</tbody>
</table>
# BTEC Japanese-English Supplied Corpus Statistics

<table>
<thead>
<tr>
<th></th>
<th>Japanese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>train</strong> sentences</td>
<td>20 000</td>
<td></td>
</tr>
<tr>
<td>words</td>
<td>209 012</td>
<td>160 427</td>
</tr>
<tr>
<td>singletons</td>
<td>4 108</td>
<td>2 956</td>
</tr>
<tr>
<td>vocabulary</td>
<td>9 277</td>
<td>6 932</td>
</tr>
<tr>
<td><strong>dev</strong> sentences</td>
<td>506</td>
<td></td>
</tr>
<tr>
<td>words</td>
<td>4 374</td>
<td>3 595</td>
</tr>
<tr>
<td><strong>test</strong> sentences</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>words</td>
<td>4 370</td>
<td>–</td>
</tr>
</tbody>
</table>
BTEC Japanese-English Unrestricted Data Track Corpus Statistics

- additional resources:
  - full BTEC 1 Japanese-English corpus
  - Spoken Language Database (dialogs, hotel reservation domain)
- kindly provided by ATR

<table>
<thead>
<tr>
<th></th>
<th>Japanese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sentences</td>
<td>240 672</td>
<td></td>
</tr>
<tr>
<td>words</td>
<td>1 974 407</td>
<td>1 770 190</td>
</tr>
<tr>
<td>singletons</td>
<td>8 975</td>
<td>3 658</td>
</tr>
<tr>
<td>vocabulary</td>
<td>26 037</td>
<td>14 301</td>
</tr>
<tr>
<td>dev</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sentences</td>
<td>506</td>
<td></td>
</tr>
<tr>
<td>words</td>
<td>3 515</td>
<td>3 595</td>
</tr>
<tr>
<td>test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sentences</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>words</td>
<td>3 794</td>
<td>–</td>
</tr>
</tbody>
</table>
### Official Evaluation Results

<table>
<thead>
<tr>
<th>Language Data Track Pair</th>
<th>Automatic Evaluation</th>
<th>Subj. Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mWER [%]</td>
<td>mPER [%]</td>
</tr>
<tr>
<td>CE Small</td>
<td>45.6</td>
<td>39.0</td>
</tr>
<tr>
<td>JE Small</td>
<td>41.9</td>
<td>33.8</td>
</tr>
<tr>
<td>Unrestricted</td>
<td>30.6</td>
<td>24.9</td>
</tr>
</tbody>
</table>

- balanced fluency/adequacy scores
- NIST score has the highest correlation with subjective ratings
Rescoring Improvements - Chinese-English

- error rates and scores on the development corpus (CSTAR 2003 test set)
- best overall performance achieved when optimizing the model scaling factors with respect to the NIST score
- base model scaling factors optimized using a narrow beam
- \( n \)-best lists created using a broader beam
- each added feature results in performance gain

<table>
<thead>
<tr>
<th>System</th>
<th>Error Rates</th>
<th>Accuracy Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mWER [%]</td>
<td>mPER [%]</td>
</tr>
<tr>
<td>baseline</td>
<td>55.2</td>
<td>45.6</td>
</tr>
<tr>
<td>broad beam</td>
<td>53.4</td>
<td>45.3</td>
</tr>
<tr>
<td>+ IBM-1 lexicon</td>
<td>50.9</td>
<td>42.1</td>
</tr>
<tr>
<td>+ deletion model</td>
<td>50.6</td>
<td>42.2</td>
</tr>
<tr>
<td>+ 9-gram LM</td>
<td>50.6</td>
<td>42.2</td>
</tr>
<tr>
<td></td>
<td>BLEU [%]</td>
<td>NIST</td>
</tr>
<tr>
<td>baseline</td>
<td>34.8</td>
<td>7.76</td>
</tr>
<tr>
<td>broad beam</td>
<td>33.6</td>
<td>7.63</td>
</tr>
<tr>
<td>+ IBM-1 lexicon</td>
<td>36.4</td>
<td>8.06</td>
</tr>
<tr>
<td>+ deletion model</td>
<td>37.1</td>
<td>8.07</td>
</tr>
<tr>
<td>+ 9-gram LM</td>
<td>38.0</td>
<td>8.14</td>
</tr>
</tbody>
</table>
Rescoring Improvements - Japanese-English

- error rates and scores on the development corpus (CSTAR 2003 test set)
- ITG reordering constraints in search improve the translation quality

<table>
<thead>
<tr>
<th>System</th>
<th>Error Rates</th>
<th>Accuracy Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mWER [%]</td>
<td>mPER [%]</td>
</tr>
<tr>
<td>baseline</td>
<td>48.7</td>
<td>38.6</td>
</tr>
<tr>
<td>+ ITG constraints</td>
<td>45.1</td>
<td>36.0</td>
</tr>
<tr>
<td>+ broad beam</td>
<td>49.5</td>
<td>37.3</td>
</tr>
<tr>
<td>+ IBM-1 lexicon</td>
<td>44.6</td>
<td>35.7</td>
</tr>
<tr>
<td>+ deletion model</td>
<td>43.2</td>
<td>34.7</td>
</tr>
<tr>
<td>+ 5-gram LM</td>
<td>42.6</td>
<td>34.2</td>
</tr>
</tbody>
</table>
Conclusions

• translation system based on loglinear model combination
• additional knowledge sources easily integrated as features
• phrasal context and local word reorderings are important
  ⇒ captured in the alignment templates model
• direct optimization of base models using minimum error training of model scaling factors
• an additional deletion model feature penalizes too short translations
• scaling factors for additional features optimized using $n$-best lists of translation hypotheses
• optimization of the RWTH system with respect to the NIST score seems to correspond best to subjective evaluation criteria
• on the BTEC Chinese-English and Japanese-English tasks, translations of good quality were produced