The ITC-irst Statistical Machine Translation System for IWSLT-2004

N. Bertoldi, R. Cattoni, M. Cettolo, M. Federico

ITC-irst

Centro per la Ricerca Scientifica e Tecnologica
I-38050 Povo (Trento), Italy

{bertoldi,cattoni,cettolo,federico}@itc.it
Outline

- The ITC-irst SMT System
  - Log-linear Model
  - Phrase-based Model
  - Decoding
  - System Architecture

- Experiments for IWSLT-2004
  - Selection of Training Data
  - Chinese Segmentation
  - Official Results
Log-linear model for SMT

Maximum Entropy framework for word-alignment MT approach:

\[ e^* = \arg \max_e \sum_a \Pr(e, a | f) \approx \arg \max_e \max_a \Pr(e, a | f) \tag{1} \]

Pr(e, a | f) is determined through real valued feature functions \( h_i(e, f, a), i = 1 \ldots M \), and takes the parametric form:

\[ p_\lambda(e, a | f) = \frac{\exp\{\sum_i \lambda_i h_i(e, f, a)\}}{\sum_{e,a} \exp\{\sum_i \lambda_i h_i(e, f, a)\}} \tag{2} \]

Example: feature functions of IBM Model 4:

\[
\begin{align*}
h_1(e, f, a) &= \log \Pr(e) \quad &\text{(target language model)} \\
h_2(e, f, a) &= \log \Pr(\phi | e) \quad &\text{(fertility model)} \\
h_3(e, f, a) &= \log \Pr(\tau | e, \phi) \quad &\text{(lexicon model)} \\
h_4(e, f, a) &= \log \Pr(\pi | e, \phi, \tau) \quad &\text{(distortion model)}
\end{align*}
\]
Phrase-based model

- a phrase is a sequence of one or more words (no semantic or syntactic meaning)
- one-to-one correspondence between phrases
- source words may be not translated (into $\tilde{e}_0$)
- insertion of target phrases without translation
- all models at phrase level except language model (at word level)
- frequency-based distributions
- statistics collected from a word alignment (e.g. produced by GIZA++)
Decoding

- **approximate search criterion:** \( \hat{e}^* \approx \arg \max_{\hat{e}} \max_a \sum_i \lambda_i h_i(\hat{e}, f, a) \)

- DP-based algorithm

- search progresses synchronously along the target string
  (decisions are taken when generating target phrase)

- search ends when all source positions are covered

- optimal final theory is chosen among all complete theories

- beam search: threshold pruning, histogram pruning

- garbaging of theories without extensions

- constraints on the length of the source and target phrases
System Architecture: Run-Time

MODEL PARAMETERS
- lexicon distributions
- fertility ""
- distortion ""
- LM
- scaling factors

Preprocessing → PREPROCESSED
   src

Decoder
   src

POSTPROCESSED

BEST HYPOTHESIS
   tgt

BEST TRANSLATION
   tgt

Postprocessing
System Architecture: Training

Phase 1:
Phrase–based Model Training

Phase 2:
Minimum Error Training

PREPROCESSED TRAINING SET

WORD ALIGNMENTS

PREPROCESSED DEVELOPMENT SET

PHRASE–BASED MODEL PARAMETERS
- lexicon distributions
- fertility
- distortion
- LM

Phrase Extraction

Decoder

Evaluator

SCORE

Phase 1:
Phrase–based Model Training

Phase 2:
Minimum Error Training

Simplex

Bertoldi, et al.  IWSLT-2004 Workshop  Kyoto, 30 September 2004
Experiments

- Chinese-English track (all the three data conditions)
- no optimization on the post-processing
- BLEU score for data selection and minimum error training
ITC-irst SMT Experiments

Preprocessing

- tokenization (EN)*
- dp-based Chinese segmentation (CH)*
- rule-based recognition of time and numerical expressions (CH, EN):
  weekdays, month names, percentages, cardinals, ordinals
- lower case text (EN)
- ignored unknown Chinese words
- split of long sentences (test)

* when needed
## Selection of Training Data

<table>
<thead>
<tr>
<th>System name</th>
<th>extra data</th>
<th>BLEU</th>
<th>NIST</th>
<th>MWER</th>
<th>MPER</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td></td>
<td>0.3001</td>
<td>7.0157</td>
<td>50.8</td>
<td>41.5</td>
</tr>
<tr>
<td>lm-btec</td>
<td>BTEC</td>
<td>0.3509</td>
<td>7.5099</td>
<td>47.2</td>
<td>38.1</td>
</tr>
<tr>
<td>lm-db1</td>
<td>BTEC, DB1</td>
<td>0.3466</td>
<td>7.4475</td>
<td>47.6</td>
<td>38.3</td>
</tr>
<tr>
<td>lm-db2</td>
<td>BTEC, DB2</td>
<td>0.3460</td>
<td>7.4427</td>
<td>47.1</td>
<td>38.3</td>
</tr>
<tr>
<td>tm-btec</td>
<td>BTEC</td>
<td>0.4311</td>
<td>8.5336</td>
<td>42.0</td>
<td>33.3</td>
</tr>
<tr>
<td>tm-db3</td>
<td>BTEC, DB3</td>
<td>0.4574</td>
<td>8.7890</td>
<td>39.7</td>
<td>30.5</td>
</tr>
</tbody>
</table>

- DB1: news corpora
- DB2: press releases of Hong Kong Special Administrative Region
- DB3: selection of corpora from NIST MT-EVAL 2004 competition (large data condition)
Chinese Segmentation

1. Supplied:
   - Chinese segmentation as provided in the supplied training/test corpora

2. Special:
   - Chinese segmentation from scratch
   - word-frequency list (7K) extracted from the supplied training corpus

3. Full:
   - Chinese segmentation from scratch
   - word-frequency list (44K) provided by LDC
## Official Results: Objective Scores

<table>
<thead>
<tr>
<th>Data Condition</th>
<th>Segmentation</th>
<th>BLEU</th>
<th>NIST</th>
<th>MWER</th>
<th>MPER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>training</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supplied</td>
<td>Supplied</td>
<td>0.3156</td>
<td>7.1604</td>
<td>53.1</td>
<td>45.3</td>
</tr>
<tr>
<td></td>
<td>Special</td>
<td>0.3493</td>
<td>7.0973</td>
<td>50.8</td>
<td>43.0</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional</td>
<td>Supplied</td>
<td>0.3499</td>
<td>7.5199</td>
<td>51.0</td>
<td>43.3</td>
</tr>
<tr>
<td></td>
<td>Special</td>
<td>0.3514</td>
<td>7.3958</td>
<td>49.7</td>
<td>42.0</td>
</tr>
<tr>
<td></td>
<td>Full</td>
<td>0.3490</td>
<td>6.6185</td>
<td>51.9</td>
<td>44.5</td>
</tr>
<tr>
<td>Unrestricted</td>
<td>Full</td>
<td>0.3774</td>
<td>7.0880</td>
<td>50.0</td>
<td>43.4</td>
</tr>
<tr>
<td></td>
<td>Special</td>
<td>0.4118</td>
<td>7.0908</td>
<td>47.7</td>
<td>41.0</td>
</tr>
<tr>
<td></td>
<td>Full</td>
<td>0.4409</td>
<td>7.2413</td>
<td>45.7</td>
<td>39.3</td>
</tr>
</tbody>
</table>

(∗) marked for subjective evaluation
## Official Results: Subjective Scores

<table>
<thead>
<tr>
<th>Data Condition</th>
<th>Segmentation</th>
<th>Fluency</th>
<th>Adequacy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Test</td>
<td></td>
</tr>
<tr>
<td>Supplied</td>
<td>Special</td>
<td>Special</td>
<td>3.120</td>
</tr>
<tr>
<td>Additional</td>
<td>Supplied</td>
<td>Special</td>
<td>3.256</td>
</tr>
<tr>
<td>Unrestricted</td>
<td>Full</td>
<td>Full</td>
<td>3.776</td>
</tr>
</tbody>
</table>
THE END
Decoding: Expansion, Recombination and Pruning