The ITC-irst SMT System for IWSLT-2005

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

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica
38050 Povo (Trento), Italy
Log-Linear Model Approach to SMT

Maximum Entropy framework for the word-alignment MT approach:

\[
    e^* = \arg \max_e \max_a \Pr(e, a \mid f) = \arg \max_e \max_a \sum_i \lambda_i h_i(e, f, a) \quad (1)
\]

where \( f = \text{source} \), \( e = \text{target} \), \( a = \text{alignment} \), and \( h_i(e, f, a) \) are suitable feature functions.

Advantages:

- directly models the posterior probability (discriminative model)
- does not rely on probability factorizations with independence assumptions
- is mathematically sound and allows to add any kind of feature function
- includes any IBM model as a special case
- minimum error training to estimate free parameters \((\lambda_i)\)
Phrase-based Model

- A phrase is a sequence of one or more words without semantic/syntactic meaning.

Generative process:
1. cover new source positions (distortion)
2. link to target phrase (fertility, lexicon)
3. add target phrase (language model)
4. untranslated words ($\tilde{e}_0$-fertility, lexicon)

Search is over strings of phrases:

$$\tilde{\mathbf{e}}^* = \arg \max_{\mathbf{e}} \max_{\mathbf{a}} \sum_{i} \lambda_i h_i(\tilde{\mathbf{e}}, \mathbf{f}, \mathbf{a})$$
Two Pass Search Strategy

First Pass:

- Log-linear Model
- Dynamic programming algorithm
- Beam search decoder:
  - threshold and histogram pruning
- Non-monotone search constraints
  - max number of vacancies on the left (MVN)
  - max distance from left-most vacancy (MVD)

Second Pass:

- Extraction of 1,000-best
- Log-linear Model
- Re-ranking algorithm
Two Pass Search Strategy

First Pass feature functions:

- Target 3-gram LM
- Fertility model target phrases
- Direct phrase-based lexicon
- Inverse phrase-based lexicon
- Negative distortion
- Positive distortion
- $\tilde{e}_0$ fertility
- $\tilde{e}_0$ permutation
Training of Phrase-based model

Phrase-based model (baseline):

- Word-alignment: union of direct and inverse IBM alignments (GIZA++, \(1^5 H^5 3^4 4^4 5^4\))

- Phrase-extraction: max length 8, filtering (length or punctuation mismatches)

- Feature estimation: lexicon, fertility models (… by freq smoothing …)

- Monotone search: MVD=0

Improvements by exploiting Competitive Linking Algorithm (Melamed, 2000):

- **CLA translation lexicon** added to data before word-alignment

- **CLA word-alignments** added to IBM word alignments before phrase-extraction

- **Re-segmented Chi/Jap data** added to training data before word-alignment (in-house tool)
Experimental Results: First Pass

- Task: Supplied Data Condition
- Lang: Chinese, Japanese, Arabic
- Test set: IWSLT 2004
- Dev set: CSTAR 2003
- BLEU%: no-case with punctuation
- No weight optimization
- Non-monotone search:
  - MVD=4 MVN=3 Arabic
  - MVD=6 MVN=5 Chinese
  - MVD=7 MVN=6 Japanese

<table>
<thead>
<tr>
<th>System</th>
<th>Chi2Eng</th>
<th>Jap2Eng</th>
<th>Ara2Eng</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>35.82</td>
<td>33.82</td>
<td>51.01</td>
</tr>
<tr>
<td>+CLA translation lexicon</td>
<td>36.28</td>
<td>35.78</td>
<td>52.84</td>
</tr>
<tr>
<td>+CLA alignments</td>
<td>37.59</td>
<td>38.77</td>
<td>54.14</td>
</tr>
<tr>
<td>+re-segmented data</td>
<td>38.29</td>
<td>38.97</td>
<td>–</td>
</tr>
<tr>
<td>+chunked data</td>
<td>–</td>
<td>39.59</td>
<td>–</td>
</tr>
<tr>
<td>+non-monotone search</td>
<td>42.51</td>
<td>44.66</td>
<td>56.40</td>
</tr>
</tbody>
</table>
CLA alignments vs. IBM Alignments

- IBM alignments are many-to-one
- CLA alignments are one-to-one
- CLA alignments have higher precision
- CLA alignments allow for more phrase-pairs

Despite past work (Och & Ney, 2003) showed that quality of CLA alignments is poorer than for IBM Model 1, we found that such alignments work indeed well for phrase-based SMT.
Phrase extraction from IBM and CLA alignments

In this real example, the CLA alignment allows to extract the useful phrase “where is”.

Two Pass Search Strategy

source string

decoder

WG

N-best

Rescoring

target string

☐ : extractor (from WG)

Second Pass feature functions:

– IBM model 1 lexicon score
– IBM model 3 lexicon score
– CLA lexicon score
– Question feature
– Frequency of n-grams within n-best
– ratio of target source lengths
– 2-gram target LM
– 4-gram target LM
– 5-gram target LM
New Feature Functions in Re-scoring

The following statistics are computed on each entry of the 1000-best list:

- **CLA alignment score**
  Integrates the CLA associative score over all possible word alignments between source and target, similarly to how is done for IBM Model 1 re-scoring

- **Question tag**
  Triggers a binary feature when the string ends with a question mark and starts with one of the following words: what, which, who, when, how, do, did, ...

- **N-gram frequency**
  Counts the frequencies of its n-grams (n=1,2,3,4) within the full n-best list and sums them up according to a linear combination.
Experimental Results: Re-scoring Stage

- Task: Supplied Data Condition
- Lang: Chinese, Japanese, Arabic
- BLEU%: no-case with punctuation
- Test set: IWSLT 2004
- Dev set: CSTAR 2003
- Optimization: BLEU% + 4 * NIST
- N-best 1000

<table>
<thead>
<tr>
<th>System</th>
<th>Chi2Eng</th>
<th>Jap2Eng</th>
<th>Ara2Eng</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decoder</td>
<td>42.51</td>
<td>44.66</td>
<td>56.40</td>
</tr>
<tr>
<td>IBM model-1</td>
<td>42.31</td>
<td>44.48</td>
<td>56.00</td>
</tr>
<tr>
<td>IBM model-3</td>
<td>41.53</td>
<td>44.97</td>
<td>56.16</td>
</tr>
<tr>
<td>CLA score</td>
<td>42.42</td>
<td>45.20</td>
<td>56.31</td>
</tr>
<tr>
<td>question tag</td>
<td>42.81</td>
<td>45.83</td>
<td>56.66</td>
</tr>
<tr>
<td>n-grams</td>
<td>43.71</td>
<td>46.19</td>
<td>56.89</td>
</tr>
<tr>
<td>target length</td>
<td>41.11</td>
<td>41.00</td>
<td>50.87</td>
</tr>
<tr>
<td>2-grams LM</td>
<td>44.06</td>
<td>45.34</td>
<td>56.07</td>
</tr>
<tr>
<td>4-grams LM</td>
<td>45.88</td>
<td>45.51</td>
<td>56.72</td>
</tr>
<tr>
<td>5-grams LM</td>
<td>45.72</td>
<td>45.81</td>
<td>56.61</td>
</tr>
<tr>
<td>+all features</td>
<td>47.99</td>
<td>51.01</td>
<td>57.94</td>
</tr>
</tbody>
</table>
Conclusions

Main performance improvements came from:

- **Integration of IBM and CLA word-alignments** at different levels:
  - Translation lexicon used to constrain IBM alignments
  - Phrase-extraction performed on both CLA and IBM word-alignments

- **Use of multiple word segmentations** for Chinese, Japanese

- **New feature functions** used for n-best re-scoring:
  - Associative score from CLA
  - Frequency of n-grams in n-best list
  - High order language models (4-gram 5-gram)

- **Optimization of non-monotone search constraints**
The End ... Thank You!