Statistical Machine Translation of Euparl data by using Bilingual N-grams

Rafael E. Banchs
Josep M. Crego
Adrià de Gispert
Patrik Lambert
José B. Mariño
Outline

- Overview
- Bilingual N-gram translation model
- Training from parallel corpus
- Additional feature functions
  - Target LM and word penalty
  - IBM model1 lexicon model
- Decoder
- Experiments and ACL'05 Shared Task results
- Conclusions
Overview

- Log-linear combination of multiple Statistical Models
  - Bilingual N-gram translation model
  - Target Language Model
  - Word Penalty
  - IBM 1 lexicon model (src → trg, trg → src)
- Derived from maximum entropy approach

\[
\hat{t}_1^I = \arg \max_{t'_1^I} \sum_{m=1}^{M} \lambda_m h_m(t'_1^I, s'_1^J)
\]

- Weights optimization on development set (simplex alg.)
Bilingual N-gram model

- Standard N-gram of bilingual units (tuples)

\[
h_{TM}(t,s) = \log \prod_{n=1}^{N} p((t,s)_n \mid (t,s)_{n-1}, (t,s)_{n-2})
\]

\[
(t,s)_n = (t_{i(n)}...t_{i(n)+I(n)}, s_{j(n)}...s_{j(n)+J(n)})
\]

Training

- Parallel Corpus (preproc.)
- Word alignment
- Symm.
- Tuple extraction
- N-gram estimation

SRILM LM toolkit

GIZA++ union, intersection

feature leading the translation
Tuple extraction

- Tuples are bilingual units with
  - one / more source words
  - zero / one / more target words

- Conditions of extraction:
  1. Monotonic segmentation of the bilingual pair
  2. The tuple cannot be decomposed into smaller units without violating 1

Subset of phrases, unique under these conditions

- Details:

Example
Tuple extraction

- Tuples are bilingual units with
  - one / more source words
  - zero / one / more target words

- Conditions of extraction:
  1. Monotonic segmentation of the bilingual pair
  2. The tuple cannot be decomposed into smaller units without violating 1

Subset of phrases, unique under these conditions

- Details:
  - **Source NULL** words (not allowed)

Example:

```
(f_1 f_2, e_1) (f_3, e_2 e_3 e_4) (f_4 f_5 f_6, e_5 e_6)
```
Tuple extraction

- Tuples are bilingual units with
  - one / more source words
  - zero / one / more target words
- Conditions of extraction:
  1. Monotonic segmentation of the bilingual pair
  2. The tuple cannot be decomposed into smaller units without violating 1

Subset of phrases, unique under these conditions

- Details:
  - **Source NULL** words (not allowed)
  - **Embedded words** (unigram dictionary dictionary is extracted)
Additional feature models (1)

- **N-gram Target Language Model**

  $3\text{-gram}$
  $$h_{LM}(t,s) = h_{LM}(t) = \log \prod_{n=1}^{I} p(t_i | t_{i-1}, t_{i-2})$$

- **Word Penalty**

  Compensates the LM preference for short translations
  $$h_{WP}(t,s) = I$$

**SRILM LM toolkit**
Additional feature models

- IBM 1 lexicon model

\[
h_{IBM1}(t,s)_n = \log \frac{1}{(I'+1)^{J'}} \prod_{j=1}^{J'} \sum_{i=0}^{I'} p(t_i | s_j)
\]

- IBM 1 inverse lexicon model

\[
h_{IBM-1}(t,s)_n = \log \frac{1}{(J'+1)^{I'}} \prod_{i=1}^{I'} \sum_{j=0}^{J'} p(s_j | t_i)
\]

Special Case: certain pairs may not be represented for both directions (considered equal)
MARIE Decoder

- Beam-search strategy based on dynamic programming
- Takes the previous 5 models into account
- Pruning methods
  - Threshold pruning
  - Histogram pruning
  - Hypothesis recombination
- **Monotone search** was used for all EuParl experiments
- Weights of each feature optimized according to BLEU
  - Based on simplex algorithm
Experiments: Preprocessing

- All pairs  *sent pairs ( > 100 words ) or ( ratio > 2.4 ) excluded*

- French-English  *apostrophes re-tokenised*
  - _[cdjlmnst]_' → [cdjlmnst]
  - (qu / jusqu / lorsqu / quelqu / puisqu / quoiqu / presqu)_' → $1'$

- Standard GIZA++ alignments  $^{15}H^53^34^3$

<table>
<thead>
<tr>
<th>src lang</th>
<th># words</th>
<th># vocab</th>
<th>tuple vocab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>15.67 M</td>
<td>113.5 K</td>
<td>1.28 M</td>
</tr>
<tr>
<td>French</td>
<td>14.84 M</td>
<td>78.4 K</td>
<td>1.17 M</td>
</tr>
<tr>
<td>German</td>
<td>15.20 M</td>
<td>204.9 K</td>
<td>1.39 M</td>
</tr>
<tr>
<td>Finnish</td>
<td>11.22 M</td>
<td>389.2 K</td>
<td>1.49 M</td>
</tr>
</tbody>
</table>

+ *sparseness*
ACL'05 Workshop Results

- Results on TEST set

<table>
<thead>
<tr>
<th>lang pair</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>es – en</td>
<td>0.3007</td>
</tr>
<tr>
<td>fr – en</td>
<td>0.3020</td>
</tr>
<tr>
<td>de – en</td>
<td>0.2426</td>
</tr>
<tr>
<td>fi – en</td>
<td>0.2031</td>
</tr>
</tbody>
</table>

- Translation quality is highly dependent on vocab. size
- Finnish and especially German translations suffer from bad order

Whether data sparseness or reordering is more critical has not been explored yet
ACL'05 Workshop Results (2)

- Results on DEVELOPMENT set

- Baseline (only bilingual N-gram model) vs Full (5 model combination)

<table>
<thead>
<tr>
<th>lang pair</th>
<th>Baseline</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>es – en</td>
<td>0.2588</td>
<td>0.3004</td>
</tr>
<tr>
<td>fr – en</td>
<td>0.2547</td>
<td>0.2938</td>
</tr>
<tr>
<td>de – en</td>
<td>0.1844</td>
<td>0.2350</td>
</tr>
<tr>
<td>fi – en</td>
<td>0.1526</td>
<td>0.1989</td>
</tr>
</tbody>
</table>

- Significant improvements are produced by the additional features

*The separate contribution of each feature model is being explored*
Conclusions

- Statistical MT based on a log-linear combination of 5 models
  - Bilingual N-gram of tuples as translation model
  - Target LM and Word Penalty
  - IBM model 1 (forward and inverse)
- **Monotone** beam-search decoding (dyn. programming)
- Further research
  - More experimentation on impact of separate feature models
  - Reordering techniques
Thanks for attention

Questions or comments are welcome ...