The ISI/USC MT System for IWSLT 2004

Ignacio Thayer, Emil Ettelaie, Kevin Knight, Daniel Marcu, Dragos Stefan Munteanu, Franz Joseph Och*, Quamrul Tipu

* Now at Google, Inc.
Overview

• ISI/USC MT System
  – Overview
  – Model components
    • Simpler version of 2004 NIST Evaluation System
  – Training data

• Results
MT as Noisy Channel

• Translate source sentence $f$ into target sentence $e$

• Noisy Channel
  – $P(e)$ - language model
  – $P(f|e)$ - translation model
  – $P(e|f) = P(e)P(f|e)/P(f)$

• Translation is search
  – $\text{argmax}_e P(e|f) = \text{argmax}_e P(e)P(f|e)$
Log-Linear Model

- Translate source sentence $f$ into target sentence $e$
- **Direct Model**
  - Feature functions $h_m(e, f)$
  - Feature weight $\lambda_m$
  - $P(e|f) = \exp(\sum M \lambda_m h_m(e, f)) * Z(f)$
- **Translation is search**
  - $\arg\max_e P(e|f) = \arg\max_e \sum M \lambda_m h_m(e, f)$
Log-Linear Model

MT = feature function combination

Chinese sentence → Trigram LM, Phrase Lexicon → MT → English sentence

Word Lexicon, Reordering, ...
Training

• Feature functions trained individually
  – Specific training criterion for each FF
    • Phrase Probability: Relative Frequency
    • Language Model: Smoothed ML
    • ...

• Feature function weights are optimized to increase BLEU score
Minimum Error Rate Training

Translate Development Corpus

Measure BLEU Score

Update Model Weights To Reduce Translation Error

Alignment Template Model

• Corpus is word aligned
  – Uni-directional word alignments are merged

• Phrase pairs are collected
  – A phrase is only collected if words on both sides are only aligned to each other

• Probability determined by relative frequency
  – \( p(e|f) = \frac{C(e,f)}{C(f)} \)
Language Model

• Smoothed trigram
  – Kneser-Ney smoothing

• SRI Language Modelling Toolkit
Other Feature Functions

• 10 other feature functions used for scoring
  – Length Bonus - encourage longer sentences
  – Jump Penalty - discourage non-monotonicity
  – ...
  – Full list in paper

• Fewer feature functions that NIST 2004 system
Search

• Dynamic programming beam-search
• Generate translation hypothesis word-by-word
• Heuristic rest-cost estimate
• Reordering constraints:
  – < 8 word jumps
Training Data - Supplied

- 20K lines BTEC corpus J-to-E, C-to-E
- LM trained on English half
Training Data - Additional

- 20K lines BTEC corpus C-to-E (x5)
  - Re-segmented with LDC segmenter
- 6 of allowable LDC corpora
- LM trained on English half
- LM trained on 800M words news text
- Punctuation removal
  - No other rule-based translations/postprocessing
Training Data - Unrestricted

- 20K lines BTEC corpus C-to-E (x5)
- 167M words political+news data (NIST eval corpora)
- LM trained on English half
- LM trained on 800M words news text
- Punctuation removal
- No minimum error training
  - Model weights from “Additional” system were used.
## BLEU Results

<table>
<thead>
<tr>
<th></th>
<th>C-to-E</th>
<th>J-to-E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplied</td>
<td>37.42</td>
<td>40.08</td>
</tr>
<tr>
<td>Additional</td>
<td>44.05*</td>
<td>N/S</td>
</tr>
<tr>
<td>Unrestricted</td>
<td>24.3**</td>
<td>N/S</td>
</tr>
</tbody>
</table>

* previously reported as 31.16

** no minimum-error rate training
Conclusion

• Applied our translation system to speech expressions
• Excited to learn more about spoken-language translation