The University of Washington Machine Translation System for the IWSLT 2007 Competition

Katrin Kirchhoff & Mei Yang

Department of Electrical Engineering
University of Washington, Seattle, USA
Overview

- Two systems, I-EN and AR-EN
- Main focus on exploring use of out-of-domain data and Arabic word segmentation

- Basic MT System
- Italian-English system
  - Out-of-domain data experiments
- Arabic-English system
  - Semi-supervised Arabic segmentation
Basic MT System

- Phrase-based SMT system
- Translation model: log-linear model
  \[ e^* = \arg \max_e p(e \mid f) = \arg \max_e \left\{ \sum_{k=1}^{K} \lambda_k \phi_k(e, f) \right\} \]
- Feature functions:
  - 2 phrasal translation probabilities
  - 2 lexical translation scores
  - Word count penalty
  - Phrase count penalty
  - LM score
  - Distortion score
  - Data source feature
Basic MT System

- Word alignment
  - HMM-based word-to-phrase alignment [Deng & Byrne 2005] as implemented in MTTK
  - Comparable in performance to GIZA++
- Phrase extraction:
  - Method by [Och & Ney 2003]
Basic MT System

- **Decoding/Rescoring:**
  - Minimum-error rate training to optimize weights for first pass (optimization criterion: BLEU)
  - MOSES decoder
  - First pass: up to 2000 hypotheses/sentence
  - Additional features for rescoring
    - POS-based language model
    - Rank-based feature [Kirchhoff et al, IWSLT06]
  - → Final 1-best hypothesis
Basic MT System

- Postprocessing:
  - Hidden-ngram model [Stolcke & Shriberg 1996] for punctuation restoration:
    \[
    P(e_1, ..., e_T) \approx \prod_{t=1}^{T} P(e_t \mid e_{t-1}, ..., e_{t-n+1})
    \]
    Event set E: words and punctuation signs

- Noisy-channel model for truecasing
  - SRILM *disambig* tool
Basic MT System

- Spoken-language input:
  - IWSLT 06: mixed results using confusion network input for spoken language
  - No specific provisions for ASR output this year!
Italian-English System
Data Resources

- **Training:**
  - BTEC (all data sets from previous years, 160k words)
  - Europarl Corpus (parliamentary proceedings, 17M words)

- **Development:**
  - IWSLT 2007 (SITAL) dev set (development only, no training on this set)
  - Split into dev set (500 sentences) and held-out set (496 sentences)

- BTEC name list
Preprocessing

- Sentence segmentation into smaller chunks based on punctuation marks
- Punctuation removal and lowercasing
- Automatic re-tokenization on English and Italian side:
  - For N most frequent many-to-one alignments (N = 20), merge multiply aligned words into one
    - E.g. *per piacere* – *please* → *per_piacere*
  - Improves noisy word alignments
Translation of names/numbers

- Names in the BTEC name list were translated according to list, not statistically
- Small set of hand-coded rules to translate dates and times
Out-of-vocabulary words

- To translate OOV words, map OOV forms to all words in training data that do not differ in length by more than 2 characters
- Mapping done by string alignment
- For all training words with edit distance < 2, reduplicate phrase table entries with word replaced by OOV form
- Best-matching entry chosen during decoding
- Captures misspelled or spoken-language specific forms: *senz’*, *undic’*, *quant’*, etc.
Using out-of-domain data

- Experience in IWSLT 2006:
  - additional English data for language model did not help
  - out-of-domain data for translation model (IE) was useful for text but not for ASR output
- Adding data:
  - second phrase table trained from Europarl corpus
  - Use phrase tables from different sources jointly
  - each entry has *data source feature*: binary feature indicating corpus it came from
Using out-of-domain data

- Phrase coverage (%) and translation results on IE 2007 dev set

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>BLEU/PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTEC</td>
<td>78.3</td>
<td>29.6</td>
<td>6.7</td>
<td>1.3</td>
<td>0.2</td>
<td>18.9/55.1</td>
</tr>
<tr>
<td>Europarl</td>
<td>83.9</td>
<td>37.0</td>
<td>6.4</td>
<td>0.7</td>
<td>0.1</td>
<td>18.5/55.4</td>
</tr>
<tr>
<td>combined</td>
<td>85.8</td>
<td>39.9</td>
<td>9.4</td>
<td>1.7</td>
<td>0.2</td>
<td>20.7/53.5</td>
</tr>
</tbody>
</table>
## Using out-of-domain data

Diagnostic experiments: importance of matched data vs. data size

<table>
<thead>
<tr>
<th>Training set</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>500 sentences from SITAL</td>
<td>Small amount of matched data</td>
</tr>
<tr>
<td>BTEC training corpus</td>
<td>Moderate amount of domain-related by stylistically different data</td>
</tr>
<tr>
<td>Europarl</td>
<td>Large amount of mismatched data</td>
</tr>
<tr>
<td>BTEC &amp; Europarl</td>
<td>Combination of sources</td>
</tr>
<tr>
<td>BTEC &amp; Europarl &amp; SITAL</td>
<td></td>
</tr>
</tbody>
</table>
Using out-of-domain data

Performance on held-out SITAL data (%BLEU/PER)

<table>
<thead>
<tr>
<th>Training set</th>
<th>Text</th>
<th>ASR output</th>
</tr>
</thead>
<tbody>
<tr>
<td>500 sentences from SITAL</td>
<td>28.0/46.8</td>
<td>26.1/49.0</td>
</tr>
<tr>
<td>BTEC training corpus</td>
<td>18.9/55.1</td>
<td>16.6/57.1</td>
</tr>
<tr>
<td>Europarl</td>
<td>18.5/55.4</td>
<td>17.3/56.4</td>
</tr>
<tr>
<td>BTEC &amp; Europarl</td>
<td>20.7/53.3</td>
<td>18.6/55.3</td>
</tr>
<tr>
<td>BTEC &amp; Europarl &amp; SITAL</td>
<td>30.1/41.9</td>
<td>27.7/44.8</td>
</tr>
</tbody>
</table>
## Effect of different techniques

<table>
<thead>
<tr>
<th>Technique</th>
<th>BLEU(%)/PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>18.9/55.1</td>
</tr>
<tr>
<td>OOD data</td>
<td>20.7/53.4</td>
</tr>
<tr>
<td>Rescoring</td>
<td>22.0/52.7</td>
</tr>
<tr>
<td>Rule-based number trans.</td>
<td>22.6/52.2</td>
</tr>
<tr>
<td>Postprocessing</td>
<td>21.2/50.4</td>
</tr>
</tbody>
</table>
Arabic-English System
Data Resources

• Training
  • BTEC data, without dev4 and dev5 sets (160k words)
  • LDC parallel text (Newswire, MTA, ISI) (5.5M words)

• Development
  • BTEC dev4 + dev5 sets

• Buckwalter stemmer
Preprocessing

- Chunking based on punctuation marks
- Conversion to Buckwalter format
- Tokenization
  - Linguistic tokenization
  - Semi-supervised tokenization
Linguistic Tokenization

• Columbia University MADA/TOKAN tools:
  • Buckwalter stemmer to suggest different morphological analyses for each word
  • Statistical disambiguation based on context
  • Splitting off of word-initial particles and definite article
• - Involves much human labour
• - difficult to port to new dialects/languages
Semi-supervised tokenization

- Based on [Yang et al. 2007]: segmentation for dialectal Arabic

1. Affix list + set of segmented words
2. Initial segmentation
3. Unsegmented text
4. New list of segmented words and stems
Semi-supervised tokenization

- Needs fewer initial resources
- Produces potentially more consistent segmentations
- Once trained, much faster than linguistic tools

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU/PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>MADA/TOKAN</td>
<td>22.5/50.5</td>
</tr>
<tr>
<td>SemiSup – initialized on MSA</td>
<td>23.0/50.7</td>
</tr>
<tr>
<td>SemiSup – initialized on Iraqi Arabic</td>
<td>21.6/51.1</td>
</tr>
</tbody>
</table>
**Effect of out-of-domain data**

Phrase coverage rate (%) and translation performance on dev5 set, clean text

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>BLEU/PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTEC</td>
<td>69.3</td>
<td>34.3</td>
<td>12.9</td>
<td>3.6</td>
<td>1.2</td>
<td>0.3</td>
<td>22.5/50.5</td>
</tr>
<tr>
<td>News</td>
<td>60.2</td>
<td>30.4</td>
<td>11.5</td>
<td>2.7</td>
<td>0.9</td>
<td>0.2</td>
<td>----</td>
</tr>
<tr>
<td>combined</td>
<td>82.6</td>
<td>46.7</td>
<td>20.1</td>
<td>5.6</td>
<td>1.9</td>
<td>0.5</td>
<td>24.6/47.5</td>
</tr>
</tbody>
</table>
# Effect of different techniques

<table>
<thead>
<tr>
<th>Technique</th>
<th>BLEU(%)/PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>22.5/50.5</td>
</tr>
<tr>
<td>OOD data</td>
<td>24.6/48.2</td>
</tr>
<tr>
<td>Rescoring</td>
<td>24.6/47.5</td>
</tr>
<tr>
<td>Postprocessing</td>
<td>23.4/48.5</td>
</tr>
</tbody>
</table>

Dev5 set, clean text
## Evaluation results

Performance on eval 2007 sets

<table>
<thead>
<tr>
<th></th>
<th>BLEU(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IE – clean text</td>
<td>26.5</td>
</tr>
<tr>
<td>IE – ASR output</td>
<td>25.4</td>
</tr>
<tr>
<td>AE – clean text</td>
<td>41.6</td>
</tr>
<tr>
<td>AE – ASR output</td>
<td>40.9</td>
</tr>
</tbody>
</table>
Conclusions

- Adding out-of-domain data helped in both clean text and ASR conditions
- Importance of stylistically matched data
- Semi-supervised word segmentation for Arabic comparable to supervised segmentation, uses fewer resources
- Cross-dialectal bootstrapping of segmenter possible