The CMU-UKA Statistical Machine Translation Systems for IWSLT 2007

Ian Lane, Andreas Zollmann, Thuy Linh Nguyen, Nguyen Bach, Ashish Venugopal, Stephan Vogel, Kay Rottmann, Ying Zhang, Alex Waibel
Overview

- Overview of submission systems
- Research Topics Investigated
  - Topic-Aware Spoken Language Translation
  - Morphological-Decomposition for Arabic SMT
  - Comparison of Punctuation-Recovery Approaches
Submission Systems
Submission Systems (“diversity”)

- Submissions made for three language pairs
- All systems based on phrase-based SMT
- Each language-pair focused on specific research area

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>System Description</th>
<th>Rank (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japanese → English</td>
<td>SMT with Punctuation Recovery / Topic-based N-best-list rescoring</td>
<td>1</td>
</tr>
<tr>
<td>Chinese → English</td>
<td>Syntax Augmented SMT</td>
<td>3</td>
</tr>
<tr>
<td>Arabic → English</td>
<td>SMT with Morphological Decomposition</td>
<td>7</td>
</tr>
</tbody>
</table>

(1) Spoken language translation task - ASR (BLEU)
Japanese Submission System

<table>
<thead>
<tr>
<th>Training Corpora</th>
<th>IWSLT-training, IWSLT-dev1-3, Tanaka</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpora-size</td>
<td>200k sentence pairs, 2M words</td>
</tr>
<tr>
<td>Phrase-Extraction</td>
<td>PESA [Vogel05]</td>
</tr>
<tr>
<td>LMs</td>
<td>6-gram SA-LM</td>
</tr>
<tr>
<td></td>
<td>4-gram interpolated n-gram LM</td>
</tr>
<tr>
<td>Reordering Window</td>
<td>6</td>
</tr>
<tr>
<td>Decoder</td>
<td>STTK (phrase-based SMT) [Vogel03]</td>
</tr>
</tbody>
</table>

- Punctuation estimated on source-side via HELM
- N-best candidates rescored: Topic-Confidence Scores
Chinese Submission System

<table>
<thead>
<tr>
<th>Training Corpora</th>
<th>IWSLT-training, IWSLT-dev1-3,5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpora-size</td>
<td>67k sentence pairs</td>
</tr>
<tr>
<td>Rule-Extraction</td>
<td>Giza++, Pharaoh, Stanford-parser</td>
</tr>
<tr>
<td>Decoder</td>
<td>SAMT [<a href="http://www.cs.cmu.edu/~zollmann/samt">www.cs.cmu.edu/~zollmann/samt</a>]</td>
</tr>
</tbody>
</table>

- Identical to IWSLT 2006 submission system
  - Improved efficiency and robustness decoder
    “to handle GALE size data”
  - Slight increase in training data
- See IWSLT 2006 paper for detailed system description
### Arabic Submission System

<table>
<thead>
<tr>
<th>Training Corpora</th>
<th>IWSLT-training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpora-size</td>
<td>20k sentence pairs</td>
</tr>
<tr>
<td>Phrase-Extraction</td>
<td>Giza++, Pharoah</td>
</tr>
<tr>
<td>Decoder</td>
<td>STTK (phrase-based SMT) [Vogel03]</td>
</tr>
</tbody>
</table>

- **Morphological decomposition** performed using [Smith05]
- **30% of morphemes discarded** to obtain source/target ratio close to 1
Research Topics

- **Topic-aware SLT**
  - Apply utterance-level topic constraints for SLT

- **Morphological-Decomposition for Arabic SMT**
  - Decompose Arabic words into morphemes
  - Discard “un-necessary” morphemes before translation

- Comparison of Punctuation Recovery Techniques
  (described in paper)
Topic-aware SLT
Topic-aware SLT

• Previous work have focused on document level adaptation for translation of monologue data
  • Bi-LSA: Adaptation of Target-LM [Tam07]
  • Adaptation of IBM-1 Lexicon [Tam07]
  • Bi-TAM: Incorporate *topic* during alignment [Zhao06]

• Investigate approach, appropriate for spoken dialogue (applicable to small training corpora)

• Apply independently to each utterance
Topic-aware SLT

• Apply topic-constraints within SLT
  → Detect topic of discourse and apply topic-constraints during translation

• Investigate two additional feature-functions
  • Topic-Dependent LM Score
  • Topic-Confidence Scores

• Rescore N-best trans. candidates incorporating above scores
Description of Scores

**Topic-Dependent LM Score**
- Topic-specific LM should better discriminate between acceptable and bad translations
- Add additional Topic-Dependent LM score

**Topic Confidence Score**
- No constraint to maintain topic consistency within translation hypothesis
- Visual inspecting identified the following:
  - “Good” translation hypotheses typically obtained high topic-confidence score (for a single topic)
  - “Bad” translations typically obtained low-confidence scores for all topics
Topic-Dependent LM Scores

1. Select topic of utterance by 1-best hypo.
2. Generate additional score by applying TD-LM to each hypothesis
3. Re-rank N-best hypotheses based on log-lin. $\Sigma$ model scores
1. Calculate topic confidence score [0,1] for each topic class
2. Re-rank N-best hypotheses based on log-lin. Σ model scores
   (features used during decoding (10) + $M$ topic confidence scores)
Experimental Evaluation

• **Topic Class Definition**
  - Training corpora split into eight classes
    - Hierarchical clustering, minimize global perplexity

• **Topic Models**
  - SVM classification models trained for each class
    - **Features:** word, word-pairs and 3-grams
  - TD-LMs trained for each topic class

• **Tuning / Evaluation Sets**
  - MERT Set: IWSLT06-dev.
  - Eval. Set: IWSLT06-eval, IWSLT07-eval
Effectiveness on ’06 Eval. Set

- **Baseline**: JE phrase-based SMT system (described earlier)

**TDLM**: Topic-Dependent LMs

**TC**: Topic Confidence Scores

- Both TDLM and TC feature sets improve translation performance (0.0022 and 0.0027 BLEU-points respectively)

→ Use Topic-Confidence scores in submission system
Effectiveness on ‘07 Eval. Set

**TDLM**: Topic-Dependent LMs  
**TC**: Topic Confidence Scores

- Slight degradation in BLEU-score on 2007 Evaluation-Set (0.4990 → 0.4828)
- ‘06 Eval.-set typically contained multiple sentences per utterance
  → Maintains topic-consistency between sentences (mismatch with ‘07 Eval. set)
Morphological-Decomposition for Arabic SMT
Morphological-Decomposition for Arabic SMT

- Traditional word-alignment models assume similar number of source/target tokens
- For diverse language-pairs significant mismatch
  - Highly agglomerative language (Arabic)
  - Non-agglomerative language (English)
- Decompose Arabic words into prefix/stem/suffix morphemes
  - Also improve translation coverage
- Able to translate unseen Arabic words at Morpheme-level
Morphological-Decomposition for Arabic SMT

- Prefix / stem / suffix of an Arabic word often corresponds to individual English word

  **Prefix:**
  - conjunction: *wa* → and
  - article: *Al* → the
  - preposition: *li* → to/for

  **Suffix:**
  - Pronoun: *hm* → their/them

- Some specific morphemes are redundant in A→E trans. → can be discarded during translation

  **Suffix:**
  - Gender: *f* → female singular
  - Case marker, number, voice, etc..
Proposed Approach

• Previous works [Habash06] used manually defined rules to remove inflectional features before translation

• Data driven approach to discard non-informative morphemes

1. Perform full morphological decomposition on Arabic-side
2. Align training corpora: Arabic morpheme-level / English word-level
3. Discard morphemes with zero-fertility > $\theta_{th}$
   Morphemes not aligned to any English word $\rightarrow$ high zero-fertility
   Morphemes typically aligned to a English word $\rightarrow$ low zero-fertility
Experimental Evaluation

• **Topic Class Definition**
  • Training corpora split into eight classes
    • Hierarchical clustering, minimize global perplexity []

• **Topic Models**
  • SVM classification models trained for each class
    • **Features**: word, word-pairs and 3-grams
  • TD-LMs trained for each topic class

• **Tuning / Evaluation Sets**
  • MERT Set: IWSLT06-dev.
  • Eval. Set: IWSLT06-eval, IWSLT07-eval
Morpheme Removal (fertility)

- From 158k Arabic wrds obtain 294k morph. (190k English wrds)
- Manually set $\theta_{th}$ to discard 40% of morphemes

- Discarding morphemes with high zero-fertility normalizes source/target ratio
- Shifts fertility peak > 1.0
Morpheme Removal (Trans. Quality)

- Manually set $\theta_{th}$ to discard $\%$ of morphemes

- Discarding 30-40% of morphemes obtains highest BLEU score
- Improved BLEU 0.5573 $\rightarrow$ 0.5631 (IWSLT05 held-out eval.-set)
Conclusions
Conclusions

- Developed evaluation systems for 3 language-pairs
  - Each language-pair focused on specific research topic

- **Punctuation Recovery**
  - Best performance obtained with source-side HELM estimation

- **Topic-aware SLT**
  - Significant improvement in performance obtained for multi-sentence utterances (IWSLT 2006 evaluation set)
  - Topic-Classification Scores more effective than TD-LM

- **Morphological- Decomposition for Arabic SMT**
  - Improved BLEU by applying morphological decomposition and discarding 30% morphemes with highest zero-fertility
Thank you
Other Slides
Punctuation Recovery for SLT
## Punctuation Recovery for SLT

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>97.8%</td>
<td>96.8%</td>
<td>97.3%</td>
</tr>
<tr>
<td>82.1%</td>
<td>44.2%</td>
<td>57.5%</td>
</tr>
<tr>
<td>96.4%</td>
<td>95.9%</td>
<td>96.2%</td>
</tr>
<tr>
<td>71.8%</td>
<td>43.6%</td>
<td>54.3%</td>
</tr>
<tr>
<td>100%</td>
<td>63.9%</td>
<td>77.9%</td>
</tr>
</tbody>
</table>
Topic-aware SLT

![Bar Chart](Image)

- **Manual Transcriptions**
- **1-best ASR Hypothesis**

**Punctuation Recovery Approach**

- **Source**
- **Target**
- **SMT**
- **Manual**

**BLEU score (lower case)**

- 0.2000
- 0.2100
- 0.2200
- 0.2300
- 0.2400
- 0.2500