The MIT-LL/AFRL IWSLT-2009 MT System

Wade Shen, Brian Delaney, A. Ryan Aminzadeh
Tim Anderson and Ray Slyh
2 December 2009

This work is sponsored by the United States Air Force Research Laboratory under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.
Outline

• IWSLT-2009 System Architecture

• Better Arabic Morphology Processing
  – CoMMA

• Domain Adaptation Overview
  – Unsupervised and Semi-supervised Adaptation
  – Human-in-the Loop Adaptation
• Standard Statistical Architecture

• Developed in-house to support SMT experiments
  – Framework for experiments with low-resource languages
  – Test-bed for S2S MT system

• Most components are home-grown
  – Phrase Training/Minimum Error Rate Training
  – Moses and FST decoders used, comparable performance

• Participated in Arabic/Turkish ⇄ English BTEC Data track
Phrase Based FST Decoder


- The target language hypothesis is the best path through the following transducer:

\[ E = I \circ P \circ D \circ T \circ L \]

- where,
  - \( I \) = source language input acceptor
  - \( P \) = phrase segmentation transducer
  - \( D \) = weighted phrase swapping transducer
  - \( T \) = weighted phrase translation transducer (source phrases to target words)
  - \( L \) = weighted target language model acceptor

- Apply phrase swapping twice for long distance reordering

- OOV words are inserted during decoding as parallel links to \( P, D, T, \) and \( L \) models.

- Allows for direct decoding on pruned ASR lattices
System Combination

- Generate consensus networks using round-robin alignment, where each system gets to be the skeleton alignment
- Take union of all consensus networks and apply a language model
- Weight optimization on a development set using n-best lists
- Final combination on unseen data using optimized system weights
Outline

• IWSLT-2009 System Architecture

• Better Arabic Morphology Processing
  – CoMMA

• Domain Adaptation Overview
  – Unsupervised and Semi-supervised Adaptation
  – Human-in-the Loop Adaptation
## Arabic Preprocessing

### AP5 Review

<table>
<thead>
<tr>
<th>Preprocessing Method</th>
<th>Mean BLEU on dev6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (No normalization or AP5)</td>
<td>42.06</td>
</tr>
<tr>
<td>Remove all diacritics except tanween, no AP5</td>
<td>49.40</td>
</tr>
<tr>
<td>Remove all diacritics, no AP5</td>
<td>50.39</td>
</tr>
<tr>
<td>Remove all diacritics, apply AP5</td>
<td>53.55</td>
</tr>
</tbody>
</table>

- “Diacritics” removed:
  - Short vowels
  - **Sukuun**: Marks absence of sort vowel
  - **Shadda**: Marks consonant gemination (i.e., doubling)
  - **Tanween**: Case markers for indefinite forms & other uses
  - **Tatweel**: Stretches letters in Arabic typography (not a true diacritic)
- AP5 segments the following from stems:
  - **Prefixes**: al-, bi-, fa-, ka-, li-, wa-
  - **Suffixes**: Attached pronouns
CoMMA Processing for Arabic

- **Observation:** *With limited training data more morphological processing seems to help, less with more training data*

- **Count Mediated Morphological Analysis**
  - Modification to AP5: decide segmentation based on counts

- **Given a count threshold t, and a vocabulary W**

- **Foreach w in |W|**
  - Apply AP5 diacritic normalization procedure
  - If count(w) < t
    - Apply AP5 segmentation of clitics, etc.
  - Else don’t segment
### CoMMA Experiments

<table>
<thead>
<tr>
<th>COMMA Threshold</th>
<th>BLEU Score</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev6</td>
<td>Dev7</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>50.00</td>
<td>51.94</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>53.92</td>
<td>54.29</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>53.14</td>
<td>54.64</td>
<td></td>
</tr>
<tr>
<td>2,000</td>
<td>54.02</td>
<td>54.57</td>
<td></td>
</tr>
<tr>
<td>10,000</td>
<td>53.33</td>
<td>54.48</td>
<td></td>
</tr>
</tbody>
</table>

**Baseline (No Tokenization)**

**CoMMA**

**AP5 (all tokens segmented)**

- AP5 and CoMMA results in 7-8% relative improvement
- CoMMA only slightly better than AP5, +0.5–1.5 BLEU in system combination
Outline

• IWSLT-2009 System Architecture

• Better Arabic Morphology Processing
  – CoMMA

• Domain Adaptation Overview
  – Unsupervised and Semi-supervised Adaptation
  – Human-in-the Loop Adaptation
Cross Domain Adaptation Overview

- **Observations from past work**
  - SMT performs best when training and test data are **matched**
  - Adding large volumes of out-of-domain data to training **does not improve performance**

- **Adaptation**
  - **GOAL:** *Optimally port general purpose (out-of-domain) models to specific domain with limited in-domain data*

- **NOTE:** *Adapted Systems not used in IWSLT BTEC submissions*
Data

• General purpose data:
  – 500k Arabic-English parallel data from ISI automatically extracted parallel corpus
  – Domain: newswire data

• In-domain (adaptation) data:
  – 20k IWSLT-2009 BTEC Arabic-English training set
  – Domain: travel
Adaptation of Phrase-based MT Models

Semi-supervised

Initial Data

- General Purpose Parallel Data
- In-Domain Source Data

Translation & Evaluation

- Human Judges
- MT Output
- Adaptation Algorithms

Adaptation

- Adapted MT Model
- High-Quality MT Translations

Train

Translate
Adaptation of Phrase-based MT Models

Human-in-the-Loop

Initial Data
- General Purpose Parallel Data
- In-Domain Source Data

Translation & Evaluation
- GP Model
- MT Output
- Human Judges

Adaptation
- Adaptation Algorithms
- Adapted Model
- Human Translator Corrections

Poor-Quality MT Translations
- Human Translator Corrections
Selection of In-domain Adaptation Data

- General purpose models used to translate the IWSLT ’09 training set
-Translations ranked using METEOR as a proxy for a human judge
-Ranked sentences divided into octiles and used for experiments:
  - Semi-supervised adaptation: Use top scoring octiles for adaptation
    Goal: is to use best in-domain target data
  - Human-in-the-loop adaptation: Use bottom scoring octiles for adaptation
    Goal: is to correct worst in-domain target data (active learning paradigm)
Adaptation Approaches

Language Model Adaptation

- Optimized for BLEU
- Trained on:
  - Semi-supervised: Machine translations of IWSLT training set
  - Human-in-the-Loop: Reference translations of IWSLT training set
Adaptation Approaches

Phrase Table Adaptation

Phrase Table_{GP} \rightarrow MAP-Based Adaptation^* \rightarrow Phrase Table_{IWSLT+GP}

Phrase Table_{IWSLT}

Language Model_{GP} \rightarrow \lambda_1 \rightarrow \text{SCORE} \rightarrow \lambda_2

*Based on approaches described in:
Phrase Table MAP Adaptation

- Interpolated phrase table probabilities are computed using the following equation:

\[
\hat{p}(s \mid t) = \lambda p_{\text{in-domain}}(s \mid t) + (1 - \lambda) p_{\text{gp}}(s \mid t)
\]

- \(p_{\text{in-domain}}\): probability estimate from in-domain models
- \(p_{\text{gp}}\): probability estimate from general purpose models
- \(\lambda\): interpolation coefficient computed using the following equation:

\[
\lambda = \frac{N_{\text{in-domain}}(s,t)}{N_{\text{in-domain}}(s,t) + \tau}
\]

- \(\tau\): Fixed-value MAP relevance factor
- \(N_{\text{in-domain}}(s,t)\): observed count of phrase pair \((s, t)\)
Experimental Results

Semi-supervised Adaptation

Semi-supervised Training Experiments (IWSLT09 dev7)

Best semi-supervised Adaptation

Unsupervised Adaptation

BLEU Score

Top X Octiles of Training Set Scores

Phrase Table and LM Adaptation
In-Domain Only
LM Adaptation
Phrase Table Adaptation
Baseline (GP Model only)
Experimental Results

Human-in-the-Loop Adaptation

Human-in-the-Loop Experiments (IWSLT09 dev7)

Gains from Adaptation Methods:
- More improvement correcting BLEU
- 1/8 data correction = +13 BLEU improvement

Bottom X Octiles of Training Set Scores

GP Baseline

Phrase Table and LM Adaptation
In-Domain Only
LM Adaptation
Phrase Table Adaptation

Air Force Research Laboratory
MIT Lincoln Laboratory
## Experimental Results

### Best System Scores

<table>
<thead>
<tr>
<th>System</th>
<th>dev7</th>
<th>eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>GP</td>
<td>23.06</td>
<td>21.35</td>
</tr>
<tr>
<td>GP + Unsupervised LM + PT Adaptation</td>
<td>25.74</td>
<td>23.86</td>
</tr>
<tr>
<td>GP + Semi-supervised LM + PT Adaptation (Top quartile)</td>
<td>27.19</td>
<td>25.89</td>
</tr>
<tr>
<td>IWSLT '09 Baseline</td>
<td>54.63</td>
<td>52.69</td>
</tr>
<tr>
<td>GP + Human-in-the-Loop LM + PT Adaptation</td>
<td>56.57</td>
<td>56.11</td>
</tr>
</tbody>
</table>
Conclusions

• Morphological processing is critical
  – +4 BLEU for Turkish using Bilkent Analyzer
  – +3.5-4 BLEU for Arabic using AP5

• CoMMa gains in system combination
  – Multiple CoMMa systems (20, 200, 2000): +0.5-1.5 BLEU over AP5

• Unsupervised Adaptation
  – LM: +1.5 BLEU, PT: +0.5 BLEU
  – Combined: +2.5-3.0 BLEU (15% relative) compared to GP only

• Semi-supervised Adaptation
  – Gains +1.5-2 BLEU over Unsupervised, only ¼ of total data
  – But requires human judgement

• Human-in-the-Loop Adaptation
  – +2-3.5 BLEU using all IWSLT data
  – +13 BLEU using 1/8th of total data
  – Gains from LM and PT are non-additive