HKUST Statistical Machine Translation Experiments for IWSLT 2007

Yihai Shen  Chi-kiu Lo  Marine Carpuat  Dekai Wu

HKUST
Human Language Technology Center
Department of Computer Science
University of Science and Technology
Hong Kong

{shenyh, jackielo, marine, dekai}@cs.ust.hk
The HKUST submission
Goals for our second IWSLT participation

- Experiment with the open-source Moses decoder, focusing primarily on Chinese-English text translation
  - on various data sets and input conditions
    - Chinese-English text translation task
    - Challenge task on spontaneous speech cancelled by organizers
  - on various language pairs from different language families
    - Arabic-English, Chinese-English, Italian-English, Japanese-English

- Systematically compare Moses against the closed-source Pharaoh decoder
  - used by HKUST for IWSLT-2006
The HKUST submission
Secondary goals for contrastive experiments

- Obtain preliminary indications on performance with...
  - (semantics) integration of our recent WSD-for-SMT model [Carpuat & Wu 2007] with Moses (not Pharoah)
  - (syntax) our BITG decoder [Wu 1996] substituted for Moses

... while holding all else constant
Outline

- System description
- Experimental setup
  - Chinese-English
  - Other language pairs
- Results
- Contrastive experiments
  - (semantics) Phrase Sense Disambiguation: WSD for SMT
  - (syntax) Bracketing ITG decoder
System description
Experiments using several SMT decoders

- Decoders
  - Pharaoh [Koehn 2004]
  - Moses [Koehn 2007]
  - Moses [Koehn 2007] + WSD-for-SMT [Carpuat & Wu 2007]
  - Bracketing ITG [Wu 1996]

- Common assumptions of the controlled experiments
  - Phrasal bilexicon
  - Log-linear model
  - Phrases/words represented using surface forms only
    - did not use Moses’ factored representation option
System description
Common phrasal bilexicons used

- Learned from bidirectional IBM4 word alignments
  - produced by GIZA++ [Och & Ney 2002]

- Base features used [Koehn 2003]:
  - conditional translation probabilities in both directions
  - lexical weights derived from word translation probabilities

- Allowed phrase lengths up to 20 words
  - short sentences in a well-defined domain
System description
Common phrasal bilexicons used

- Compared two phrase extraction methods:
  - intersection
    - uses strict intersection of bidirectional word alignments
  - grow-diag-final
    - expands alignment by adding directly neighboring alignment points in diagonal neighborhood

- grow-diag-final produced better BLEU scores
  - typically around 0.5 points higher
System description

Language model

- Standard n-gram language models
  - trained using SRI LM toolkit [Stolcke 2002]

- Chinese-English: mixture*
  - 4-gram LM trained on BTEC English
  - 3-gram LM trained on English Gigaword

- Arabic-English, Italian-English, Japanese-English:
  - 3-gram LM trained on BTEC English

- Same LMs used for all experiments*

*except that BITG decoding used only a 3-gram LM trained on BTEC English
Outline

- System description
- Experimental setup
  - Chinese-English
  - Other language pairs
- Results
- Contrastive experiments
  - (semantics) Phrase Sense Disambiguation: WSD for SMT
  - (syntax) Bracketing ITG decoder
Experimental setup

IWSLT tasks

- Chinese-English text translation only
  - Challenge task (correct recognition vs. read speech vs. spontaneous speech) was cancelled by the organizers

- Text and read speech translation
  - Arabic-English
  - Italian-English
  - Japanese-English
Experimental setup
Minimal language-specific preprocessing

- **English** data was tokenized and case-normalized
- **Italian** data was processed as if it were English
- **Chinese** data was word segmented using LDC segmenter
- **Japanese** data was used directly as provided
- **Arabic**
  - Converted to Buckwalter romanization scheme
  - Tokenized with ASVMT Morphological Analysis toolkit [Diab 2005]
Experimental setup

Improving the sentence segmentation

- The original sentence segmentation is not optimal for training
- Re-segmenting the sentences consistently improves BLEU score

<table>
<thead>
<tr>
<th>IWSLT-07 data set</th>
<th># sentences</th>
<th># sentences after resegmentation</th>
<th>BLEU with original sentences</th>
<th>BLEU after resegmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE devtest1</td>
<td>506</td>
<td>546</td>
<td>41.09</td>
<td>42.05</td>
</tr>
<tr>
<td>CE devtest2</td>
<td>500</td>
<td>543</td>
<td>42.43</td>
<td>43.76</td>
</tr>
<tr>
<td>CE devtest2</td>
<td>506</td>
<td>558</td>
<td>51.86</td>
<td>53.51</td>
</tr>
</tbody>
</table>
**Experimental setup**

**Training corpus statistics**

- Corpora for Chinese and Japanese are twice as large as for Arabic and Italian.
- The English side of corpus for Arabic and Italian is a subset.

<table>
<thead>
<tr>
<th>Training data statistics</th>
<th>Chinese-English</th>
<th>Arabic-English</th>
<th>Italian-English</th>
<th>Japanese-English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of bisentences</td>
<td>39,953</td>
<td>19,972</td>
<td>19,972</td>
<td>39,953</td>
</tr>
<tr>
<td>Vocabulary size (input language)</td>
<td>11,178</td>
<td>25,152</td>
<td>17,917</td>
<td>12,535</td>
</tr>
<tr>
<td>Vocabulary size (English output)</td>
<td>18,992</td>
<td>13,337</td>
<td>13,337</td>
<td>18,992</td>
</tr>
</tbody>
</table>
Outline

- System description
- Experimental setup
  - Chinese-English
  - Other language pairs
- Results
- Contrastive experiments
  - (semantics) Phrase Sense Disambiguation: WSD for SMT
  - (syntax) Bracketing ITG decoder
Results

Official (buggy) results

- Submitted runs were buggy
  (arising from accidental errors in combining models and parameters)

<table>
<thead>
<tr>
<th>IWSLT07 task</th>
<th>Clear Transcription</th>
<th>ASR Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese-English</td>
<td>34.26</td>
<td>N/A</td>
</tr>
<tr>
<td>Arabic-English</td>
<td>19.51</td>
<td>14.20</td>
</tr>
<tr>
<td>Italian-English</td>
<td>17.02</td>
<td>17.02</td>
</tr>
<tr>
<td>Japanese-English</td>
<td>40.51</td>
<td>32.49</td>
</tr>
</tbody>
</table>

- Chinese-English: 34.26
  (range among 9 primary submissions: 19.34 - 40.77)
## Results
Updated results after removing bugs

<table>
<thead>
<tr>
<th>IWSLT07 data set</th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
<th>METEOR no synonyms</th>
<th>TER</th>
<th>WER</th>
<th>PER</th>
<th>CDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE devtest1 (buggy)</td>
<td>45.49</td>
<td>7.78</td>
<td>66.11</td>
<td>64.50</td>
<td>36.13</td>
<td>41.68</td>
<td>36.25</td>
<td>37.10</td>
</tr>
<tr>
<td>CE devtest1</td>
<td><strong>46.23</strong></td>
<td><strong>8.00</strong></td>
<td><strong>68.01</strong></td>
<td><strong>66.41</strong></td>
<td><strong>36.18</strong></td>
<td><strong>41.35</strong></td>
<td><strong>36.12</strong></td>
<td><strong>37.14</strong></td>
</tr>
<tr>
<td>CE devtest2 (buggy)</td>
<td>48.23</td>
<td>8.32</td>
<td>68.98</td>
<td>67.22</td>
<td>34.99</td>
<td>40.78</td>
<td>34.45</td>
<td>35.43</td>
</tr>
<tr>
<td>CE devtest2</td>
<td><strong>49.77</strong></td>
<td><strong>8.82</strong></td>
<td><strong>71.88</strong></td>
<td><strong>69.85</strong></td>
<td><strong>34.47</strong></td>
<td><strong>40.12</strong></td>
<td><strong>33.41</strong></td>
<td><strong>34.58</strong></td>
</tr>
<tr>
<td>CE devtest3 (buggy)</td>
<td>56.44</td>
<td>9.26</td>
<td>76.57</td>
<td>74.47</td>
<td>29.40</td>
<td>34.16</td>
<td>28.86</td>
<td>33.02</td>
</tr>
<tr>
<td>CE devtest3</td>
<td><strong>58.29</strong></td>
<td><strong>9.61</strong></td>
<td><strong>78.48</strong></td>
<td><strong>76.28</strong></td>
<td><strong>28.29</strong></td>
<td><strong>32.76</strong></td>
<td><strong>27.62</strong></td>
<td><strong>29.15</strong></td>
</tr>
<tr>
<td>CE test (buggy)</td>
<td>34.26</td>
<td>34.04</td>
<td>6.18</td>
<td>58.28</td>
<td>56.50</td>
<td>45.53</td>
<td>49.15</td>
<td>44.17</td>
</tr>
<tr>
<td>CE test</td>
<td><strong>35.12</strong></td>
<td><strong>6.51</strong></td>
<td><strong>60.47</strong></td>
<td><strong>58.57</strong></td>
<td><strong>44.89</strong></td>
<td><strong>48.30</strong></td>
<td><strong>43.40</strong></td>
<td><strong>41.50</strong></td>
</tr>
</tbody>
</table>
## Results

**Updated results after removing bugs**

<table>
<thead>
<tr>
<th>IWSLT07 data set</th>
<th>BLEU buggy submitted</th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
<th>METEOR no synonyms</th>
<th>TER</th>
<th>WER</th>
<th>PER</th>
<th>CDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE devtest1 (buggy)</td>
<td>45.49</td>
<td>7.78</td>
<td>66.11</td>
<td>64.50</td>
<td>36.13</td>
<td>41.68</td>
<td>36.25</td>
<td><strong>37.10</strong></td>
<td></td>
</tr>
<tr>
<td>CE devtest1</td>
<td><strong>46.23</strong></td>
<td><strong>8.00</strong></td>
<td><strong>68.01</strong></td>
<td><strong>66.41</strong></td>
<td><strong>36.18</strong></td>
<td><strong>41.35</strong></td>
<td><strong>36.12</strong></td>
<td>37.14</td>
<td></td>
</tr>
<tr>
<td>CE devtest2 (buggy)</td>
<td>48.23</td>
<td>8.32</td>
<td>68.98</td>
<td>67.22</td>
<td>34.99</td>
<td>40.78</td>
<td>34.45</td>
<td>35.43</td>
<td></td>
</tr>
<tr>
<td>CE devtest2</td>
<td><strong>49.77</strong></td>
<td><strong>8.82</strong></td>
<td><strong>71.88</strong></td>
<td><strong>69.85</strong></td>
<td><strong>34.47</strong></td>
<td><strong>40.12</strong></td>
<td><strong>33.41</strong></td>
<td><strong>34.58</strong></td>
<td></td>
</tr>
<tr>
<td>CE devtest3 (buggy)</td>
<td>56.44</td>
<td>9.26</td>
<td>76.57</td>
<td>74.47</td>
<td>29.40</td>
<td>34.16</td>
<td>28.86</td>
<td>33.02</td>
<td></td>
</tr>
<tr>
<td>CE devtest3</td>
<td><strong>58.29</strong></td>
<td><strong>9.61</strong></td>
<td><strong>78.48</strong></td>
<td><strong>76.28</strong></td>
<td><strong>28.29</strong></td>
<td><strong>32.76</strong></td>
<td><strong>27.62</strong></td>
<td><strong>29.15</strong></td>
<td></td>
</tr>
<tr>
<td>CE test (buggy)</td>
<td>34.26</td>
<td>34.04</td>
<td>6.18</td>
<td>58.28</td>
<td>56.50</td>
<td>45.53</td>
<td>49.15</td>
<td>44.17</td>
<td>41.53</td>
</tr>
<tr>
<td>CE test</td>
<td><strong>35.12</strong></td>
<td><strong>6.51</strong></td>
<td><strong>60.47</strong></td>
<td><strong>58.57</strong></td>
<td><strong>44.89</strong></td>
<td><strong>48.30</strong></td>
<td><strong>43.40</strong></td>
<td><strong>41.50</strong></td>
<td></td>
</tr>
</tbody>
</table>

**our own scoring tools give lower BLEU scores than the official IWSLT scoring**

HKUST Human Language Technology Center

Shen, Lo, Carpuat & Wu

IWSLT 2007
Results
Moses almost always outperforms Pharoah

- Varied many settings and pre-/post-processing steps (bilexicons, LMs, ...) to obtain experimental runs under many conditions

<table>
<thead>
<tr>
<th>Run No.</th>
<th>Pharaoh</th>
<th>Moses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41.14</td>
<td>41.17</td>
</tr>
<tr>
<td>2</td>
<td>41.65</td>
<td>41.70</td>
</tr>
<tr>
<td>3</td>
<td>42.05</td>
<td>42.16</td>
</tr>
<tr>
<td>4</td>
<td>43.40</td>
<td>43.55</td>
</tr>
<tr>
<td>5</td>
<td>41.92</td>
<td>42.26</td>
</tr>
<tr>
<td>6</td>
<td>42.80</td>
<td>43.19</td>
</tr>
<tr>
<td>7</td>
<td>43.76</td>
<td>44.28</td>
</tr>
<tr>
<td>8</td>
<td>44.17</td>
<td>44.64</td>
</tr>
<tr>
<td>9</td>
<td>51.64</td>
<td>52.19</td>
</tr>
<tr>
<td>10</td>
<td>52.15</td>
<td>52.59</td>
</tr>
<tr>
<td>11</td>
<td>53.51</td>
<td>53.64</td>
</tr>
<tr>
<td>12</td>
<td>53.87</td>
<td>53.53</td>
</tr>
</tbody>
</table>
Outline

- System description
- Experimental setup
  - Chinese-English
  - Other language pairs
- Results
- Contrastive experiments
  - (semantics) Phrase Sense Disambiguation: WSD for SMT
  - (syntax) Bracketing ITG decoder
Contrastive experiments (semantics)
Phrase Sense Disambiguation: WSD for SMT

- Today’s SMT makes little use of source-language context
- In contrast, WSD approaches generalize across rich contextual features to assign context-dependent probabilities to senses

- Earlier negative results: [Carpuat & Wu 2005]
  - Surprisingly, Senseval WSD models do not help translation quality when integrated into a word-based SMT model

- New: Using PSD, we repurpose the WSD models for SMT in our newer fully phrasal model: [Carpuat & Wu EMNLP, MT-Summit, TMI 2007]
  - Words are phrasal, just as in traditional lexicography
  - WSD “senses” are exactly same as SMT translation candidates
  - WSD training data is exactly same as SMT training data
  - WSD scores are added to log linear model feature set
  - Feature engineering is exactly inherited from Senseval WSD models
Contrastive experiments (semantics)
The HKUST WSD System

Proved highly effective at Senseval-3
- Placed first on Chinese lexical sample
- Placed second on Multilingual lexical sample (translation)
- 71.4% on English lexical sample (median 67.2, best 72.9)

Classifier ensemble:
- naïve Bayes [Yarowsky & Florian 2002]
- maximum entropy [Klein & Manning 2002]
- boosting [Carreras et al. 2002; Wu et al. 2002]: we use boosted decision stumps
- Kernel PCA model [Wu et al. 2004]
**Contrastive experiments** *(semantics)*

Contextual features in HKUST WSD system

- Feature set includes:
  - Bag-of-words context
  - Position sensitive local collocational features
  - Syntactic features

- A WSD model using these features yielded the best classification accuracy in Yarowsky & Florian [2002]
Contrastive experiments (semantics)
PSD improved Moses... just like Pharoah

- Encouraging preliminary indication
- Consistent with our larger EMNLP-CoNLL results [Carpuat & Wu 2007]

<table>
<thead>
<tr>
<th>Run No.</th>
<th>Pharaoh</th>
<th>Moses</th>
<th>WSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41.14</td>
<td><strong>41.17</strong></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>41.65</td>
<td>41.70</td>
<td><strong>43.47</strong></td>
</tr>
<tr>
<td>3</td>
<td>42.05</td>
<td>42.16</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>43.40</td>
<td><strong>43.55</strong></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>41.92</td>
<td><strong>42.26</strong></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>42.80</td>
<td><strong>43.19</strong></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>43.76</td>
<td><strong>44.28</strong></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>44.17</td>
<td><strong>44.64</strong></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>51.64</td>
<td><strong>52.19</strong></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>52.15</td>
<td><strong>52.59</strong></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>53.51</td>
<td><strong>53.64</strong></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td><strong>53.87</strong></td>
<td>53.53</td>
<td></td>
</tr>
</tbody>
</table>
Outline

- System description
- Experimental setup
  - Chinese-English
  - Other language pairs
- Results
- Contrastive experiments
  - (semantics) Phrase Sense Disambiguation: WSD for SMT
  - (syntax) Bracketing ITG decoder
Contrastive experiments \textit{(syntax)}
Decoding under the ITG Hypothesis

- Intrinsically imposes ITG constraints on permutations/reorderings

[Wu 1995]
Contrastive experiments (syntax)
Bracketing ITG decoder

- Basic decoding algorithm is polynomial-time $O(n^7)$ [Wu 1996]
- Current version uses beam search
- Current version integrates trigram LM
  - Note: did not use 4-gram LM or Gigaword 3-gram LM, so has less information than the Moses and Pharoah models
- Phrase-based SMT’s distortion feature replaced by BITG permutation score
- All other factors controlled to be the same as Moses and Pharoah
  - Note: did not yet take advantage of any additional syntactic or other information naturally integrated into ITGs
Contrastive experiments (syntax)
BITG decoding competitive with Moses

- Again, encouraging preliminary indications

<table>
<thead>
<tr>
<th>Run No.</th>
<th>Pharaoh</th>
<th>Moses</th>
<th>WSD</th>
<th>BITG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41.14</td>
<td>41.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>41.65</td>
<td>41.70</td>
<td>43.47</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>42.05</td>
<td>42.16</td>
<td></td>
<td>43.04</td>
</tr>
<tr>
<td>4</td>
<td>43.40</td>
<td>43.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>41.92</td>
<td>42.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>42.80</td>
<td>43.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>43.76</td>
<td>44.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>44.17</td>
<td>44.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>51.64</td>
<td>52.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>52.15</td>
<td>52.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>53.51</td>
<td>53.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>53.87</td>
<td>53.87</td>
<td>53.53</td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

- We have described experiments at HKUST focusing primarily on the Chinese-English task
  - also reported results on 3 other language pairs from different language families

- On Chinese-English, both our Pharaoh and Moses based systems achieved good performance

- Moses almost always outperforms Pharaoh
  - across a wide variety of experimental conditions

- Preliminary indications from contrastive experiments:
  - our WSD-for-SMT model improves Moses too
  - plain vanilla BITG decoding appears competitive with Moses