The LIG Arabic / English Speech Translation System at IWSLT07

Laurent BESACIER,
Amar MAHDHAOUI,
Viet-Bac LE

LIG*/GETALP (Grenoble, France)
Laurent.Besacier@imag.fr

* Former name : CLIPS
OUTLINE

1 Baseline MT system
- Task, data & tools
- Restoring punctuation and case
- Use of out-of-domain data
- Adding a bilingual dictionary

2 Lattice decomposition for CN decoding
- Lattice to CNs
- Word lattices to sub-word lattices
- What SRI-LM does
- Our algo.
- Examples in arabic

3 Speech translation experiments
- Results on IWSLT06
- Results on IWSLT07 (eval)
OUTLINE

1 Baseline MT system
- Task, data & tools
- Restoring punctuation and case
- Use of out-of-domain data
- Adding a bilingual dictionary
Task, data & tools

- First participation to IWSLT
  - A/E task
  - Conventional phrase-based system using Moses+Giza+sri-lm

- Use of IWSLT-provided data (20k bitext) except
  - A 84k A/E bilingual dictionary taken from http://freedict.cvs.sourceforge.net/freedict/eng-ara/
  - The buckwalter morphological analyzer
  - LDC’s Gigaword corpus (for english LM training)
Restoring punctuation and case

- 2 separated punct. and case restoration tools built using *hidden-ngram* and *disambig* commands from sri-lm
  - => restore MT outputs

<table>
<thead>
<tr>
<th></th>
<th>(1) train with case &amp; punct</th>
<th>(2) train without case &amp; punct</th>
<th>(3) train with restored case &amp; punct</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev06</td>
<td>0.2341</td>
<td>0.2464</td>
<td>0.2298</td>
</tr>
<tr>
<td>tst06</td>
<td>0.1976</td>
<td>0.1948</td>
<td>0.1876</td>
</tr>
</tbody>
</table>

Option (2) kept
Use of out-of-domain data

- Baseline in-domain LM trained on the english part of A/E bitext
- Interpolated LM between Baseline and Out-of-domain (LDC gigaword) : 0.7/0.3

<table>
<thead>
<tr>
<th></th>
<th>In domain LM No MERT</th>
<th>Interpolated in-domain and out-of-domain LM No MERT</th>
<th>Interpolated in-domain and out-of-domain LM MERT on dev06</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev06</td>
<td>0.2464</td>
<td>0.2535</td>
<td>0.2674</td>
</tr>
<tr>
<td>tst06</td>
<td>0.1948</td>
<td>0.2048</td>
<td>0.2050</td>
</tr>
</tbody>
</table>
Adding a bilingual dictionary

- A 84k A/E bilingual dictionary taken from http://freedict.cvs.sourceforge.net/freedict/eng-ara/
- Directly concatenated to the training data + retraining + retuning (mert)

<table>
<thead>
<tr>
<th></th>
<th>No bilingual dict.</th>
<th>Use of a bilingual dict.</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev06</td>
<td>0.2674</td>
<td>0.2948</td>
</tr>
<tr>
<td>tst06</td>
<td>0.2050</td>
<td>0.2271</td>
</tr>
</tbody>
</table>

Submitted MT system (from verbatim trans.)
OUTLINE

2 Lattice decomposition for CN decoding
- Lattice to CNs
- Word lattices to sub-word lattices
- What SRI-LM does
- Our algo.
- Examples in arabic
Lattice to CNs

- Moses allows to exploit CN as interface between ASR and MT
- Example of word lattice and word CN
Word lattices to sub-word lattices

- **Problem**: word graphs provided for IWSLT07 do not have necessarily word decomposition compatible with the word decomposition used to train our MT models
  - Word units vs sub-word units
  - Different sub-word units used
- Need for a lattice decomposition algorithm
What SRI-LM does

- Example: CANNOT splitted into CAN and NOT
- \textit{-split-multiwords} option of \textit{lattice-tool}
  - First node keeps all the information
  - new nodes have null scores and zero-duration
Proposed lattice decomposition algorithm (1)

- identify the arcs of the graph that will be split (decompoundable words)
- each arc to be split is decomposed into a number of arcs that depends on the number of subword units
- the start / end times of the arcs are modified according to the number of graphemes into each subword unit
- so are the acoustic scores
- the first subword of the decomposed word is equal to the initial LM score of the word, while the following subwords LM scores are made equal to 0

Freely available on http://www-clips.imag.fr/geod/User/viet-bac.le/outils/
Proposed lattice decomposition algorithm (2)
Examples in arabic

Word lattice
Examples in arabic

Sub-Word lattice
OUTLINE

3 Speech translation experiments
- Results on IWSLT06
- Results on IWSLT07 (eval)
Results on IWSLT06

- Full CN decoding (subword CN as input)
  - obtained after applying our word lattice decomposition algorithm
  - all the parameters of the log-linear model used for the CN decoder were retuned on dev06 set
- “CN posterior probability parameter” to be tuned

<table>
<thead>
<tr>
<th></th>
<th>(1) verbatim</th>
<th>(2) 1-best</th>
<th>(3) cons-dec</th>
<th>(4) full-cn-dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev06</td>
<td>0.2948</td>
<td>0.2469</td>
<td>0.2486</td>
<td>0.2779</td>
</tr>
<tr>
<td>tst06</td>
<td>0.2271</td>
<td>0.1991</td>
<td>0.2009</td>
<td>0.2253</td>
</tr>
</tbody>
</table>

ASR secondary  ASR primary
## Results on IWSLT07 (eval)

<table>
<thead>
<tr>
<th></th>
<th>clean verbatim</th>
<th>ASR 1-best</th>
<th>ASR full-cn-dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eva07</td>
<td>0.4135</td>
<td>0.3644</td>
<td>0.3804</td>
</tr>
</tbody>
</table>

### AE ASR

<table>
<thead>
<tr>
<th></th>
<th>BLEU score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1XXXXX</td>
<td>0.4445</td>
</tr>
<tr>
<td>2XXXXX</td>
<td>0.4429</td>
</tr>
<tr>
<td>3XXXXX</td>
<td>0.4092</td>
</tr>
<tr>
<td>4XXXXX</td>
<td>0.3942</td>
</tr>
<tr>
<td>5XXXXX</td>
<td>0.3908</td>
</tr>
<tr>
<td>6LIG_AE_ASR_primary_01</td>
<td><strong>BLEU score = 0.3804</strong></td>
</tr>
<tr>
<td>7XXXXX</td>
<td>0.3756</td>
</tr>
<tr>
<td>8XXXXX</td>
<td>0.3679</td>
</tr>
<tr>
<td>9XXXXX</td>
<td>0.3644</td>
</tr>
<tr>
<td>10XXXXX</td>
<td>0.3626</td>
</tr>
<tr>
<td>11XXXXX</td>
<td>0.1420</td>
</tr>
</tbody>
</table>