Evaluating Syntax-Driven Approaches to Phrase Extraction for Machine Translation

Ankit Srivastava
CNGL, DCU

Sergio Penkale
CNGL, DCU

Declan Groves
Traslán

John Tinsley
NCLT, DCU

3rd International Workshop on Example-Based Machine Translation, 12-13 November, 2009: Dublin, Ireland
about

- EBMT-influenced data sources used in a PB-SMT model (Moses)
  - Marker Hypothesis
  - Parallel Treebanks
- Lessons learned from work carried out over a number of years at DCU
- Focus on techniques for supplementing Moses phrases with syntactically motivated phrases
Supplementing with Syntax-driven Phrases
November 12, 2009 :: Dublin, Ireland

pb-smt system

- Moses framework [Koehn et al., 2007]

Translation model
- Heuristics-based phrase extraction from bidirectional word alignments
- Syntactically-motivated phrase extraction: marker / treebank
**moses phrases: an example**

Official journal of the European Communities  
Journal officiel des Communautés européennes

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>Official journal</td>
<td>Journal officiel</td>
</tr>
<tr>
<td>Official journal of</td>
<td>Journal officiel des</td>
</tr>
<tr>
<td>Official journal of the</td>
<td>Journal officiel des \</td>
</tr>
<tr>
<td>European Communities</td>
<td>Communautés européennes</td>
</tr>
<tr>
<td>of</td>
<td>des</td>
</tr>
<tr>
<td>of the European Communities</td>
<td>des Communautés européennes</td>
</tr>
<tr>
<td>the European Communities</td>
<td>Communautés européennes</td>
</tr>
<tr>
<td>European</td>
<td>européennes</td>
</tr>
</tbody>
</table>
marker-based

- Chunk sentences on encountering a ‘marker’ word
  - Founded on the Marker Hypothesis [Green, 1979]
  - Marker words are closed class of lexemes / morphemes
  - Each marker word associated with a marker category (tag)
  - 7 marker categories identified. E.g. DET, PREP, PRON
  - Each marker chunk must contain at least 1 non-marker word

- Align bilingual marker chunks
  - Use marker tag and relative positions in the sentence
  - Use cognate and MI scores

- Obtain marker-based phrase pairs
marker-based: an example

That is almost a personal record for me this autumn
C’est pratiquement un record personnel pour moi cet automne

<DET>That is almost <DET>a personal record <PREP>for <PRON>me <DET>this autumn
<DET>C’est pratiquement <DET>un record personnel <PREP>pour <PRON>moi <DET>cet automne

That is almost ↔ C’est pratiquement
a personal record ↔ un record personnel
for me this autumn ↔ pour moi cet automne
**marker-based: direct**

- Merging phrase pairs in a single phrase table
- Fr-En Europarl data: (3-gram lang model, Pharaoh decoder)
- System performance as training data increases
- 13% new phrases added via marker-based phrases

![Graph showing BLEU, NIST, METEOR scores for different phrase counts.]

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.3079</td>
<td>0.5590</td>
<td>0.6025</td>
</tr>
<tr>
<td>1-counter</td>
<td>0.3078</td>
<td>0.5775</td>
<td>0.6024</td>
</tr>
</tbody>
</table>

- Es-En Europarl data: (200K train, 5-gram lang model, Moses decoder)
treebank-based

- Monolingual parsing of sentences
  - Parse both sides
  - Requires constituency-structure parsers
- Align bilingual parse trees
  - Requires a sub-tree aligner [Zhechev & Way, 2008]
- Get aligned phrases
  - Extract surface-level chunks
- Also implemented using dependency structure
  - Using off-the-shelf dependency parsers
  - Head percolation of constituency trees [Magerman, 1995]
treebank-based: an example [con]

the green witch

NP
  DET
  the
  ADJP
  JJ
  green
  N
  witch

la bruja verde

NP
  DET
  la
  ADJP
  N
  JJ
  bruja
  verde
Supplementing with Syntax-driven Phrases

November 12, 2009 :: Dublin, Ireland

treebank-based: an example [con]

the green witch ↔ la bruja verde

the green witch ↔ la bruja verde

green witch ↔ bruja verde

the ↔ la

green ↔ verde

witch ↔ bruja
treebank-based: direct

- En-ES Europarl data: (700K train, 5-gram lang model, Moses decoder)
- Moses (Baseline), Constituency (Syntax)
- Merging phrase pairs in a single phrase table
- 24M phrases in Baseline Vs 6M phrases in Syntax
- 4.87% overlap between Moses and Syntax
- 16.79% new phrases added

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.3341</td>
<td>7.0765</td>
<td>0.5739</td>
</tr>
<tr>
<td>+Syntax</td>
<td>0.3397</td>
<td>7.0891</td>
<td>0.5782</td>
</tr>
<tr>
<td>Syntax_only</td>
<td>0.3153</td>
<td>6.8187</td>
<td>0.5598</td>
</tr>
</tbody>
</table>
treebank-based: direct

- Fr-En Europarl data: (100K train, 5-gram lang model, Moses decoder)
- Moses (B), Constituency (C), Dependency (D), Percolated (P)
- Merging phrase pairs in a single phrase table (1 / 2 / 3 / 4)
- Compare sizes of B with C/D/P
- Overlap between tables

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.2850</td>
<td>7.00</td>
<td>0.5783</td>
</tr>
<tr>
<td>B + C</td>
<td>0.2950</td>
<td>7.10</td>
<td>0.5855</td>
</tr>
<tr>
<td>B + D</td>
<td>0.2930</td>
<td>7.08</td>
<td>0.5843</td>
</tr>
<tr>
<td>B + P</td>
<td>0.2945</td>
<td>7.10</td>
<td>0.5854</td>
</tr>
<tr>
<td>B + C + D</td>
<td>0.2929</td>
<td>7.09</td>
<td>0.5848</td>
</tr>
<tr>
<td>B + C + P</td>
<td>0.2949</td>
<td>7.10</td>
<td>0.5850</td>
</tr>
<tr>
<td>B + D + P</td>
<td>0.2939</td>
<td>7.09</td>
<td>0.5849</td>
</tr>
<tr>
<td>B + C + D + P</td>
<td>0.2940</td>
<td>7.09</td>
<td>0.5849</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CON(C)</td>
<td>0.2564</td>
<td>6.55</td>
<td>0.5526</td>
</tr>
<tr>
<td>DEP(D)</td>
<td>0.2524</td>
<td>6.59</td>
<td>0.5465</td>
</tr>
<tr>
<td>PERC(P)</td>
<td>0.2587</td>
<td>6.59</td>
<td>0.5563</td>
</tr>
<tr>
<td>C + D</td>
<td>0.2632</td>
<td>6.69</td>
<td>0.5556</td>
</tr>
<tr>
<td>C + P</td>
<td>0.2637</td>
<td>6.62</td>
<td>0.5605</td>
</tr>
<tr>
<td>D + P</td>
<td>0.2657</td>
<td>6.74</td>
<td>0.5583</td>
</tr>
<tr>
<td>C + D + P</td>
<td>0.2690</td>
<td>6.75</td>
<td>0.5614</td>
</tr>
</tbody>
</table>
recap

- Baseline system (Moses):
  - `source_phrase ||| target_phrase ||| [feature_value]
- Alternate phrase pairs
  - Marker-based: `src ||| tgt`
  - Treebank-based (con, dep): `src ||| tgt`
- Experiments on direct combination
  - Merging phrase pairs and re-estimating probabilities

*Other ways to supplement the Moses phrase table with alternate phrase segmentation approaches*
Combining strategies

- Direct combination ✓
- Weighted combination
- Prioritised combination
- Feature-based
- System combination
Weighted combination

- Instead of simple merging, add ‘n’ copies of a type of phrase pair
- This modifies the relative frequency of the syntax-based phrase pairs
- Generally does not improve over direct combination
- Experiments on adding n copies of marker-based phrases
- Experiments on adding n copies of constituency-based phrases

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
<th>EBMT%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.3079</td>
<td>7.5590</td>
<td>0.6025</td>
<td>24.21</td>
</tr>
<tr>
<td>1-count</td>
<td>0.3078</td>
<td>7.5775</td>
<td>0.6024</td>
<td>23.47</td>
</tr>
<tr>
<td>2-count</td>
<td>0.3076</td>
<td>7.5582</td>
<td>0.6020</td>
<td>23.64</td>
</tr>
<tr>
<td>4-count</td>
<td>0.3071</td>
<td>7.5609</td>
<td>0.6015</td>
<td>24.34</td>
</tr>
<tr>
<td>8-count</td>
<td>0.3083</td>
<td>7.5969</td>
<td>0.6018</td>
<td>26.64</td>
</tr>
<tr>
<td>16-count</td>
<td>0.3042</td>
<td>7.5386</td>
<td>0.5986</td>
<td>29.71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline+Syntax</td>
<td>0.3397</td>
<td>7.0891</td>
<td>0.5782</td>
</tr>
<tr>
<td>+Syntax x2</td>
<td>0.3386</td>
<td>7.0813</td>
<td>0.5776</td>
</tr>
<tr>
<td>+Syntax x3</td>
<td>0.3361</td>
<td>7.0584</td>
<td>0.5756</td>
</tr>
<tr>
<td>+Syntax x5</td>
<td>0.3377</td>
<td>7.0829</td>
<td>0.5771</td>
</tr>
<tr>
<td>Half-weights</td>
<td>0.3404</td>
<td>7.1050</td>
<td>0.5792</td>
</tr>
</tbody>
</table>
Prioritized combination

- A phrase pair consists of \texttt{src \|\| tgt \|\| [feature_value]}

- Alternative to direct combination
  - Prioritize set A over set B
  - Add only those B phrase pairs when src not in A

- Experiments on baseline & constituency
  - No improvements over direct combination

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.3341</td>
<td>7.0765</td>
<td>0.5739</td>
</tr>
<tr>
<td>+Syntax</td>
<td>0.3397</td>
<td>7.0891</td>
<td>0.5782</td>
</tr>
<tr>
<td>Syntax_only</td>
<td>0.3153</td>
<td>6.8187</td>
<td>0.5598</td>
</tr>
<tr>
<td>Syntax Prioritised</td>
<td>0.3339</td>
<td>6.9887</td>
<td>0.5723</td>
</tr>
<tr>
<td>Baseline Prioritised</td>
<td>0.3381</td>
<td>7.0835</td>
<td>0.5789</td>
</tr>
</tbody>
</table>
feature-based combination

- A phrase pair consists of src ||| tgt ||| [feature_value]
- Add a new feature
  - Binary: type of phrase pair
  - MERT tuning assigns weight like other features
- Merging like direct combination
- Experiments on marker-based
  - Improvements in translation quality

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
<th>EBMT%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.3079</td>
<td>7.5590</td>
<td>0.6025</td>
<td>24.21</td>
</tr>
<tr>
<td>1-count</td>
<td>0.3078</td>
<td>7.5775</td>
<td>0.6024</td>
<td>23.47</td>
</tr>
<tr>
<td>Feature</td>
<td>0.3111</td>
<td>7.6004</td>
<td>0.6055</td>
<td>35.09</td>
</tr>
</tbody>
</table>
System combination

- So far, all methods have altered how phrases merged into one phrase table
- An alternative is to combine translated sentences (after decoding) rather than phrase pairs (during training)
- Use MBR-CN system combination [Du et al., 2009]
  - Experiments on B/C/D/P
  - Output sentences are unique enough to profit
  - 7.16% relative (4 systems), 12.3% relative (15 systems)

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MBR (4 systems)</td>
<td>0.2952</td>
<td>6.85</td>
<td>0.5784</td>
</tr>
<tr>
<td>CN (4 systems)</td>
<td>0.3070</td>
<td>7.06</td>
<td>0.5852</td>
</tr>
<tr>
<td>MBR (15 systems)</td>
<td>0.3260</td>
<td>7.32</td>
<td>0.6050</td>
</tr>
<tr>
<td>CN (15 systems)</td>
<td>0.3251</td>
<td>7.33</td>
<td>0.6039</td>
</tr>
</tbody>
</table>
Lessons learned

- Syntax-based phase pairs are a unique knowledge source
  - Overlap between phrase pairs
- Using only syntax-based phrases deteriorates
  - Large coverage of PB-SMT method
- Supplementing PB-SMT with syntax-based helps
  - Explored 5 different strategies for combining
  - System combination helps the most
- Decrease in gains as training data increases
Lessons learned:

- Syntax-based phase pairs
  - Overlap between phrase pairs
- Using only syntax-based phrases
  - Large coverage of PB-SMT
- Supplementing PB-SMT with syntax-based phrases
  - Explored 5 different strategies
  - System combination helps
- Decrease in gains as training data increases
Examined a number of different phrase segmentation approaches for MT

Explored ways of using linguistic information (borrowed from EBMT research) in a PB-SMT system

Level of improvement is dependent on amount of training data

Useful for languages with limited training data and MT systems with a smaller footprint

Difficult to improve the PB-SMT alignment / extraction / decoding pipeline without significant remodeling
thank you!

Questions?

Contact info
- Declan: dgroves @ traslan.ie
- Sergio: spenkale @ computing.dcu.ie
- John: jtinsley @ computing.dcu.ie
- Ankit: asrivastava @ computing.dcu.ie
Bonus Slide: Sample Output

REF: Does the commission intend to seek more transparency in this area?

MOSES: Will the commission ensure that more than transparency in this respect?

CON: The commission will the commission ensure greater transparency in this respect?

DEP: The commission will the commission ensure greater transparency in this respect?

PERC: Does the commission intend to ensure greater transparency in this regard?