Chunk-based Verb Reordering in VSO Sentences for Arabic-English SMT

Arianna Bisazza, Marcello Federico
FBK-irst Trento, Italy

WMT 2010, Uppsala, 15-16 July 2010
Introduction

- English word order: Subject-Verb-Object
- Arabic: both SVO and VSO
- Common errors in phrase-based SMT outputs:
  - wrong order of syntactic constituents
  - verbless sentences
Outline

- Reordering patterns in Arabic-English
- Chunk-based verb reordering: technique and analysis
- Impact of VSO sentences on translation quality
- Chunk-based reordering lattices
Outline

- Reordering patterns in Arabic-English
- Chunk-based verb reordering: technique and analysis
- Impact of VSO sentences on translation quality
- Chunk-based reordering lattices
Reordering patterns in Arabic-English

VSO sentence: Arabic verb *anticipated* wrt English

The Moroccan monarch King Mohamed VI *renewed* his support to the project of the French President
Reordering patterns in Arabic-English

VSO sentence: Arabic verb *anticipated* wrt English

Several local, one long reordering involving the verb

Typical phrase-based SMT outputs:

*The Moroccan monarch King Mohamed VI ___ his support to...*

*He renewed the Moroccan monarch King Mohamed VI his support to...*
Previous works

(Habash '07; Crego&Habash '08; Elming&Habash '09)

・ preprocess source data to approximate target word order
・ address all reorderings
・ deterministic reordering => 1 most probable permutation
・ non-deterministic => word reordering lattices

Our work:

・ only one class of reorderings
・ mixed approach: deterministic for train, lattices for test
Reordering patterns in Arabic-English

Working hypothesis:

*uneven distribution of reordering phenomena*
Reordering patterns in Arabic-English

Working hypothesis:

*uneven distribution of reordering phenomena*

Many local

Few global
Reordering patterns in Arabic-English

Working hypothesis:

*uneven distribution of reordering phenomena*

Many local
- adjectives follow nouns
- head-initial genitive constructions (*idafa*)

Example =>

```
 dAjrp  AlElAqAt  AlEAmph

 the Public Relations Department
```

Few global
Reordering patterns in Arabic-English

Working hypothesis:

*uneven distribution of reordering phenomena*

Many local

- adjectives follow nouns
- head-initial genitive constructions (*idafa*)

Example =>

```
dA)rp  AlE1AqAt  AlEAmp
the Public Relations Department
```

Few global

- Verb-Subject-Object sentences
Reordering patterns in Arabic-English

VSO sentences:
moving verb after subject simplifies reordering

The Moroccan monarch King Mohamed VI renewed his support to the project of the French President

Other (local) reorderings:
handled inside phrases or through distortion

WMT 2010, Uppsala
A. Bisazza, M. Federico
Outline

- Reordering patterns in Arabic-English
- **Chunk-based verb reordering: technique and analysis**
- Impact of VSO sentences on translation quality
- Chunk-based reordering lattices
Chunk-based verb reordering

– Simplifying assumptions:

1) verb reordering only between shallow syntax chunks;
2) no overlap between consecutive verb movements
Chunk-based verb reordering

– Simplifying assumptions:
  1) verb reordering only between shallow syntax chunks;
  2) no overlap between consecutive verb movements

– Possible movements:
  move verb chunk...

CH0  CH1  CH2  CH3  CH4  CH5  CH6  CH7  CH8  CH9
Chunk-based verb reordering

– Simplifying **assumptions:**
  1) verb reordering only between shallow syntax chunks;
  2) no overlap between consecutive verb movements

– Possible **movements:**
  move verb chunk...
  ...or verb chunk + next chunk (e.g. adverbials)
  by up to X chunks to the right
Chunk-based verb reordering

Best movement:

minimizes distortion wrt English translation

The Moroccan monarch King Mohamed VI renewed his support to the project of the French President
Chunk-based verb reordering: corpus analysis

Distribution by movement length

Max shift length (in nb. of chunks)

Percentage of verbs

Intersection of GIZA++ alignments
Manual alignments

eval08-NW
Gale-NW
Chunk-based verb reordering: corpus analysis

Distribution by movement length

=> Good coverage (≥ 99.5%) with max movement length 6
Outline

- Reordering patterns in Arabic-English
- Chunk-based verb reordering: technique and analysis
- Impact of VSO sentences on translation quality
- Chunk-based reordering lattices
Impact of VSO sentences on MT quality

- Baseline: Moses, 30M words newswire from NIST09
Impact of VSO sentences on MT quality

- Baseline: Moses, 30M words newswire from NIST09
- Shallow syntax chunking: AMIRA (Diab&al.2004)
- Verb-reorder training and devset, re-train whole system
Impact of VSO sentences on MT quality

- Baseline: Moses, 30M words newswire from NIST09
- Shallow syntax chunking: AMIRA (Diab&al.2004)
- Verb-reorder training and devset, re-train whole system
- Verb-reorder test aligned with reference *(oracle)*
- Tested with different Distortion Limits (DL) from 2 to 10 and wide beam search
Impact of VSO sentences on MT quality

%BLEU scores on Eval08-NW (MERT on Dev06-NW):
Impact of VSO sentences on MT quality

%BLEU scores on Eval08-NW (MERT on Dev06-NW):

Verb reordering of training data only => positive effect
(9% more phrases extracted)
Impact of VSO sentences on MT quality

%BLEU scores on Eval08-NW (MERT on Dev06-NW):

Verb reordering of training and test => further gain (+1.2 with 1/3 of sentences modified)

Verb reordering of training data only => positive effect (9% more phrases extracted)
Impact of VSO sentences on MT quality

%BLEU scores on Eval08-NW (MERT on Dev06-NW):

- Verb reordering of training and test => further gain (+1.2 with 1/3 of sentences modified)
- Verb reordering of training data only => positive effect (9% more phrases extracted)
- Relaxing the DL to high values doesn’t help
Impact of VSO sentences on MT quality

To resume:

- VSO sentences affect negatively phrase-based SMT
- Specific models needed to handle verb reordering of test
Outline

- Reordering patterns in Arabic-English
- Chunk-based verb reordering: technique and analysis
- Impact of VSO sentences on translation quality
- **Chunk-based reordering lattices**
Chunk-based reordering lattices

- Word lattices: represent input ambiguities (segmentation, decompounding, ... ordering)

- Thanks to our assumptions, we build compact reordering lattices and run non-monotonic decoding on them

- Double strategy
  - for global reordering: lattices
  - for local reordering: standard (phrase-internal and distortion)
Word-based vs Chunk-based lattices

EN: Official sources confirmed that there was a link between the attacks.
Word-based vs Chunk-based lattices


**EN:** Official sources confirmed that there was a link between the attacks.

1+6 reordering paths
Word-based vs Chunk-based lattices

EN: Official sources confirmed that there was a link between the attacks.

1+6 reordering paths

1+3 reordering paths
Word-based vs Chunk-based lattices

AR: w \[>kt\]_{VP} [mSAdr rsmyp]_{NP} [ wjwd rAbT]_{NP} [ byn AlAEtd'A'At]_{PP}.

EN: Official sources confirmed that there was a link between the attacks.

1+6 reordering paths

1+3 reordering paths

Chunk-to-word expansion
Chunk-based reordering lattices

Lattice representation of the rule:

“move 1 or 2 chunks by up to 6 chunk positions right”
Chunk-based reordering lattices

Lattice representation of the rule:

“move 1 or 2 chunks by up to 6 chunk positions right”
Chunk-based reordering lattices

Lattice representation of the rule:

“move 1 or 2 chunks by up to 6 chunk positions right”

Simple edge weighting scheme:

\[
\begin{align*}
1 & \quad \text{for the plain order path} \\
0.25 & \quad \text{for the reordering paths}
\end{align*}
\]
Chunk-based reordering lattices: Evaluation

- Same Moses-based system
- Training and tuning on verb-reordered data
- Non-monotonic decoding of word lattices by Dyer et al. (2008)
Abstract

In Arabic to English phrase-based statistical machine translation a large number of syntactic disfluencies are due to wrong long-range reordering of the verb in VSO sentences, where the verb is anticipated with respect to the English word order. In this paper we propose a chunk-based reordering technique to automatically detect and displace clause-initial verbs in the Arabic side of a word-aligned parallel corpus. This method is applied to preprocess the training data and to collect statistics about verb movements. From this analysis specific verb reordering lattices are then built on the test sentences before decoding them. The application of our reordering methods on the training and test sets results in consistent BLEU score improvements on the NIST-MT 2010 Arabic-English benchmark.

Acknowledgments

This work was supported by the EuroMatrixPlus project (IST-342831) which is funded by the European Commission under the Seventh Framework Programme for Research and Technological Development.

System DL | eval08nw | reo08nw | eval09nw
--- | --- | --- | ---
baseline 6 | 43.10 | 46.90 | 48.13
reord. training +
plain input 6 | 43.67 | 46.64 | 48.53
lattice 4 | 44.04 | 47.51 | 48.96
oracle reord. 4 | 44.36 | 48.25 | 49.26

Table 2. BLEU scores of baseline and reordered system on plain test and on reordering lattices.
## Chunk-based Reordering Lattices: Evaluation

%BLEU scores on Eval08-NW, Reo08-NW (specific set containing only VSO sentences) and Eval09-NW:

<table>
<thead>
<tr>
<th>System</th>
<th>DL</th>
<th>eval08nw</th>
<th>reo08nw</th>
<th>eval09nw</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>6</td>
<td>43.10</td>
<td>46.90</td>
<td>48.13</td>
</tr>
<tr>
<td>reord. training +</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>plain input</td>
<td>6</td>
<td>43.67</td>
<td>46.64</td>
<td>48.53</td>
</tr>
<tr>
<td>lattice</td>
<td>4</td>
<td>44.04</td>
<td>47.51</td>
<td>48.96</td>
</tr>
<tr>
<td>oracle reord.</td>
<td>4</td>
<td>44.36</td>
<td>48.25</td>
<td>49.26</td>
</tr>
</tbody>
</table>

+0.9/0.6/0.8% abs. improvement
Chunk-based reordering lattices: Evaluation

%BLEU scores on Eval08-NW, Reo08-NW (specific set containing only VSO sentences) and Eval09-NW:

<table>
<thead>
<tr>
<th>System</th>
<th>DL</th>
<th>eval08nw</th>
<th>reo08nw</th>
<th>eval09nw</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>6</td>
<td>43.10</td>
<td>46.90</td>
<td>48.13</td>
</tr>
<tr>
<td>reord. training +</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>plain input</td>
<td>6</td>
<td>43.67</td>
<td>46.64</td>
<td>48.53</td>
</tr>
<tr>
<td>lattice</td>
<td>4</td>
<td>44.04</td>
<td>47.51</td>
<td>48.96</td>
</tr>
<tr>
<td>oracle reord.</td>
<td>4</td>
<td>44.36</td>
<td>48.25</td>
<td>49.26</td>
</tr>
</tbody>
</table>

+0.9/0.6/0.8% abs. improvement

Gap between baseline and oracle is largely but not totally filled
Conclusions

• We have focused on a class of significant reorderings
• analysed their distribution and measured their impact on SMT
• developed techniques to:
  ➢ verb-reorder the parallel data,
  ➢ represent likely verb movements in test sentences
• Positive results ( +0.8% BLEU on Nist09-NW ) but further improvement possible
Conclusions

Future work will include:

• devising more discriminative weighting scheme for lattices
• evaluating with reordering-specific metrics by Birch & al. (2010)
• developing linguistically informed reordering constraints, alternative to lattices
Appendices
Improved MT outputs

**src:** w A$Ar AIsnAtrw AIY dEm h m$srwEA ErD ElY mjls Al$ywX

**ref:** The Senator referred to his support for a project proposed to the Senate

**base MT:** The Senator to support projects presented to the Senate

**new MT:** Senator noted his support projects presented to the Senate

---

**src:** mn jAnb h hdd >bw m$SEb EbdAlwdwd Amyr Al$qAEdp b blAd Almgrb AlAslAmy fy nfs Al$sryT b AlqyAm b sslp AEtdA’At w >EmAl <rhAbyp Dd AlmSAIh w Alm&ssAt AljzA} ryp fy AlEdyd mn AlmnATq AljzA} ryp

**ref:** For his part, Abu Musab Abd al-Wadud, the commander of al-Qaeda in the Islamic Maghreb Countries, threatened in the same tape to carry out a series of attacks and terrorist actions against Algerian interests and organizations in many parts of Algeria

**base MT:** For his part threatened Abu Musab EbdAlwdwd Amir al-Qaeda Islamic Morocco country in the same tape to carry out a series of attacks and terrorist acts against the interests and the Algerian institutions in many areas of Algiers

**new MT:** For his part, Abu Musab EbdAlwdwd Amir al Qaida threatened to Morocco Islamic country in the same tape to carry out a series of attacks and terrorist acts against the interests of the Algerian and institutions in many areas of Algiers

---

**src:** w ymdt Alm$rwE 500 km mtr w yrbT Almdyntyn Almqdstyn b mdynp jdp ElY sAHl AlbHr Al>Hmr .

**ref:** The project is 500 kilometers long and connects the two holy cities with the city of Jeddah on the Red Sea coast.

**base MT:** It extends the project 500 km and linking the two holy cities in the city of Jeddah on the Red Sea coast.

**new MT:** The project extends 500 km, linking the two holy cities in the city of Jeddah on the Red Sea coast.
Evaluation (also on No-reo-08)

%BLEU scores on:

Eval08-NW, Reo08-NW (only sentences needing reordering), No-reo08-NW (the other sentences), and Eval09-NW:

<table>
<thead>
<tr>
<th>System</th>
<th>DL</th>
<th>eval08nw</th>
<th>reo08nw</th>
<th>no-reo08nw</th>
<th>eval09nw</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>6</td>
<td>43.10</td>
<td>46.90</td>
<td>40.68</td>
<td>48.13</td>
</tr>
<tr>
<td>reord. training +</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>plain input</td>
<td>6</td>
<td>43.67</td>
<td>46.64</td>
<td>41.79</td>
<td>48.53</td>
</tr>
<tr>
<td>lattice</td>
<td>4</td>
<td>44.04</td>
<td>47.51</td>
<td>41.83</td>
<td>48.96</td>
</tr>
<tr>
<td>oracle reord.</td>
<td>4</td>
<td>44.36</td>
<td>48.25</td>
<td>41.79</td>
<td>49.26</td>
</tr>
</tbody>
</table>

Table 2: BLEU scores of baseline and reordered system on plain test and on reordering lattices with no-reo08.
Evaluation (with L-weights)

%BLEU scores on:

Eval08-NW, Reo08-NW (only sentences needing reordering),
No-reo08-NW (the other sentences), and Eval09-NW

Length-based edge weighting scheme (L-weights):

<table>
<thead>
<tr>
<th>System</th>
<th>DL</th>
<th>eval08nw</th>
<th>reo08nw</th>
<th>no-reo08nw</th>
<th>eval09nw</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>6</td>
<td>43.10</td>
<td>46.90</td>
<td>40.68</td>
<td>48.13</td>
</tr>
<tr>
<td>reord. training +</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>plain input</td>
<td>6</td>
<td>43.67</td>
<td>46.64</td>
<td>41.79</td>
<td>48.53</td>
</tr>
<tr>
<td>lattice</td>
<td>4</td>
<td>44.04</td>
<td>47.51</td>
<td>41.83</td>
<td>48.96</td>
</tr>
<tr>
<td>lattice(Lweights)</td>
<td>4</td>
<td>44.18</td>
<td>47.40</td>
<td>42.13</td>
<td>49.06</td>
</tr>
<tr>
<td>oracle reord.</td>
<td>4</td>
<td>44.36</td>
<td>48.25</td>
<td>41.79</td>
<td>49.26</td>
</tr>
</tbody>
</table>