Comparing Constituency and Dependency Representations for SMT Phrase Extraction

Sylwia Ozdowska
(joint work with Mary Hearne and John Tinsley)

23rd July 2008
Phrase-pair Extraction and Syntax

- Standard technique to induce translation models (string-based phrase-pairs) not syntax-aware (Koehn et al. 03)

  Q: What about adding syntax into the phrase extraction process?

- Combining constituency-based phrase pairs with string-based phrase pairs improves translation quality (Tinsley et al. 07)

  Q: What about using dependency-based phrase pairs?
    - What is the value of replacing and/or combining string-based methods with syntax-based methods for PB-SMT?
    - What are the relative merits of using constituency-annotated vs. dependency-annotated training data?
Constituency parses

- Constituency parses are context-free phrase-structure trees.
- They make explicit syntactic constituents such as noun phrases (e.g. *the commission*), verb phrases (e.g. *intends to take...*), prepositional phrases (e.g. *to protect it*), etc.

```
The Commission
  
intends
  
to
  
  take
  
  measures
  
    to
    
    protect
    
    it
```
Dependency parses make explicit the relationships between the words in terms of heads and dependents and possibly the nature of the relationship, *i.e.* subject (*e.g.* Commission depends on intends via the subject relation), object (*e.g.* measures depends on take via the object relation), etc.

The Commission intends to take measures to protect it
Recap: constituency parses and dependency parses

Constituency parses

```
The Commission intends to take measures to protect it
```

Dependency parses

```
The Commission intends to take measures to protect it
```
 Parsing

- **Corpus**
  - JOC English–French parallel corpus (Chiao et al. 06)
  - 8,759 aligned sentences (7722 training pairs and 1000 test pairs)

- **Constituency annotation: Bikel’s statistical parser (Bikel 02)**
  - English: training on the Penn II Treebank (Marcus et al. 96)
  - French: training on the Modified French Treebank (Schulter & Genabith 07)

- **Dependency annotation: Syntex parsers (Bourigault et al. 05)**
String-based alignment and phrase extraction (STR)

- Giza++ (Och & Ney 03), Moses (Koehn et al. 07)
- Extraction of word and phrase-pairs consistent with the word alignment

La Commission a l’intention de prendre des mesures pour assurer sa protection

The Commission intends to take measures to protect it

a l’intention ⇔ intends
a l’intention de ⇔ intends to
commission a l’intention de ⇔ commission intends to
la commission a l’intention de ⇔ the commission intends to
Syntax-based alignment and phrase-pair extraction

- TreeAligner (Tinsley et al. 07)
- Extraction of all string pairs dominated by linked constituents
de prendre des mesures pour assurer sa protection ⇔ to take measures to protect it
pour assurer sa protection ⇔ to protect it
des mesures ⇔ measures
La Commission a l’intention de prendre des mesures pour assurer sa protection

The Commission intends to take measures to protect it

a l’intention de prendre des mesures pour assurer sa protection ⇔

intends to take measures to protect it

sa protection ⇔ protect it

mesures ⇔ measures
Recap: **STR, CON and DEP**

- **STR**
  
  *La Commission a l'intention de prendre des mesures pour assurer sa protection*

  *The Commission intends to take measures to protect it*

- **CON** (Bikel’s parser)

  - *La Commission a l’intention de prendre des mesures pour assurer sa protection*
  - *The Commission intends to take measures to protect it*

- **DEP** (Syntex)

  - *La Commission a l’intention de prendre des mesures pour assurer sa protection*
  - *The Commission intends to take measures to protect it*
Translation

- 1000 sentences
- Moses decoder (Koehn et al. 07)
- Different phrase-table configurations
  - STR|CON|DEP: each phrase table used individually
  - STR+(CON|DEP): combination of STR with either CON or DEP
  - STR+CON+DEP: combination of all phrase-pairs
## Results

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>STR</td>
<td>30.35</td>
<td>62.62</td>
<td>64.32</td>
</tr>
<tr>
<td>CON</td>
<td>29.82</td>
<td>62.96</td>
<td>63.53</td>
</tr>
<tr>
<td>DEP</td>
<td>29.80</td>
<td>63.07</td>
<td>64.00</td>
</tr>
<tr>
<td>STR+CON</td>
<td>31.88</td>
<td>65.04</td>
<td>65.59</td>
</tr>
<tr>
<td>STR+DEP</td>
<td>31.97</td>
<td>65.07</td>
<td>65.70</td>
</tr>
<tr>
<td>STR+CON+DEP</td>
<td>31.90</td>
<td>65.10</td>
<td>65.57</td>
</tr>
</tbody>
</table>

- Conflicting results across evaluation metrics
- Replacing STR with CON|DEP: translation quality — —
- Combining STR with CON|DEP: translation quality ++
- Translation quality CON < Translation quality DEP
Analysis

- **Phrase-pair length**
  - shorter phrase-pairs > greater impact on translation quality
  - unique phrase-pair average length: 9.98 for CON vs. 5.67 for DEP

- **Phrase-pair coverage**
  - higher alignment coverage > lower phrase-pair quality
  - linked phrase-pairs:
    66,601 En and 67,280 Fr for CON
    64,904 En and 64,135 Fr for DEP

- **Parsers’ performance**
  - different monolingual parses > differing translation accuracies
  - parsers’ f-scores:
    90% En and 80% Fr for CON
    82% En and 89% Fr for DEP

- **Generated phrases**
  52% En and 46.6% Fr unique to CON
  38.9% En and 42.7% Fr unique to DEP
Conclusions
- Translation quality improves when combining STR with either CON or DEP
- General trend towards preferring dependency-based phrase-pairs over constituency-based phrase-pairs

Future work
- Gain further insights into the reasons why translation quality varies
- Analyse the relative impact of the different types of constituents for which phrase-pairs are extracted
- Scale up the experiments