Tree-based Alignment and Translation

John Tinsley, Ventsislav Zhechev, Mary Hearne and Andy Way

June 18, 2007
Outline

Parallel Treebanks

Automatic Tree-to-Tree Alignment

Future Work
Outline

Parallel Treebanks

Automatic Tree-to-Tree Alignment

Future Work
Parallel treebanks

A parallel treebank comprises:

- sentence pairs
- parsed
- word-aligned
- tree-aligned

(Volk & Samuelsson, 2004)
Parallel treebanks

Some areas of application for parallel treebanks are:

- training for data-driven machine translation
- knowledge source for transfer-rule induction
- reference for phrase-alignment
- knowledge source for corpus-based translation studies
- knowledge source for studies in contrastive linguistics
Automatic parallel treebank acquisition

Required elements:

- sentence pairs: increasing availability of bitexts, both manually and automatically constructed
- parsed: increasing availability of treebanks and trainable parsing resources
- word-aligned: availability of tools such as Giza++ and Moses
- tree-aligned: ???
Outline

Parallel Treebanks

Automatic Tree-to-Tree Alignment

Future Work
Principles:

- independence with respect to language pair and constituent labelling schema;
- preservation of the given tree structures;
- minimal external resources required;
- word-level alignments not fixed \textit{a priori}.
Tree-alignment algorithm (Tinsley et al., 2007)

Well-formedness criteria:

- a node can only be linked once;
- crossing constraints:
  - descendants of a source linked node may only link to descendants of its target linked counterpart;
  - ancestors of a source linked node may only link to ancestors of its target linked counterpart.
Tree-alignment algorithm (Tinsley et al., 2007)

Alignment algorithm:

- hypothesise initial alignments: each source node can link to any target node and vice versa;
- assign a score to each hypothesised alignment;
- select a set of links meeting the well-formedness criteria according to a greedy search.
Tree-alignment algorithm (Tinsley et al., 2007)
Tree-alignment algorithm (Tinsley et al., 2007)

Computing hypothesis scores:
Assume tree pair \(<S,T>\), hypothesis \(<s,t>\), the following strings and GIZA++ / Moses word-alignment probabilities.

\[
\begin{align*}
  s_l &= s_i \ldots s_{ix} & s_l^c &= S_1 \ldots s_{i-1}s_{ix+1} \ldots S_m \\
  t_l &= t_j \ldots t_{jx} & t_l^c &= T_1 \ldots t_{j-1}t_{jx+1} \ldots T_n
\end{align*}
\]

String correspondence score: 
\[
\alpha(x|y) = \prod_{j=1}^{|x|} \frac{\sum_{i=1}^{|y|} P(x_j|y_i)}{|y|}
\]

Hypothesis score: 
\[
\gamma(\langle s,t \rangle) = \alpha(s_l|t_l) \alpha(t_l|s_l) \alpha(s_l^c|t_l^c) \alpha(t_l^c|s_l^c)
\]
## Evaluation

<table>
<thead>
<tr>
<th>Configs</th>
<th>all links</th>
<th>lexical links</th>
<th>non-lexical links</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>scr1</td>
<td>0.6686</td>
<td>0.7733</td>
<td>0.5615</td>
</tr>
<tr>
<td>scr2</td>
<td>0.6736</td>
<td>0.7844</td>
<td>0.5722</td>
</tr>
<tr>
<td>scr1_sp1</td>
<td>0.6763</td>
<td><strong>0.8026</strong></td>
<td>0.5733</td>
</tr>
<tr>
<td>scr2_sp1</td>
<td><strong>0.6781</strong></td>
<td>0.7927</td>
<td><strong>0.5781</strong></td>
</tr>
</tbody>
</table>
### Translation Evaluation

<table>
<thead>
<tr>
<th>Configs</th>
<th>Bleu</th>
<th>NIST</th>
<th>Meteor</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>manual</td>
<td>0.5444</td>
<td>7.0701</td>
<td>0.7302</td>
<td>70.4167</td>
</tr>
<tr>
<td>scr1</td>
<td>0.5163</td>
<td>6.8685</td>
<td>0.7242</td>
<td>73.7500</td>
</tr>
<tr>
<td>scr2</td>
<td>0.5280</td>
<td>6.8869</td>
<td>0.7297</td>
<td>73.5417</td>
</tr>
<tr>
<td>scr1_sp1</td>
<td>0.5252</td>
<td>6.8842</td>
<td>0.7285</td>
<td>75.3125</td>
</tr>
<tr>
<td>scr2_sp1</td>
<td>0.5258</td>
<td>6.8506</td>
<td>0.7268</td>
<td>73.6458</td>
</tr>
</tbody>
</table>
Outline

Parallel Treebanks

Automatic Tree-to-Tree Alignment

Future Work
Improving alignment quality

- alternative scoring strategies
- alternative search strategies
- using monolingual parse n-best lists for increased flexibility
- larger-scale translation-based evaluation
Using tree-alignments for MT

Aligner output:

```
 VP  VP  V  PP
  VP    V   NP
    V  ADJ N
      D  ADJ N
          sur D N
              ADJ
```

the Save As button le bouton Enregistrer Sous

Resource extraction:

```
click      →      cliquez sur

click X   →      cliquez sur X

click NP  →      cliquez sur NP

VP → click NP  :  VP → cliquez sur NP
```
Generating and using alternative outputs

- generalise the existing aligner to the
  - string-to-tree
  - tree-to-string
  - string-to-string

cases, and develop corresponding resource-extraction methods.