Using alignment of
tectogrammatical trees in
phrase-based machine
translation

David Mareček
marecek@ufal.mff.cuni.cz

Statistical Machine Translation seminar
May 11, 2009, Prague
Outline

• Alignment of tectogrammatical trees and its advantages

• GIZA++ word alignment and symmetrization methods

• Combining various alignments

• Testing usability of these alignments on the MOSES toolkit
Alignment of Tectogrammatical Trees

Tectogrammatical tree = rooted dependency tree where only autosemantic (content) words have their own nodes.

Advantage of tectogrammatical alignment over word alignment:
• Functional words (e.g. articles, prepositions, auxiliary verbs, modal verbs ...), that are often problematic to align (they can have different functions in different languages), don’t have their own node in the tectogrammatical layer – we needn’t align them.
• The tree structure may help.

Disadvantages:
• Errors in tagging and parsing often causes errors in the alignment.
I have always been convinced that Milosevic should have been put on trial in Belgrade.

Vždy jsem byl přesvědčen, že Miloševič by měl být souzen v Bělehradě.
AER and IAA

• AER - Alignment error-rate
  • heuristic metric for word-alignment (Och and Ney, 2003)

• IAA - Inter-annotator alignment

• Our testset: 2,500 manually aligned sentences

<table>
<thead>
<tr>
<th>IAA</th>
<th>tool</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>word-alignment</td>
<td>89.6 %</td>
<td>13.2 %</td>
</tr>
<tr>
<td>tecto-alignment</td>
<td>94.6 %</td>
<td>7.3 %</td>
</tr>
</tbody>
</table>
Tecto-alignment on the Surface

• We link all content words that were linked in tectogrammatical trees.
• We can link also functional words:
  • Two words are linked if the content words they belong to are also linked.
GIZA++ Word alignment

- Sentences are lemmatized before running GIZA++.
- IBM models create a many-to-one mapping – asymmetric output.
- We run GIZA++ in both directions and symmetrize the outputs.
Symmetrization methods

union

intersection

grow-diag-final
Combining alignments

• We combine word alignments acquired from GIZA++ and tecto-alignment together

• Our goal is to get an alignment with lower AER or higher BLEU when applied to MOSES

• We used this combinations:
  • GIZA_int U tecto_lex
  • GIZA_gdfa U tecto_lex
  • GIZA_gdfa ∩ tecto_lex+aux
  • (GIZA_gdfa ∩ tecto_lex+aux) U GIZA_int
Running MOSES

• Direction EN → CZ

• Training, tuning and testing data from WMT08

• MERT tuning
## Results

<table>
<thead>
<tr>
<th>alignment</th>
<th>density</th>
<th>AER</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>giza_int</td>
<td>0.62</td>
<td>13.2</td>
<td>12.4</td>
</tr>
<tr>
<td>giza_gdfa</td>
<td>0.92</td>
<td>20.3</td>
<td>12.9</td>
</tr>
<tr>
<td>giza_int U tecto_lex</td>
<td>0.74</td>
<td>10.7</td>
<td>12.3</td>
</tr>
<tr>
<td>giza_gdfa U tecto_lex</td>
<td>1.00</td>
<td>17.8</td>
<td>12.7</td>
</tr>
<tr>
<td>tecto_lex+aux</td>
<td>1.43</td>
<td>37.8</td>
<td>8.9</td>
</tr>
<tr>
<td>giza_gdfa ∩ tecto_lex+aux</td>
<td>0.60</td>
<td>18.4</td>
<td>12.3</td>
</tr>
<tr>
<td>( giza_gdfa ∩ tecto_lex+aux ) U giza_int</td>
<td>0.71</td>
<td>19.6</td>
<td>12.8</td>
</tr>
</tbody>
</table>
Future Work

• Try other combinations of alignments

• Use bidirectional tectogrammatical alignment (one-to-many alignment in both directions)
  • First combine the two unidirectional alignments with GIZA++ alignments
  • Then symmetrize

• Extend MOSES phrase table - an information whether a phrase matches with tectogrammatical alignment or not