Evaluation of Sentence Alignment Systems

Sadaf Abdul-Rauf, Mark Fishel, Patrik Lambert, Sandra Noubours, Rico Sennrich

1 Introduction

- Sentence Alignment usually starts with a parallel text pre-aligned based on (meta-)textual information.
  - time stamp (subtitles)
  - speaker information (Europarl)
  - paragraphs/chapters/documents
- The more fine-grained we can align the text based on textual structure, the easier sentence alignment becomes

Project aim

- If textual structure only allows for a coarse-grained alignment, sentence alignment is more difficult.
- Example: NIST Urdu-English training data.
- Our goal: evaluate sentence alignment tools on document-aligned parallel corpora and impact on SMT

Alignment Tools

We evaluated five unsupervised sentence alignment tools

- Gale and Church (1993)
- Bilingual Sentence Aligner (Moore (2002))
- Hunalign (Varga et al. (2005))
- Gargantua (Braune and Fraser (2010))
- Bleualign (Sennrich and Volk (2010))

Gale and Church (1993)

- Probability score for each sentence pair based on sentence-length (number of characters)
- Dynamic programming to find maximum likelihood alignment
- Used for Europarl, JRC-acquis and others

Bilingual Sentence Aligner (Moore (2002))

- Combines sentence length based method with word correspondence
- First pass based on sentence length
- Second on IBM Model-1
- Sentence length method: based on the distribution of length variable
- Training of IBM Model-1: during runtime, uses alignments obtained from the first pass
- The larger corpus size the more effective (better model of distribution of word length variable and word correspondence)
Hunalign (Varga et al. (2005))

The algorithm:

• alignment matrix, diagonal ± 10%
• weights – combination of length-based (~ Gale’n’Church) and dictionary-based similarity
• if no dictionary:
  – do length-based, estimate dictionary from result
  – reiterate once

Problems:
• not designed to handle corpora of over 20k sentences
• copes by splitting larger corpora
• this causes worse dictionary estimates

Gargantua (Braune and Fraser (2010))

• alignment model similar to Moore (2002), but differences in pruning and search strategy
• according to (2010), main score improvement over Bilingual sentence aligner comes from also returning 1-N/N-1 alignments
• (Bilingual sentence aligner also calculates 1-N/N-1 alignments, but does not return them; can be hacked into program)

Bleualign (Sennrich and Volk (2010))

• based on automatic translation of source text
• dynamic programming to find path that maximizes BLEU score between target text and translation of source text
• central problem: how to get automatic translation?
  – alignment of training set with Gale and Church
  – training of SMT system with Moses
  – pruning of phrase table as in Johnson (2007)
  – translation of training set

2 Experiments

2.1 Urdu-English Task

Corpus:
• Training:
  – English: 89 323 sentences, 1 799 745 tokens
  – Urdu: 93 332 sentences, 2 026 633 tokens
  – 5282 documents (document length between 1 and 1003/878 sentences)
• Evaluation:
  – 1100 sentence pairs (29k Urdu tokens and 25k English tokens) for test and 932 for dev (26k Urdu tokens and 22k English tokens)
  – 152 documents
  – automatically aligned (Bleualign)
Coverage:

Table 1 indicates the coverage in terms of number of tokens achieved by the various aligners tested. The % columns indicate the percentage of tokens in the sentence aligned bitexts compared to the original amount. Note that the Microsoft Bilingual Aligner was used with its default precision threshold. A different threshold could have been used to obtain a higher coverage.

<table>
<thead>
<tr>
<th>Aligner</th>
<th>sentences/units (k)</th>
<th>tokens (k)</th>
<th>Urdu</th>
<th>%</th>
<th>English</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Urdu</td>
<td></td>
<td></td>
<td></td>
<td>English</td>
<td></td>
</tr>
<tr>
<td>Original</td>
<td>93.3</td>
<td>2027</td>
<td>100.0</td>
<td>1800</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>bleualign</td>
<td>65.6</td>
<td>1821</td>
<td>89.9</td>
<td>1607</td>
<td>89.3</td>
<td></td>
</tr>
<tr>
<td>galechurch</td>
<td>70.0</td>
<td>1925</td>
<td>95.0</td>
<td>1729</td>
<td>96.1</td>
<td></td>
</tr>
<tr>
<td>gargantua</td>
<td>71.1</td>
<td>1943</td>
<td>95.9</td>
<td>1737</td>
<td>96.5</td>
<td></td>
</tr>
<tr>
<td>hunalign</td>
<td>68.7</td>
<td>1950</td>
<td>96.2</td>
<td>1670</td>
<td>92.8</td>
<td></td>
</tr>
<tr>
<td>mba</td>
<td>40.3</td>
<td>902</td>
<td>44.5</td>
<td>745</td>
<td>41.4</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Coverage on Urdu-English data

SMT results:

The bitexts described in Table 1 were used to train phrase-based moses SMT systems. Table 2 shows the corresponding scores of automated MT metrics obtained. The evaluation was case-sensitive.

<table>
<thead>
<tr>
<th>Aligner</th>
<th>BLEU</th>
<th>METEOR-rank</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>bleualign</td>
<td>9.72 ±0.19</td>
<td>27.36 ±0.17</td>
<td>86.21 ±0.69</td>
</tr>
<tr>
<td>galechurch</td>
<td>8.14 ±0.04</td>
<td>26.32 ±0.04</td>
<td>87.86 ±0.33</td>
</tr>
<tr>
<td>gargantua</td>
<td>9.43 ±0.18</td>
<td>27.10 ±0.08</td>
<td>86.26 ±1.08</td>
</tr>
<tr>
<td>hunalign</td>
<td>9.11 ±0.27</td>
<td>26.77 ±0.26</td>
<td>87.03 ±0.22</td>
</tr>
<tr>
<td>mba</td>
<td>8.35 ±0.05</td>
<td>26.01 ±0.07</td>
<td>88.13 ±0.21</td>
</tr>
</tbody>
</table>

Table 2: SMT results on Urdu-English data. The values shown are the average and standard deviation over 3 MERT runs with different random seeds. The values in bold are possibly the best one taking the error range into account.

2.2 French-English Task

Corpus:

- Training:
  - NewsCommentary + BAF + Rapid
  - English: 187 213 sentences, 3 486 184 tokens
  - French: 187 656 sentences, 4 104 411 tokens
  - 3461 documents (document length between 1 and 7077/6890 sentences)

- Evaluation:
  - realized with WMT10 evaluation internal development (news-test2008) and test (newstest2009) sets

Coverage:

Table 3 indicates the coverage achieved by the various aligners tested on the French-English data.
Table 3: Coverage on French-English data

<table>
<thead>
<tr>
<th>Aligner</th>
<th>BLEU</th>
<th>METEOR-rank</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>bleualign</td>
<td>21.07</td>
<td>38.83 ± 0.15</td>
<td>61.23 ± 0.09</td>
</tr>
<tr>
<td>galechurch</td>
<td>20.64</td>
<td>38.54 ± 0.15</td>
<td>61.66 ± 0.23</td>
</tr>
<tr>
<td>gargantua</td>
<td>20.83</td>
<td>38.63 ± 0.04</td>
<td>61.13 ± 0.13</td>
</tr>
<tr>
<td>humalign</td>
<td>21.03</td>
<td>38.68 ± 0.10</td>
<td>60.94 ± 0.18</td>
</tr>
<tr>
<td>mba</td>
<td>20.91</td>
<td>38.85 ± 0.14</td>
<td>61.42 ± 0.22</td>
</tr>
</tbody>
</table>

Table 4: SMT results on French-English data. The values shown are the average and standard deviation over 3 MERT runs with different random seeds. The values in bold are possibly the best one taking the error range into account.

References


