Phrase-based alignment combining corpus cooccurrences and linguistic knowledge

Adrià de Gispert
José B. Mariño
Josep Maria Crego
Outline

- Introduction
- Proposed phrase alignment strategy
- Experimental results
- Discussion
- Further research
Outline

• **Introduction**
  – Motivation
  – Word and phrases association measures
• Proposed phrase alignment strategy
• Experimental results
• Discussion
• Further research
Motivation

- Word alignment is crucial to train SMT systems
- GIZA++ alignments are state-of-the-art, but...
  - Symmetrization strategies are non-linguistic
  - Model complexity to introduce additional knowledge
- Cooccurrence-based algorithms perform well too, but...
  - Their output must be a many-to-many alignment

Goal: phrase alignment following linguistic criteria
Word & phrase cooccurrence measures

- $\phi^2$ score, t-score, Dice, ...
- Can be computed between words but also phrases
- Phrase cooccurrence measures give complementary and stronger evidence

<table>
<thead>
<tr>
<th>please</th>
<th>favor</th>
<th>por</th>
<th>22.4</th>
<th>1.2</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>maybe</td>
<td>a</td>
<td>lo</td>
<td>23.1</td>
<td>18.2</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>lo</td>
<td>mejor</td>
<td>12.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Not efficient to compute for all possible phrase pairs
- A selection of candidate phrases is needed
Outline

- Introduction
- **Proposed phrase alignment strategy**
  - Candidate phrase selection and classification
  - Phrase-to-phrase alignment
  - Word alignment algorithm
- Experimental results
- Discussion
- Further research
Phrase alignment strategy

- Linguistically-guided selection of candidate phrases
- Verb groups and idiomatic expressions
- Add knowledge limiting cooc. counts table size
- $\phi^2$-based competitive linking until threshold
- Very-high precision required
- One-to-one word alignment with unaligned tokens
- Final global decisions on word alignment

Four stages:

1. Phrase selection (classification)
2. Phrase alignment
3. Word alignment
4. Post-processing
Candidate selection: Verbs

- **Rule-based** detection
  - Using word, POS and base form
  - Classification according to head verb base form
  - Check base forms against lists to avoid tagging errors

- Single-word verbs substituted by base form
- Reduction in cooc. table size
- Limit: Base form ambiguity not tackled
Candidate selection: Idioms

- Lists of **frequently-used idioms**
  - Spanish: 1496 idioms
  - English: 49 idioms
- No further classification
  - Compute coocs. against all other language tokens
  - Slight increase in cooc. table size

\[ \phi^2 \text{ ("idiom",x)} \]
Phrase-to-phrase alignment

- Competitive linking strategy until threshold is met
- Verb groups and idioms treated separately
- Example

\[
\begin{align*}
\phi^2 ("how many", \text{cuántas}) &= 2.5 \\
\phi^2 ("how many", \text{habitaciones}) &= 23.0 \\
\phi^2 ("how many", "BF(necesitar)") &= 33.4 \\
\phi^2 ("BF(need)", \text{cuántas}) &= 31.05 \\
\phi^2 ("BF(need)", \text{habitaciones}) &= 19 \\
\phi^2 ("BF(need)", "BF(necesitar)") &= 0.9
\end{align*}
\]
Word alignment algorithm

- One-to-one alignment
- Iterative best-first search
- Heuristic based on link probabilities
  - Initial alignment generated using $\phi^2$ scores
  - Estimate link probabilities
  - Realignment using new estimates
- Syntax-guided coherence constraint included
Outline

- Introduction
- Proposed phrase alignment strategy
- Experimental results
  - Data used
  - Partial results: phrase alignment
  - Complete AER results
- Discussion
- Further research
**Data used**

- **Verbmobil Spa-Eng corpus** 30K sentences

<table>
<thead>
<tr>
<th></th>
<th>words</th>
<th>vocab</th>
<th>singlet.</th>
<th>Lmax</th>
<th>Lavg</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>230 K</td>
<td>3.2 K</td>
<td>39 %</td>
<td>66</td>
<td>7.6</td>
</tr>
<tr>
<td>Spanish</td>
<td>220 K</td>
<td>5.0 K</td>
<td>43 %</td>
<td>66</td>
<td>7.3</td>
</tr>
</tbody>
</table>

- Preprocessing
  - Normalization of contracted forms  
    we've = we have / del = de el
  - Tagging and base form  
    Eng: TnT + wnmorph / Spa: maco+ relax
  - Date and time expressions
  - No punctuation

- Evaluation scheme with AER
  - Dev. + test sets: 100 + 400 sentences
  - Manual alignment (80% Sure, 20% Poss) stress on Recall
Partial results: phrase alignment

• Results before word alignment

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbs $\phi^2 &lt; 8$</td>
<td>8.07</td>
<td>99.02</td>
</tr>
<tr>
<td>Verbs $\phi^2 &lt; 10$</td>
<td>9.00</td>
<td>99.12</td>
</tr>
<tr>
<td>Verbs $\phi^2 &lt; 15$</td>
<td>9.68</td>
<td>98.69</td>
</tr>
<tr>
<td>Idioms $\phi^2 &lt; 5$</td>
<td>2.01</td>
<td>98.48</td>
</tr>
<tr>
<td>Idioms $\phi^2 &lt; 10$</td>
<td>3.06</td>
<td>99.00</td>
</tr>
<tr>
<td>Idioms $\phi^2 &lt; 15$</td>
<td>3.50</td>
<td>97.41</td>
</tr>
</tbody>
</table>

• Straightforward approach, but ...
  - About 10% Recall at nearly no Precision cost
  - Complementary links between Verbs and Idioms
  - Complexity reduction for word alignment algorithm
Complete AER results

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>giza++ eng2spa</td>
<td>76.99</td>
<td>93.15</td>
<td>15.51</td>
</tr>
<tr>
<td>giza++ spa2eng</td>
<td>78.75</td>
<td>94.19</td>
<td>13.94</td>
</tr>
<tr>
<td>giza++ union</td>
<td><strong>84.47</strong></td>
<td><strong>90.85</strong></td>
<td><strong>12.30</strong></td>
</tr>
<tr>
<td>giza++ intersection</td>
<td>71.27</td>
<td><strong>97.58</strong></td>
<td>17.52</td>
</tr>
</tbody>
</table>

- union: precision loss, but very high recall
Experimental results

Complete AER results

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>giza++ eng2spa</td>
<td>76.99</td>
<td>93.15</td>
<td>15.51</td>
</tr>
<tr>
<td>giza++ spa2eng</td>
<td>78.75</td>
<td>94.19</td>
<td>13.94</td>
</tr>
<tr>
<td>giza++ union</td>
<td>84.47</td>
<td>90.85</td>
<td>12.30</td>
</tr>
<tr>
<td>giza++ intersection</td>
<td>71.27</td>
<td>97.58</td>
<td>17.52</td>
</tr>
<tr>
<td>one-to-one word aligner</td>
<td>72.56</td>
<td>96.69</td>
<td>16.96</td>
</tr>
</tbody>
</table>

- union: precision loss, but very high recall
- intersection vs. one-to-one aligner
Complete AER results

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>giza++ eng2spa</td>
<td>76.99</td>
<td>93.15</td>
<td>15.51</td>
</tr>
<tr>
<td>giza++ spa2eng</td>
<td>78.75</td>
<td>94.19</td>
<td>13.94</td>
</tr>
<tr>
<td>giza++ union</td>
<td>84.47</td>
<td>90.85</td>
<td>12.30</td>
</tr>
<tr>
<td>giza++ intersection</td>
<td>71.27</td>
<td>97.58</td>
<td>17.52</td>
</tr>
<tr>
<td>one-to-one word aligner</td>
<td>72.56</td>
<td>96.69</td>
<td>16.96</td>
</tr>
<tr>
<td>phrase aligner $\phi^2 &lt; 10$</td>
<td>76.31</td>
<td>97.48</td>
<td>13.36</td>
</tr>
<tr>
<td>phrase aligner $\phi^2 &lt; 15$</td>
<td>76.88</td>
<td>97.35</td>
<td>13.20</td>
</tr>
</tbody>
</table>

- intersection vs. one-to-one aligner
- union: precision loss, but very high recall
- proposed: high-precision, much higher recall
- phrase alignment is accurate and helps word alignment algorithm to perform better
Outline

- Introduction
- Proposed phrase alignment strategy
- Experimental results
- Discussion
- Further research
Discussion

• Promising results
  – competitive results still making small use of ling. knowledge
  – open to new knowledge sources
• Evaluation in translation task
• Evaluation with other corpora
Outline

- Introduction
- Proposed phrase alignment strategy
- Experimental results
- Discussion
- Further research
Further research

- Postprocessing techniques
- Extension of phrase detection rules
  - 'Gapped' structures
- Ambiguity in classifying detected phrases
  - numbers, times, different head verbs,...
- Training data reduction

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>phrase aligner $\phi^2 &lt; 15$</td>
<td>76.88</td>
<td>97.35</td>
<td>13.20</td>
</tr>
<tr>
<td>+ Gapped verbs</td>
<td>77.67</td>
<td>97.55</td>
<td>12.85</td>
</tr>
</tbody>
</table>
Thanks for attention

Centre de Tecnologies i Aplicacions del Llenguatge i la Parla
TALP Research Center
Universitat Politècnica de Catalunya (UPC)
Barcelona