Bilingual segmentation for phrasetable pruning in Statistical Machine Translation

Germán Sanchis-Trilles    Daniel Ortiz-Martínez    Jesús González-Rubio
Jorge González           Francisco Casacuberta

{gsanchis,dortiz,jegonzalez,jgonzalez,fcn}@dsic.upv.es

Departamento de Sistemas Informáticos y Computación
Instituto Tecnológico de Informática • Universidad Politécnica de Valencia
Motivation

• Typical SMT systems require inferring huge tables of phrase pairs

• Large phrasetables lead to an elevated computational cost

• Bottleneck for the widespread application of SMT in portable devices

• Remove phrase pairs that have no influence on final translation

• Develop phrasetable pruning technique that:
  – is straightforward
  – is independent on the extraction algorithm
  – does not affect translation quality
Introduction

- Fundamental equation of SMT:

\[ \hat{e} = \arg\max_e Pr(e|f) \approx \arg\max_e \sum_{m=1}^{M} \lambda_m h_m(f, e) \]

- Current SMT systems strongly based on phrases (i.e. word sequences)

- Work performed in PB models and PBSFSTs

- Phrase-extraction obtains multiple overlapping segmentations per sentence pair
  \( \rightarrow \) reduce redundancy
Bilingual segmentation

- Selecting a single bilingual segmentation per sentence is a difficult problem.

- In SMT, bilingual segmentation can be derived from phrase-based alignment.

  * Words are aligned into phrases, building supersets.

  * The best phrase-alignment can be defined as:

    \[
    \tilde{A}_V(f, e) = \arg\max_{\tilde{\alpha}} p(\tilde{\alpha}|f, e) \tag{1}
    \]

Search problem?
Bilingual segmentation: coverage problem

pronunciarme
puedo
no,
momento
de,
asi,
cannot say anything at this stage.

EAMT 2011 • May 31, 2010
Bilingual segmentation: coverage problem

\[ \tilde{A}_V(f, e) = \arg\max_{\tilde{a}} p(\tilde{a}|f, e) \]

Search problem?
Outline

- Motivation
- Introduction
- Bilingual segmentation
  - True bilingual segmentation
  - Source-driven bilingual segmentation
- Experiments
  - Phrase-based models
  - Phrase-based SFSTs
- Conclusions
True bilingual segmentation

• If output sentence is fixed, coverage problems imply that smoothing is needed

• Use a log-linear model to control different aspects of the segmentation

\[
\tilde{A}_V(f, e) = \arg\max_{\tilde{a}} p(\tilde{a} | f, e) = \arg\max_{\tilde{a}} p(\tilde{a}, e | f)
\]

→ Not a decoding problem, since maximisation takes place only over alignments

→ However, underlying log-linear model not the same as in decoding time

• Once optimal segmentations are available, a new phrasetable can be built

• New phrases are introduced as a side-effect of smoothing
**Source-driven bilingual segmentation**

- True bilingual segmentation solves the coverage problem and fixes the output sentence.
- In translation time, such restriction may introduce an inappropriate bias:
  - Score function is modified due to smoothing.
  - New phrase pairs are introduced.
- Heuristic algorithm has proved to provide appropriate bilingual phrases.
- Relax the output sentence restriction:

  \[
  \tilde{A}_V(f, e) \approx \arg\max_{\tilde{a}, e} p(\tilde{a} | f, e) = \arg\max_{\tilde{a}, e} p(\tilde{a}, e | f) 
  \]

  \[\Rightarrow \text{SMT search problem}\]

- Output sentence is allowed to be different from reference.
- Only segments in the current phrasetable are used.
- Segmentation induced by input sentence.
Experimental setup

- Experiments conducted by means of Thot & GREAT toolkits toolkit
- Similar experiments with Moses led to the same conclusions
- Translation quality measured with BLEU, TER and speedup ($S_p = T_b/T_r$)
- Experiments conducted on Europarl
- 95% level confidence intervals were about 0.65 points in every case
## Experimental setup

<table>
<thead>
<tr>
<th>Subset features</th>
<th>De</th>
<th>En</th>
<th>Es</th>
<th>En</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>751k</td>
<td></td>
<td>731k</td>
<td></td>
</tr>
<tr>
<td>Run. words</td>
<td>15.3M</td>
<td>16.1M</td>
<td>15.7M</td>
<td>15.2M</td>
</tr>
<tr>
<td>Mean length</td>
<td>20.3</td>
<td>21.4</td>
<td>21.5</td>
<td>20.8</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>195k</td>
<td>66k</td>
<td>103k</td>
<td>64k</td>
</tr>
<tr>
<td><strong>Development</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>2000</td>
<td></td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td>Run. words</td>
<td>55k</td>
<td>59k</td>
<td>61k</td>
<td>59k</td>
</tr>
<tr>
<td>Mean length</td>
<td>27.6</td>
<td>29.3</td>
<td>30.3</td>
<td>29.3</td>
</tr>
<tr>
<td>OoV words</td>
<td>432</td>
<td>125</td>
<td>208</td>
<td>127</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>3064</td>
<td></td>
<td>3064</td>
<td></td>
</tr>
<tr>
<td>Run. words</td>
<td>82k</td>
<td>85k</td>
<td>92k</td>
<td>85k</td>
</tr>
<tr>
<td>Mean length</td>
<td>26.9</td>
<td>27.8</td>
<td>29.9</td>
<td>27.8</td>
</tr>
<tr>
<td>OoV words</td>
<td>1020</td>
<td>488</td>
<td>470</td>
<td>502</td>
</tr>
</tbody>
</table>
## Results: Phrase-based models

<table>
<thead>
<tr>
<th>Pair</th>
<th>Baseline BLEU</th>
<th>Baseline w/s</th>
<th>Source-driven BLEU</th>
<th>Source-driven w/s</th>
<th>True BLEU</th>
<th>True w/s</th>
<th>True $S_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Es–En</td>
<td>28.2</td>
<td>93</td>
<td>27.5</td>
<td>1500</td>
<td>23.8</td>
<td>380</td>
<td>4</td>
</tr>
<tr>
<td>En–Es</td>
<td>27.6</td>
<td>76</td>
<td>27.2</td>
<td>700</td>
<td>24.7</td>
<td>250</td>
<td>3</td>
</tr>
<tr>
<td>De–En</td>
<td>21.6</td>
<td>100</td>
<td>21.1</td>
<td>1500</td>
<td>17.5</td>
<td>280</td>
<td>3</td>
</tr>
<tr>
<td>En–De</td>
<td>15.2</td>
<td>46</td>
<td>15.1</td>
<td>400</td>
<td>14.7</td>
<td>170</td>
<td>4</td>
</tr>
</tbody>
</table>

- For source-driven segmentation:
  - BLEU (not significantly) lower, TER unaltered
  - Number of parameters reduced by two orders of magnitude (±2% of original)
  - Translation speed increased by a factor of 9–16

- For true segmentation:
  - Translation quality drops significantly
Results: Phrase-based SFSTs

<table>
<thead>
<tr>
<th>Pair</th>
<th>Source-driven</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>(w/s)</td>
<td>(S_p)</td>
<td>PB</td>
</tr>
<tr>
<td>Es–En</td>
<td>25.8</td>
<td>92k</td>
<td>986</td>
<td>(28.2)</td>
</tr>
<tr>
<td>En–Es</td>
<td>25.3</td>
<td>28k</td>
<td>374</td>
<td>(27.6)</td>
</tr>
<tr>
<td>De–En</td>
<td>18.8</td>
<td>41k</td>
<td>412</td>
<td>(21.6)</td>
</tr>
<tr>
<td>En–De</td>
<td>13.0</td>
<td>14k</td>
<td>309</td>
<td>(15.2)</td>
</tr>
</tbody>
</table>

- PBSFSTs require monotonic bilingual segmentation (no "baseline")
- Baseline PB models produce better translation quality (although with more models)
- Speed increased by almost two orders of magnitude (more)
Conclusions

• Technique for reducing size of phrasetables

• Select most probable phrase pairs in a Viterbi fashion

• Source-driven segmentation leads to important improvements in decoding speed
  – Subset of original phrasetable

• True bilingual segmentation provides worse translation results:
  – Smoothing techniques are introduced
  – New phrase pairs introduced (10%–50%)
  – Important role in estimation of new model parameters

• Further work needed to understand true bilingual segmentation
Questions? Comments? Suggestions?