Statistical Phrase-Based Translation

Philipp Koehn, Franz Och, Daniel Marcu

koehn@isi.edu, och@isi.edu, marcu@isi.edu

Information Sciences Institute
University of Southern California
Motivation

- Phrase-based translation is the **best way** to do statistical machine translation
  - best performance in recent DARPA evaluations
  - also fairly simple
  - tools are freely available

- How do I construct a phrase translation table?
Goals

- Compare different approaches to learn phrases
- Examine properties of phrase-based translation
- Syntax and phrases
Overview

- Evaluation framework
  - unified model
  - decoder
  - corpus

- Three methods for learning phrases
  - word-alignment induced phrases
  - syntactic phrases
  - phrase-alignment

- Experiments
Statistical Phrase-Based Translation

Model

- Bayes rule: \( \text{argmax}_e p(e|f) = \text{argmax}_e p(f|e)p(e) \)
- Foreign sentence \( f \) is segmented into \( I \) phrases \( f_1^I \)
- Each phrase is translated with \( \phi(f_i|\bar{e}_i) \)
- Phrases are reordered with \( d(\cdot) \)
- Use of language model \( p_{\text{LM}}(e) \) and word penalty \( \omega^{|e|} \)
**Decoder: Beam Search**

- Build English by hypothesis expansion
  - from left to right
  - search space exponential with sentence length
  - reduction by pruning weak hypothesis aided by future cost estimate
Evaluation on Europarl Corpus

- Collected from the European Parliament Proceedings
  - Available at http://www.isi.edu/~koehn/
  - 11 languages, 20 million words each

- Test set
  - German-English
  - 1755 sentence of length 5-15
Three Methods for Learning Phrases

- Word-alignment induced phrases
  - similar to alignment templates [Och et al., 1999]

- Syntactic phrases
  - only syntactic phrases are learned
  - same restriction as in recently proposed syntactic transfer models

- Phrase-alignment
  - joint model [Marcu and Wong, 2002]
Word Alignment Induced Phrases

- Word alignment is generated using IBM Model 4
  - bidirectional alignments \( e \rightarrow f, f \rightarrow e \)
  - intersect alignments
  - grow additional alignment points with heuristics

- Collect phrase pairs consistent with word alignment

[Och et al., 1999]
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch),
(verde, green), (Maria no, Mary did not), (no daba una bofetada, did not slap),
(daba una bofetada a la, slap the), (bruja verde, green witch),
(Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la bruja verde, Mary did not slap the),
(daba una bofetada a la bruja verde, slap the green witch),
(no daba una bofetada a la bruja verde, did not slap the green witch),
(Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)
Syntactic Phrases

- Syntactic phrases span whole constituents in parse tree
- Motivation
  - only these phrases used syntactic transfer models, e.g., [Yamada and Knight, 2002]
  - does syntax help or hurt?
- Extract syntactic phrase pairs
  - parse both sides (with statistical parsers)
  - use word alignment as before
  - limit to phrases to syntactic constituents in parse tree
**Phrase Alignment**

- **Direct Phrase Alignment of Parallel Corpus**
  
  [Marcu and Wong, 2002]

- **Generative Story**
  
  - a number of concepts are created
  
  - each concept generates a foreign and English phrase

```
Morgen  fliege ich nach Kanada zur Konferenz
```

```
Tomorrow I will fly to the conference in Canada
```
Experiments

- Comparison of core methods
- Maximum phrase length
- Lexical weighting
- Phrase extraction heuristics
- Simpler word alignment models
- Other language pairs
Comparison of Core Methods

- Same decoder, same training data, same language model
  - except for IBM Model 4: uses greedy decoder [Germann et al., 2001]

- WAIPh best, syntactic phrases very bad

- All following experiments on WAIPh only
Maximum Phrase Length

- Maximum limit on length of phrases
  - higher limit → larger phrase translation table
  - all tables still fit into memory of modern machines

<table>
<thead>
<tr>
<th>Max. Length</th>
<th>10k</th>
<th>20k</th>
<th>40k</th>
<th>80k</th>
<th>160k</th>
<th>320k</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>37k</td>
<td>70k</td>
<td>135k</td>
<td>250k</td>
<td>474k</td>
<td>882k</td>
</tr>
<tr>
<td>3</td>
<td>63k</td>
<td>128k</td>
<td>261k</td>
<td>509k</td>
<td>1028k</td>
<td>1996k</td>
</tr>
<tr>
<td>4</td>
<td>84k</td>
<td>176k</td>
<td>370k</td>
<td>736k</td>
<td>1536k</td>
<td>3152k</td>
</tr>
<tr>
<td>5</td>
<td>101k</td>
<td>215k</td>
<td>459k</td>
<td>925k</td>
<td>1968k</td>
<td>4119k</td>
</tr>
<tr>
<td>7</td>
<td>130k</td>
<td>278k</td>
<td>605k</td>
<td>1217k</td>
<td>2657k</td>
<td>5663k</td>
</tr>
</tbody>
</table>
Maximum Phrase Length (2)

- Impact of limit on translation quality
  - not much improvement if maximum length is extended beyond 3
  - independent of training corpus size

![Graph showing BLEU scores for different maximum phrase lengths against training corpus size](image-url)
Lexical Weighting

- Augment phrase translation probability $\phi(\vec{f}|\vec{e})$ with lexical translation probabilities $w(f|e)$

<table>
<thead>
<tr>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>green</td>
<td>---</td>
<td>###</td>
</tr>
<tr>
<td>witch</td>
<td>---</td>
<td>###</td>
</tr>
</tbody>
</table>

Lexical weight:

$$p_w = w(\text{la}|\text{the}) \times w(\text{bruja}|\text{witch}) \times w(\text{verde}|\text{green})$$
Lexical Weighting

- Improves translation quality

![Graph showing the improvement of BLEU score with increasing training corpus size, with and without lexical weighting.](chart.png)
Phrase Extraction Heuristics

- Recall: word alignment based on intersection of bidirectional IBM Model 4 alignments + heuristics
Phrase Extraction Heuristics (2)

- Different phrases are learned, if heuristic to create word alignment is changed.

- Variations in heuristics:
  - only to directly neighboring
  - also to diagonally neighboring
  - also to non-neighboring
  - prefer English-foreign or foreign-to-English
  - use lexical probabilities or frequencies
  - extend only to unaligned words
  - ...
Phrase Extraction Heuristics (3)

- No clear advantage to any strategy
  - large differences, but ...
  - ... depending on corpus size
  - ... depending on language pair

---

**Diagram:**

- **x-axis:** Training Corpus Size
- **y-axis:** BLEU

Legend:
- diag-and
- diag
- base
- e2f
- f2e
- union

Philipp Koehn, Franz Och, Daniel Marcu – USC/ISI
Simpler Word Alignment Models

- Using simpler IBM Models for word alignment
  - not much impact, if simpler models used
  - simpler models computationally much cheaper
Other Language Pairs

- Finding hold for other language pairs, other corpora
  - Phrase translation better than IBM Model 4
  - Lexicalization helps (about +0.01 BLEU)

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Model4</th>
<th>Phrase</th>
<th>Lex</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-German</td>
<td>0.2040</td>
<td>0.2361</td>
<td>0.2449</td>
</tr>
<tr>
<td>French-English</td>
<td>0.2787</td>
<td>0.3294</td>
<td>0.3389</td>
</tr>
<tr>
<td>English-French</td>
<td>0.2555</td>
<td>0.3145</td>
<td>0.3247</td>
</tr>
<tr>
<td>Finnish-English</td>
<td>0.2178</td>
<td>0.2742</td>
<td>0.2806</td>
</tr>
<tr>
<td>Swedish-English</td>
<td>0.3137</td>
<td>0.3459</td>
<td>0.3554</td>
</tr>
<tr>
<td>Chinese-English</td>
<td>0.1190</td>
<td>0.1395</td>
<td>0.1418</td>
</tr>
</tbody>
</table>
Conclusions

- Phrase-based translation better than word-based translation
- Limit to syntactic phrases hurts a lot
- Small phrases (up to 3 words) good enough
- Lexical weighting helpful
- Phrase extraction heuristics matter, but best heuristics vary on corpus size, language pair