Using Parsed Corpora for Estimating Stochastic Inversion
Transduction Grammars

Universidad Politécnica de Valencia

Instituto Tecnológico de Informática

Germán Sanchis, Joan Andreu Sánchez

{gsanchis,jandreu}@dsic.upv.es

May 2008
Index

1 Statistical Machine Translation ▶ 1

2 Phrase Based models ▶ 2

3 Stochastic Inversion Transduction Grammars ▶ 3

4 SITGs for phrase extraction ▶ 5

5 Experimental results ▶ 8

6 Discussion ▶ 10

7 Conclusions ▶ 11
Statistical Machine Translation

- SMT: efficient framework for building state-of-the-art MT systems.

- Problem originally defined as

  $$\hat{y} = \arg\max_y \Pr(y|x)$$
  $$= \arg\max_y \Pr(x|y) \cdot \Pr(y)$$

- In practice, $\Pr(y|x)$ is modelled using log-linear models:

  $$\hat{y} = \arg\max_y \sum_{m=1}^{M} \lambda_m h_m(x, y)$$
Phrase-Based models

- Systems implementing PB models are dominant in the state of the art.

- Basic translation units are bilingual phrases (segments), not single words.

- In training time, bilingual segments must be extracted: lots of techniques.

- Most common approach:
  - Heuristical extraction of phrases using word alignments.
  - Let be \((s, t) = (x_{i+1}, y_{k+1})\)
  - 5 models: \(p_c(s|t), p_c(t|s), lex(s|t), lex(t|s), C\).
Stochastic Inversion Transduction Grammars

- Originally proposed by Dekai Wu.
- Closely related to context-free grammars.

\[ \tau = (N, S, W_1, W_2, R, p) \]

- \( N \): set of non-terminal symbols.
- \( S \in N \): the axiom.
- \( W_1 \): finite set of terminal symbols of language 1.
- \( W_2 \): finite set of terminal symbols of language 2.
- \( R \): finite set of rules of type:
  - lexical rules: \( A \rightarrow x/\epsilon, A \rightarrow \epsilon/y, A \rightarrow x/y \).
  - direct syntactic rules \( A \rightarrow [BC] \)
  - inverse syntactic rules \( A \rightarrow \langle BC \rangle \)
- \( p \): a function that determines the probability of each rule.

- Analyse two strings simultaneously.

May 2008

LREC08

G. Sanchis
SITG example

Source tree  Rule  Target tree

```
A  \rightarrow [BC]

A  \rightarrow \langle BC \rangle

A  \rightarrow \langle BC \rangle
```

May 2008  LREC08  G. Sanchis
SITGs for phrase extraction

- Analyse two strings simultaneously.
  - Can be used to extract segments.
  - Take into account syntax-motivated restrictions.

- Original algorithm for parsing a sentence by Wu similar to CYK, $O(|x|^3|y|^3|R|)$

- Sánchez and Benedí, 2006: $O(|x||y||R|)$ when $x$ and $y$ are fully bracketed.

- Algorithm for phrase extraction:
  - Initial SITG built heuristically from word alignments.
  - Reestimation of probabilities with bracketed corpus to obtain improved SITG.
  - Training corpus parsed with SITG in order to obtain bilingual segments.
  - Inverse and direct translation probabilities:

$$p_c(s|t) = \frac{N(s, t)}{N(t)}, \quad p_c(t|s) = \frac{N(s, t)}{N(s)}$$
Phrase extraction example

Direct translation rule: \( A \rightarrow [BC] \)

\[
\begin{array}{c}
A \\
\downarrow \\
B \quad C \\
\downarrow \\
x_i \quad x_l \quad x_j \\
\downarrow \\
y_k \quad y_K \\
\downarrow \\
y_l
\end{array}
\]

\[
\Rightarrow \left\{ \{x_{i+1} \ldots x_I, y_{k+1} \ldots y_K\}, \{x_{I+1} \ldots x_j, y_{K+1} \ldots y_l\} \right\}
\]

Inverse translation rule: \( A \rightarrow \langle BC \rangle \)

\[
\begin{array}{c}
A \\
\downarrow \\
B \quad C \\
\downarrow \\
x_i \quad x_l \quad x_j \\
\downarrow \\
y_k \quad y_K \\
\downarrow \\
y_l
\end{array}
\]

\[
\Rightarrow \left\{ \{x_{i+1} \ldots x_I, y_{K+1} \ldots y_l\}, \{x_{I+1} \ldots x_j, y_{k+1} \ldots y_K\} \right\}
\]
When obtaining $\hat{T}_{x,y}$, a subtree $\hat{T}_{s,t}$ is obtained as well for a specific $(s,t)$.

This defines a joint probability $\hat{p}(s, t)$.

Given that the corpus is bracketed, different $\hat{T}_{s,t}$ may be obtained.  
$\Rightarrow$ different $\hat{p}(s, t)$ may exist.

Let be $\Omega$ the multiset of spans obtained from a training sample.

Let be $\Omega_{s,t} \subseteq \Omega$ a multiset of $(s, t)$ spans.

\[ \Rightarrow E_{\Omega}(\hat{p}(s, t)) = \frac{\sum_{\omega \in \Omega_{s,t}} \hat{p}_\omega(s, t)}{|\Omega|} \]

$\Rightarrow p_s(s|t) = \frac{E_{\Omega}(\hat{p}(s, t))}{E_{\Omega}(\hat{p}(t))}$  and  $p_s(t|s) = \frac{E_{\Omega}(\hat{p}(s, t))}{E_{\Omega}(\hat{p}(s))}$.
## Experimental results

- **Corpus: Europarl**

<table>
<thead>
<tr>
<th>Training</th>
<th>Spanish</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences</td>
<td>730K</td>
<td></td>
</tr>
<tr>
<td>Different pairs</td>
<td>716K</td>
<td></td>
</tr>
<tr>
<td>Vocabulary size</td>
<td>103K</td>
<td>64K</td>
</tr>
<tr>
<td>Average length</td>
<td>21.5</td>
<td>20.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Development</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td>Average length</td>
<td>30.3</td>
<td>29.3</td>
</tr>
<tr>
<td>Out of vocabulary</td>
<td>208</td>
<td>127</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Devtest</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td>Average length</td>
<td>30.2</td>
<td>29.0</td>
</tr>
<tr>
<td>Out of vocabulary</td>
<td>207</td>
<td>125</td>
</tr>
</tbody>
</table>
Experimental results

- Translation results for a SITG with 1, 2, 3 and 4 non-terminal symbols.

- It. 0: Heuristically obtained SITG, only $p_c(\cdot|\cdot)$

- It. 1: One estimation iteration, $p_e(\cdot|\cdot)$

- + syntactic: adding $p_s(\cdot|\cdot)$

<table>
<thead>
<tr>
<th>non terms</th>
<th>It. 0</th>
<th>It. 1</th>
<th>+ syntactic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>WER</td>
<td>BLEU</td>
</tr>
<tr>
<td>1</td>
<td>26.8</td>
<td>62.5</td>
<td>26.9</td>
</tr>
<tr>
<td>4</td>
<td>26.6</td>
<td>63.2</td>
<td>27.9</td>
</tr>
</tbody>
</table>

May 2008 LREC08 G. Sanchis
Discussion

▶ Comparatively, best result reported so far with this technique was 23.0 BLEU.

▶ Best score obtained with Moses: 31.0 BLEU.

▶ with only direct and inverse models: 29.6 BLEU vs our 27.9 / 28.9.

⇒ Not directly comparable with Moses’ best score: we have no lexical models.
⇒ Will add lexical models in the future.
⇒ Traditional PB models cannot obtain syntactic scores!
⇒ Moses best score uses 19M segment pairs, we use half that amount.

▶ Adding non-terminal symbols seems to improve.
Conclusions and ongoing/future work

▶ Conclusions:
  ■ Alternative, competitive method for phrase extraction.
  ■ Importance of parsed corpora for estimating SITG.

▶ Future work:
  ■ Add lexical probabilities.
  ■ Combine SITG’s phrase table with Moses’ phrase table.
  ■ Research ways to exploit reordering information in SITGs.
Questions? Comments? Suggestions?