The UPV Translation System for IWSLT 2009

Authors: Guillem Gascó i Mora
         Joan Andreu Sánchez Peiró

1-2 December 2009
Index

○ 1 Introduction  ▶ 1
  2 Decoding  ▶ 3
  3 SITG Training  ▶ 7
  4 Experiments  ▶ 16
  5 Conclusions  ▶ 19
Motivation and Related Work

**Motivation**
- PBT systems tend to be weak on target language fluency.
- Long-range dependencies or reorderings cannot be controlled by the n-gram language models.
- Extending PBT systems with syntactic information is difficult.
- Syntactic MT systems usually solve such problems but have a low sentence coverage.
- Solution proposed: Fill the gap between both approaches presenting a hybrid system that uses Stochastic Inversion Transduction Grammars.

**Related Work**
- Maximum entropy models for BTG statistical MT [Xiong06] [Xiong08].
- Hierarchical MT systems [Chiang 05]
- Syntax Augmented MT [Venugopal 06]
Index

1 Introduction ▶ 1

2 Decoding ▶ 3

3 SITG Training ▶ 7

4 Experiments ▶ 16

5 Conclusions ▶ 19
Theoretical Framework: Stochastic Phrasal ITG

- Phrasal Inversion Transduction Grammars \((\mathcal{N}, \Sigma, \Delta, S, \mathcal{R})\):
  - Direct Syntactic rule: \(A \rightarrow [BC]\) where \(A, B, C \in \mathcal{N}\)
  - Inverse Syntactic rule: \(A \rightarrow \langle BC \rangle\) where \(A, B, C \in \mathcal{N}\)
  - Lexical rule: \(A \rightarrow x/y\) where \(x \in \Sigma^*\) and \(y \in \Delta^*\)

\[
\sum_{B,C \in \mathcal{N}} (\Pr(A \rightarrow [BC]) + \Pr(A \rightarrow \langle BC \rangle)) + \sum_{x \in \Sigma^*} \sum_{y \in \Delta^*} (\Pr(A \rightarrow x/y) = 1
\]

- SPhITG: Stochastic extension of PhITG.
Translation Model

Translation goal:

\[(t^*, D^*) = \arg\max_{t,D} \Pr(S \xrightarrow{D} s/t)\]

Log-linear model over the derivations:

\[\Pr(D) = \prod_i h_i(D)^{\lambda_i}\]

Models used: Usual models of PBT systems and the syntactic model (probability of the syntactic SITG rules used in \(D\)).
Decoding Algorithm

- Without n-gram language model: CKY-like chart decoding.
- Using the n-gram language model: A Hypotheses stack in each cell of the chart.
- Optimization strategies: Recombination of hypotheses, beam pruning and histogram pruning.

```
   verde bruja
   
   S N

   ADJ
   verde

   sno [ ADJ NN ]

   verde bruja
   bruja verde
   
   S N

   ADJ
   bruja

   sno [ ADJ NN ]

   bruja
```

```
Index

1 Introduction ▷ 1

2 Decoding ▷ 3

3 SITG Training ▷ 7

4 Experiments ▷ 16

5 Conclusions ▷ 19
SITG Training

- Initial ITG:
  - Lexical rules from alignments, all with the same non-terminal.
  - All the possible syntactical rules using 4 non-terminals (from NT1 to NT4) with a random probability.

\[
NT1 \rightarrow [NT1 \ NT1] \ ... \ NT4 \rightarrow \langle NT4 \ NT4 \rangle
\]

- Reestimation of the SITG using the Viterbi reestimation algorithm:
  - Get the most likely ITG parse trees.
  - Estimate probabilities by counting productions and normalizing.
SITG Training

- Association of linguistic meaning (from the input) to the SITG non-terminals.
this is a simple example. # esto es un ejemplo simple.
this is a simple example. # esto es un ejemplo simple.
this is a simple example. # esto es un ejemplo simple

((this (is (a (simple example))))) # esto es un ejemplo simple
this is a simple example. # esto es un ejemplo simple.

(((this(is(a(simple example)))))# esto es un ejemplo simple
this is a simple example. # esto es un ejemplo simple.

(((this(is(a(simple example)))))). # esto es un ejemplo simple
this is a simple example. # esto es un ejemplo simple.

(( this ( is ( a ( simple example ) ) ) ) .) # esto es un ejemplo simple
Index

1 Introduction ▶ 1
2 Decoding ▶ 3
3 SITG Training ▶ 7
  ○ 4 Experiments ▶ 16
5 Conclusions ▶ 19
Experimental Results

Translation experiments over IWSLT08 Chinese-English corpus:

<table>
<thead>
<tr>
<th></th>
<th>Chinese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>19.972</td>
<td></td>
</tr>
<tr>
<td>Words</td>
<td>171,591</td>
<td>188,960</td>
</tr>
<tr>
<td>Vocabulary Size</td>
<td>8,428</td>
<td>7,182</td>
</tr>
<tr>
<td><strong>DevSet</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>489</td>
<td></td>
</tr>
<tr>
<td>Words</td>
<td>3,169</td>
<td>3,861</td>
</tr>
<tr>
<td>OOV Words</td>
<td>111</td>
<td>115</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>507</td>
<td></td>
</tr>
<tr>
<td>Words</td>
<td>3,357</td>
<td>-</td>
</tr>
<tr>
<td>OOV Words</td>
<td>97</td>
<td>-</td>
</tr>
</tbody>
</table>

Baseline System: PBT system with the same phrase table.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>% BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>41.1</td>
</tr>
<tr>
<td>Initial SITG</td>
<td>41.23</td>
</tr>
<tr>
<td>Reestimated SITG</td>
<td>41.79</td>
</tr>
<tr>
<td>SAITG</td>
<td>42.85</td>
</tr>
</tbody>
</table>
## Experimental Results

<table>
<thead>
<tr>
<th></th>
<th>this one and what ’s the difference between ?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBT</td>
<td>what ’s the difference between this with that ?</td>
</tr>
<tr>
<td>SAITG</td>
<td>how is this one different from that one ?</td>
</tr>
<tr>
<td>Ref</td>
<td>call mr. is three four one four five six seven .</td>
</tr>
<tr>
<td>PBT</td>
<td>call mr. is three six four five seven four one .</td>
</tr>
<tr>
<td>SAITG</td>
<td>the number for s nicholas is three six four five seven four one .</td>
</tr>
<tr>
<td>Ref</td>
<td>can i go to the front row ?</td>
</tr>
<tr>
<td>PBT</td>
<td>is it okay to the front row ?</td>
</tr>
<tr>
<td>SAITG</td>
<td>can i go up to the front ?</td>
</tr>
</tbody>
</table>

Some rules obtained:

- \( \Pr(\text{QP} \rightarrow [\text{CD CD}]) = 0.147 \)
- \( \Pr(\text{QP} \rightarrow \langle \text{CD CD} \rangle) = 0.046 \)
- \( \Pr(\text{QP} \rightarrow [\text{CD QP}]) = 0.284 \)
- \( \Pr(\text{QP} \rightarrow \langle \text{QP CD} \rangle) = 0.061 \)
Index

1 Introduction ▶ 1
2 Decoding ▶ 3
3 SITG Training ▶ 7
4 Experiments ▶ 16
5 Conclusions ▶ 19
Conclusions

- SITG based decoder.
- Analyzed heuristics to train the SITG.
- Phrase table from a PBT system.
- When no syntactic information is used, it is almost equivalent to a PBT.
- The use of linguistic information improve the results.
The UPV Translation System for IWSLT 2009

Authors: Guillem Gascó i Mora
Joan Andreu Sánchez Peiró
CKY-like Algorithm

\[ \delta_{ij}(A) = \max_{t} \Pr(A \Rightarrow s_i^j / t) \]

For all \( A \in N \) and \( i, j \) such that \( 0 \leq i < j \leq |s|, j - i \geq 1 \),

\[ \delta_{ij}(A) = \max(\delta_{[] ij}(A), \delta_{\langle ij}(A), \max_{t} \Pr(A \rightarrow s_i^j / t)) \]  

(1)

where

\[ \delta_{[] ij}(A) = \begin{cases} \max_{B,C \in N} \Pr(A \rightarrow [BC])\delta_{I I}(B)\delta_{I J}(C) & \text{if } j - i > 1 \\ i < I \leq j \\ 0 & \text{otherwise} \end{cases} \]  

(2)

\[ \delta_{\langle ij}(A) = \begin{cases} \max_{B,C \in N} \Pr(A \rightarrow \langle BC \rangle)\delta_{I I}(B)\delta_{I J}(C) & \text{if } j - i > 1 \\ i < I \leq j \\ 0 & \text{otherwise} \end{cases} \]  

(3)
CKY-like Algorithm

\[ \tau_{ij}(A) = \arg\max_t (\Pr(A \Rightarrow^* s^j_i/t)) \]

\[ \tau_{ij}(A) = \begin{cases} 
  t & \text{if } \Pr(A \rightarrow s^j_i/t) \text{ is the maximum in (1)} \\
  \tau_{iI}(B)\tau_{Ij}(C) & \text{if } \Pr(A \rightarrow [BC])\delta_{iI}(B)\delta_{Ij}(C) \text{ is the maximum in (1)} \\
  \tau_{Ij}(C)\tau_{iI}(B) & \text{if } \Pr(A \rightarrow \langle BC\rangle)\delta_{iI}(B)\delta_{Ij}(C) \text{ is the maximum in (1)} 
\end{cases} \quad (4) \]