Analysing Soft Syntax Features and Heuristics for Hierarchical Phrase Based Machine Translation

David Vilar, Daniel Stein, Hermann Ney

IWSLT 2008, Honolulu, Hawaii
20. October 2008

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany
1 Introduction

▶ Hierarchical phrase-based models: Generalization of phrase-based-models
  ▶ Allow for “gaps” in the phrases
  ▶ Integration of reordering in the translation model
▶ Study the effect of extraction heuristics
▶ Extension with inclusion of (soft) syntactic features
Outline

1 Introduction
2 Hierarchical Phrases
3 Heuristic Features
4 Syntactical Features
5 Experimental Results
6 Conclusions
2 Hierarchical Phrases

- **Formalization as a synchronous CFG**

- **Rules of the form** $X \rightarrow \langle \gamma, \alpha, \sim \rangle$, where:
  - $X$ is a non-terminal
  - $\gamma$ and $\alpha$ are strings of terminals and non-terminals
  - $\sim$ is a one-to-one correspondence between the non-terminals of $\alpha$ and $\gamma$

- **Example:**

  $X \rightarrow \langle 中 X^0 那个 X^1, It's the X^1 in the X^0 \rangle$
  $X \rightarrow \langle 也 要 X^0 一些 X^1, like to X^0 some X^1 too \rangle$

- **Additionally: Glue rules**

  $S \rightarrow \langle S^0 X^1, S^0 X^1 \rangle$
  $S \rightarrow \langle X^0, X^0 \rangle$
You order a toddler meal.

Illustration

meal • • • ■ • • •
toddler • • • • • ■ ■
a • • ■ • • • • •
order • ■ • • • • •
you ■ • • • • • •
did ■ • • • • • •
ha
ordinato
un
piatto
per
bambini

Alignment
Illustration

meal • • • • [square] • •
toddler • • • • [squares] • •
a • • [square] • • • •
order [square] • • • • •
you [squares] • • • • •
did [squares] • • • • •
ha • •
ordinato • •
piatto • •
per • •
bambini • •

Standard phrases
3 Heuristic Features

- Following features were tested:
  - **Paste rule**  Binary feature for rules of the form
    
    \[ X \rightarrow \langle X^0 \alpha, X^0 \beta \rangle \text{ or } X \rightarrow \langle \alpha X^0, \beta X^0 \rangle \]

  - **Hierarchical penalty**  Binary feature for hierarchical rules
  - **Number of non-terminals**  Two binary features indicating if the rule has one or two non-terminals.
  - **Extended glue rule**  added rule of the form
    
    \[ X \rightarrow \langle X^0 X^1, X^0 X^1 \rangle \]
4 Syntactical Features

- Goal: include linguistic information from a deep syntactic parser
- Idea: introduce additional soft syntactic features
- This can be done during the extraction of the phrases
  - No additional computational costs during decoding
  - Can be done both on source and target side
  - Rules are not filtered out
“Valid” syntactical phrases

► A phrase is valid when a node exists that completely covers all positions

► In order to obtain a normalized score, we add up all the counts and divide by the number of occurrences of the phrase pair

Extracted rule: $X \sim 0$ 在 哪 里 # Where is $X \sim 0$
Scoring variants

\[ m(i, j) = \text{minimum number of words to be deleted or added to a phrase, so that it fits the yield of a node} \]

Source Phrases:
- public toilet
- is the
Scoring variants

\[ m(i, j) = \text{minimum number of words to be deleted or added to a phrase,} \]

\[ \text{so that it fits the yield of a node} \]

Source Phrases:

- **public toilet** \( m(i, j) = 1 \)
- **is the**
Scoring variants

\[ m(i, j) = \text{minimum number of words to be deleted or added to a phrase, so that it fits the yield of a node} \]

Source Phrases:

- **public toilet** \( m(i, j) = 1 \)
- **is the** \( m(i, j) = 1 \)
Four count (“smoothing”) variants:

\[
c(i, j | t) := \begin{cases} 
\delta (m(i, j), 0) & \text{binary} \\
\frac{1}{m(i, j) + 1} & \text{linear} \\
\frac{1}{\exp (m(i, j))} & \text{exponential} \\
\frac{j - i}{(j - i) + m(i, j)} & \text{relative}
\end{cases}
\]
## 5 Experimental Results

**IWSLT BTEC Data (Tourist and Travel domain)**

<table>
<thead>
<tr>
<th></th>
<th>Chinese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>23 940</td>
<td></td>
</tr>
<tr>
<td>Running words</td>
<td>181 486</td>
<td>232 746</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>9 041</td>
<td>10 350</td>
</tr>
<tr>
<td><strong>Test 2004 Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>Running words</td>
<td>7 543</td>
<td>10 718</td>
</tr>
<tr>
<td>OOVs</td>
<td>96</td>
<td>154</td>
</tr>
<tr>
<td><strong>Test 2005 Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>506</td>
<td></td>
</tr>
<tr>
<td>Running words</td>
<td>8 052</td>
<td>10 828</td>
</tr>
<tr>
<td>OOVs</td>
<td>101</td>
<td>164</td>
</tr>
<tr>
<td><strong>Test 2008 Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>507</td>
<td></td>
</tr>
<tr>
<td>Running words</td>
<td>6325</td>
<td></td>
</tr>
<tr>
<td>OOVs</td>
<td>87</td>
<td></td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th></th>
<th>test04</th>
<th></th>
<th>test05</th>
<th></th>
<th>test08</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>TER</td>
<td>BLEU</td>
<td>TER</td>
<td>BLEU</td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>47.3</td>
<td>42.6</td>
<td>50.9</td>
<td>37.6</td>
<td>39.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-syntactic information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hierarch</td>
<td>48.4</td>
<td>41.9</td>
<td>51.4</td>
<td>38.1</td>
<td>39.6</td>
<td>40.8</td>
</tr>
<tr>
<td>paste</td>
<td>49.1</td>
<td>41.6</td>
<td>51.1</td>
<td>38.0</td>
<td></td>
<td>39.7</td>
</tr>
<tr>
<td>glue2</td>
<td>48.2</td>
<td>41.8</td>
<td>51.2</td>
<td>37.6</td>
<td></td>
<td>39.8</td>
</tr>
<tr>
<td>1NT2NT</td>
<td>48.4</td>
<td>42.2</td>
<td>51.8</td>
<td>37.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>syntactic information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>binary</td>
<td>47.8</td>
<td>41.7</td>
<td>51.7</td>
<td>37.5</td>
<td>40.3</td>
<td></td>
</tr>
<tr>
<td>linear</td>
<td>47.6</td>
<td>41.9</td>
<td>51.2</td>
<td>37.6</td>
<td>40.6</td>
<td></td>
</tr>
<tr>
<td>exponential</td>
<td>47.9</td>
<td>41.7</td>
<td>51.6</td>
<td>37.4</td>
<td>40.3</td>
<td></td>
</tr>
<tr>
<td>relative</td>
<td>47.3</td>
<td>42.4</td>
<td>51.5</td>
<td>37.3</td>
<td>40.2</td>
<td></td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th></th>
<th>test04 BLEU</th>
<th>test04 TER</th>
<th>test05 BLEU</th>
<th>test05 TER</th>
<th>test08 BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>47.3</td>
<td>42.6</td>
<td>50.9</td>
<td>37.6</td>
<td>39.6</td>
</tr>
<tr>
<td>non-syntactic information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hierarch + paste</td>
<td>48.5</td>
<td>42.0</td>
<td>51.9</td>
<td>37.6</td>
<td>39.6</td>
</tr>
<tr>
<td>hierarch + paste + glue2</td>
<td>49.2</td>
<td>42.5</td>
<td>50.8</td>
<td>37.5</td>
<td>39.5</td>
</tr>
<tr>
<td>hierarch + paste + glue2 + 1NT2NT</td>
<td>48.6</td>
<td>41.6</td>
<td>51.0</td>
<td>37.9</td>
<td>40.0</td>
</tr>
<tr>
<td>combination of both syntactic and non-syntactic information (all features)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>binary</td>
<td>46.9</td>
<td>42.5</td>
<td>50.6</td>
<td>38.4</td>
<td>39.9</td>
</tr>
<tr>
<td>linear</td>
<td>48.0</td>
<td>42.3</td>
<td>51.2</td>
<td>38.0</td>
<td>40.5</td>
</tr>
<tr>
<td>exponential</td>
<td>47.7</td>
<td>42.3</td>
<td>51.0</td>
<td>38.4</td>
<td>40.0</td>
</tr>
<tr>
<td>relative</td>
<td>47.8</td>
<td>42.3</td>
<td>51.0</td>
<td>38.0</td>
<td>40.3</td>
</tr>
</tbody>
</table>
## Example Translations

<table>
<thead>
<tr>
<th>Reference</th>
<th>Baseline</th>
<th>Syntactical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where is the exchange counter?</td>
<td>The currency exchange office is</td>
<td>Where is the currency exchange office?</td>
</tr>
<tr>
<td>Could you exchange it for a new one?</td>
<td>You can buy a new one?</td>
<td>Could you change it for a new one?</td>
</tr>
<tr>
<td>You can take our airport shuttle bus to pick up the car.</td>
<td>You can take our airport shuttle bus with me.</td>
<td>You can take our the airport shuttle bus come to pick it up.</td>
</tr>
</tbody>
</table>
6 Conclusions

- Analyzed heuristics for phrase extraction
- Introduced soft syntactic constraints
  - Use of source- and target-side information
  - No additional search effort
- High variability of results
  - Test on bigger corpora
- Bigger improvements when dealing with speech input (system talk tomorrow!)
- Applicable also to phrase-based systems
Analysing Soft Syntax Features and Heuristics for Hierarchical Phrase Based Machine Translation

David Vilar, Daniel Stein, Hermann Ney

IWSLT 2008, Honolulu, Hawaii
20. October 2008

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany