Czech-to-English Translation:

MT Marathon 2009
Session Preview

Jonathan Clark
Greg Hanneman

Language Technologies Institute
Carnegie Mellon University
26 January 2009
Outline

• Stat-XFER processing pipeline
• Processed Czech–English resources
• Possible workshop tasks
  – Syntactic phrase table combination methods
  – Synchronous grammar development
    • Selection of grammar rules
    • Exploration of label granularity
    • Development of manual grammars
  – Integration of morphological analysis
Stat-XFER Data Processing

- Czech Corpus
- English Corpus

Parsing

Word Alignment

- Moses Phrase Extraction
- Syntactic Phrase Extraction

Grammar Extraction

- Moses Phrase Table
- Syntactic Phrase Table
- SCFG Grammar Rules
Stat-XFER Data Processing

• Corpus:
  – Project Syndicate news data: portion of CzEng corpus (84,141 sentences)
Stat-XFER Data Processing

- Parsing:
  - Czech dependency parses by TectoMT; converted to projective c-structure
  - English c-structure parses by Stanford parser
Stat-XFER Data Processing

- Word alignment:
  - GIZA++ grow-diag-final alignment done in advance on tokenized corpus
  - Alignments computed on full CzEng corpus of 8 million sentences
Stat-XFER Data Processing

- Phrase extraction:
  - Syntactic extraction by PFA node alignment algorithm, t2ts mode
  - Non-syntactic extraction with Moses package
Stat-XFER Data Processing

- Grammar extraction:
  - Using syntactic node alignments as tree decomposition points
Final Result

• Two phrase tables, with counts:

<table>
<thead>
<tr>
<th></th>
<th>NNS</th>
<th>NNS</th>
<th>rozumem</th>
<th>brains</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>NN</td>
<td>NN</td>
<td>rozumem</td>
<td>reason</td>
</tr>
<tr>
<td>4</td>
<td>NN</td>
<td>NN</td>
<td>rozumem</td>
<td>sense</td>
</tr>
<tr>
<td>1</td>
<td>NP</td>
<td>NP</td>
<td>rozumem</td>
<td>reason</td>
</tr>
<tr>
<td>1</td>
<td>NN</td>
<td>NN</td>
<td>rozumností</td>
<td>wisdom</td>
</tr>
<tr>
<td>1</td>
<td>JJ</td>
<td>JJ</td>
<td>rozumnou</td>
<td>sensible</td>
</tr>
<tr>
<td>1</td>
<td>ADJP</td>
<td>ADJP</td>
<td>rozumnou měrou jisté</td>
<td>reasonably certain</td>
</tr>
<tr>
<td>1</td>
<td>NP</td>
<td>NP</td>
<td>rozumnou politiku</td>
<td>sensible policy</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>PHR</th>
<th>PHR</th>
<th>rozumem</th>
<th>brains</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>PHR</td>
<td>PHR</td>
<td>rozumem</td>
<td>reason</td>
</tr>
<tr>
<td>4</td>
<td>PHR</td>
<td>PHR</td>
<td>rozumem</td>
<td>sense</td>
</tr>
<tr>
<td>2</td>
<td>PHR</td>
<td>PHR</td>
<td>rozumem</td>
<td>sense .</td>
</tr>
<tr>
<td>1</td>
<td>PHR</td>
<td>PHR</td>
<td>rozumem , a že</td>
<td>brains ; and that</td>
</tr>
<tr>
<td>1</td>
<td>PHR</td>
<td>PHR</td>
<td>rozumem , pokud</td>
<td>sense if</td>
</tr>
<tr>
<td>1</td>
<td>PHR</td>
<td>PHR</td>
<td>rozumem , pokud ne</td>
<td>sense if not</td>
</tr>
</tbody>
</table>
Final Result

• Three suffix-array language models
  – Target side of Project Syndicate corpus
  – … + more monolingual English data
  – … + target side of public CzEng corpus

• WMT tuning, development, and test sets

• = Baseline Stat-XFER system ready to analyse and expand
Outline

• Stat-XFER processing pipeline
• Processed Czech–English resources
• Possible workshop tasks
  – Syntactic phrase table combination methods
  – Synchronous grammar development
    • Selection of grammar rules
    • Exploration of label granularity
    • Development of manual grammars
  – Integration of morphological analysis
Phrase Table Combination

• Combination of non-syntactic and syntactic phrase pairs
  – Direct combination and syntax prioritization
Synchronous Grammars: Rule Selection

- Rule learning yields huge grammars
- Decoding with millions of abstract rules is intractable
- Open Question: How do we select the best grammar rules with regard to translation quality and decoding speed?
Synchronous Grammars: Label Granularity

• Rule learning assigns non-terminal and POS labels from input parse trees

• Input labels are believed appropriate…
  – For a given single language
  – According to a particular theory of grammar

• Open Question: How do we expand or collapse these labels so that they are appropriate for translating a particular language pair?
Synchronous Grammars: Czech Example

• Subject moves in English translation
• Verbs in past tense cannot be associated with modifiers in present tense

Proti odmítnutí se zitra Petr
against dismissal AUX-REFL tomorrow Peter

v práci rozhodl protestovat
of work decided to protest

“Peter decided to protest against the dismissal of work tomorrow.”

* Example from Bojar and Lopez, “Tree-based Translation,” MT Marathon Presentation 2008
Synchronous Grammars: Manual Grammar Writing

- Stat-XFER supports LFG-style unification
- Feature structures for unification can also be provided by the morphology server

\[
\text{VP::VP : [ADV VP] -> [VP ADV]}
\]

\[
\begin{align*}
(X1::Y2) \\
(X2::Y1) \\
(*tgsrule* 0.2) \\
(*sgtrule* 0.6) \\
((X0 \text{ tense}) = (X1 \text{ tense})) \\
((X0 \text{ tense}) = (X2 \text{ tense}))
\end{align*}
\]
Czech Morphology: Example

- Czech words include clitics and inflectional morphology, marking meanings such as gender and number

nerožumím
ne+rozum+ím
NEG+understand+1SG
“I do not understand”
Czech Morphology in Stat-XFER

- Stat-XFER allows external morphology server to segment and annotate words at runtime
- Ambiguous word segmentations can be encoded as a lattice
- Must segment all training data, then rebuild phrase table & language model
(Your Idea Here)

• Any ideas about applying the statistical transfer framework to Czech–English translation are welcome!