AX

Lexico-syntactic Information from Corpora
Josefův Důl, 21.1.2003

Ondřej Bojar
Motivation

- Distant goals:
  - Machine translation, grammar checking, text summary, ...

- Steps in processing sentences:
  - Morphological analysis (~ done)
  - Syntactic analysis (difficult, lexico-syntactic information needed)
  - Deeper analyses (more difficult, out of scope of this presentation)
Information for Verbs

- Verbs “organize” the sentence, let’s start with them.
- Verbs “require” specific (types of) complements (modifications).
- Observed frame = list of sons in a dependency tree.
- Surface frame = list of “typical” complements of (a type of) a verb. (e.g. protestovat proti Čemu, not protestovat kdy)
- Steps in automatic extraction of surface frames:
  - Obtain observed frames.
  - Filter the complements to obtain surface frames.
  - (Further inference to obtain valency frames.)

Surface ~ Subcategorization frame, Subcategorization ≠ Valency frame
Data Available

- Prague Dependency Treebank (PDT)
  - Syntactic annotation (manual).
  - ~1.5 million words, ~98 thousand sentences.
  - Observed frames “free of charge”.

- Czech National Corpus (CNC)
  - Morphological annotation (automatically disambiguated).
  - ~100 million words, ~1.8 million sentences.
  - Observed frames are hard to extract

(Sample input sentence: Nové předpisy vyžadují mj. též povinnost určit u každého kvantitativního výsledku měření jeho nejistotu.)
## Data for Individual Verbs

<table>
<thead>
<tr>
<th>Verbs</th>
<th>Observations in CNC (min, $\emptyset$, max)</th>
<th>PDT Observations</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 11 253 207</td>
<td>1 737</td>
<td>být</td>
</tr>
<tr>
<td>2</td>
<td>1 2 175 254</td>
<td>513</td>
<td>mít</td>
</tr>
<tr>
<td>3</td>
<td>2 570 234, 851 099, 5, 1 131 965</td>
<td>116, 159, 5, 203</td>
<td>moci, muset</td>
</tr>
<tr>
<td>4</td>
<td>21 140 362, 243 575, 3, 522 307</td>
<td>21, 106, 3, 524</td>
<td>říci, chtít, jít, dát, uvést</td>
</tr>
<tr>
<td>5</td>
<td>53 68 535, 92 773, 1, 126 411</td>
<td>2, 45, 2, 133</td>
<td>čekat, zůstat, znamenat</td>
</tr>
<tr>
<td>6</td>
<td>99 40 248, 50 606, 4, 68 004</td>
<td>1, 31, 8, 99</td>
<td>představit, věnovat, vyjít</td>
</tr>
<tr>
<td>7</td>
<td>164 23 011, 30 707, 7, 40 210</td>
<td>1, 18, 7, 79</td>
<td>číst, přicházet, končit</td>
</tr>
<tr>
<td>8</td>
<td>317 10 970, 15 861, 7, 22 982</td>
<td>1, 10, 5, 83</td>
<td>dokončit, svědčit, přejít</td>
</tr>
<tr>
<td>9</td>
<td>818 3 551, 6 137, 0, 10 966</td>
<td>1, 4, 5, 29</td>
<td>uznávat, reprezentovat</td>
</tr>
<tr>
<td>10</td>
<td>20 800 0, 241, 7, 3 548</td>
<td>0, 1, 8, 79</td>
<td>ztroskotat, rýsovat</td>
</tr>
</tbody>
</table>
More Syntactically Annotated Data

• Simple scheme:
  • (Get more texts, e.g. from the Internet.)
  • Morphologically anotate and disambiguate the sentences.
  • Employ one of the parsers available for Czech to get the dependency trees.
  • Extract the desired lexico-syntactic information.
• But current parsers are not accurate enough (we shall see).
**Better Annotated Data**

- Parsers are not accurate enough, so:
  - Morphologically anotate (+disambiguate) the sentences.
  - Choose “simple” sentences, easier to parse.
  - Employ one of the parsers.
- The filtration must be *linguistically motivated*.
  - Keep sentences containing the observed phenomenon.
  - Filter out all the sentences where the phenomenon is “hidden” and/or interferes with other phenomena.
  - (Partial parsing needed for some of the decisions.)
- Goal dependent filtration $\Rightarrow$ a “scripting” language would help creating different filters $\Rightarrow$ system AX.
AX

- Goals:
  - Easy formulation of filters to reject sentences.
  - Easy implementation of partial syntactic analysis of input sentences.
- Overall scheme:
  - Input sentence = list of feature structures.
    1 input word = 1 fs of (ambiguous) morphological information.
  - The script is a pipeline of filters and rewriting rules.
  - Blocks of rules update the lists of feature structures, generate different “readings” of the sentence.
    Not limited to CFG or CSG. Nondeterministic rules allowed.
  - Filters reject unacceptable lists of feature structures.
Sample rules and filters

# Find a sublist of fss matching given RE
# And replace with output fs or fss or copy \1, \2...
rule "Simple noun phrase:" nounph --> (adv{0,2} adj)* noun ::
    adj = [cat-adj], noun = [cat-noun], adv = [cat-adv],
    adj.case = noun.case, # check constrains on rule
    adj.num = noun.num,  # check constrains on rule
    adj.gend = noun.gend, # check constrains on rule
    nounph = noun,       # fill the output fs
end

filter "Reject readings containing two or more verbs
or a conjunction:"
    | ^.* [cat-conj] .* $
end
Filtration to Observe Verb Frames

- Reject complicated punctuation and numbers (colon...).
- Combine analytical verb forms (*usnuli jsme*).
- Ignore sentences with more autosematic verbs.
- Fold simple coordination, reject complicated.
- Accept only sentences with simple structure:
  H or H1=H2 or H
  \[ \text{VV} \quad \text{VV} \]
- (Optional) Reject sentences with “suspicious” word order patterns (WOP).

⇒ “very simple sentences, vss.” (~15–20% of sents. in corpora)
## Parsers on vss

<table>
<thead>
<tr>
<th>Correct Dependencies</th>
<th>Words</th>
<th>Statistical</th>
<th>Hand-made</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Collins</td>
<td>Zeman</td>
</tr>
<tr>
<td>All Sentences</td>
<td>126 030</td>
<td>82,51 %</td>
<td>69,15 %</td>
</tr>
<tr>
<td>Very simple sentences</td>
<td>20 028</td>
<td>87,70 %</td>
<td>79,40 %</td>
</tr>
<tr>
<td>... and no suspicious WOP</td>
<td>11 030</td>
<td>87,89 %</td>
<td>79,31 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correct Sentences</th>
<th>Sentences</th>
<th>Statistical</th>
<th>Hand-made</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Collins</td>
<td>Zeman</td>
</tr>
<tr>
<td>All Sentences</td>
<td>7 319</td>
<td>30,95 %</td>
<td>15,00 %</td>
</tr>
<tr>
<td>Very simple sentences</td>
<td>1 786</td>
<td>47,14 %</td>
<td>29,00 %</td>
</tr>
<tr>
<td>... and no suspicious WOP</td>
<td>1 113</td>
<td>52,83 %</td>
<td>34,41 %</td>
</tr>
</tbody>
</table>

- 5–10 % better measured by correct dependencies.
- 15–20 % better (≈twice) measured by sentences without a mistake.
**Parsers on vss (contd.)**

<table>
<thead>
<tr>
<th>Observed Frames Correct</th>
<th>Verbs</th>
<th>Statistical</th>
<th>Hand-made</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Collins</td>
<td>Zeman</td>
</tr>
<tr>
<td><strong>All Sentences</strong></td>
<td>16 329</td>
<td>55,32 %</td>
<td>33,11 %</td>
</tr>
<tr>
<td><strong>Very simple sentences</strong></td>
<td>2 472</td>
<td>61,37 %</td>
<td>41,87 %</td>
</tr>
<tr>
<td><strong>. . . and no suspicious WOP</strong></td>
<td>1 546</td>
<td>64,68 %</td>
<td>47,02 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No Sons Missing</th>
<th>Verbs</th>
<th>Statistical</th>
<th>Hand-made</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Collins</td>
<td>Zeman</td>
</tr>
<tr>
<td><strong>All Sentences</strong></td>
<td>16 329</td>
<td>73,73 %</td>
<td>55,86 %</td>
</tr>
<tr>
<td><strong>Very simple sentences</strong></td>
<td>2 472</td>
<td>78,52 %</td>
<td>67,76 %</td>
</tr>
<tr>
<td><strong>. . . and no suspicious WOP</strong></td>
<td>1 546</td>
<td>81,44 %</td>
<td>71,28 %</td>
</tr>
</tbody>
</table>

- 10–15% better measured by correct observed frames.
Summary

• We need syntactic lexicons.
• Treebanks are excellent source, but:
  • expensive to build.
  • small (i.e. many lexemes are rare).
• Use “non-tree” corpora (bigger):
  • + parsers available to obtain syntactic annotation.
  • but select sentences containing an example of the observed phenomenon easier to analyze.
• AX is the system to filter sentences.
  • In fact powerful enough to build partial or full parsers.