Automated Extraction of Lexico-Syntactic Information

Ondřej Bojar
obo@cuni.cz

June 17, 2004

Outline

- Motivation: Why syntactic lexicons?
- The two goals:
  - Extending monolingual syntactic lexicons.
  - Providing translation dictionaries with syntactic information.
- Summary and further research.

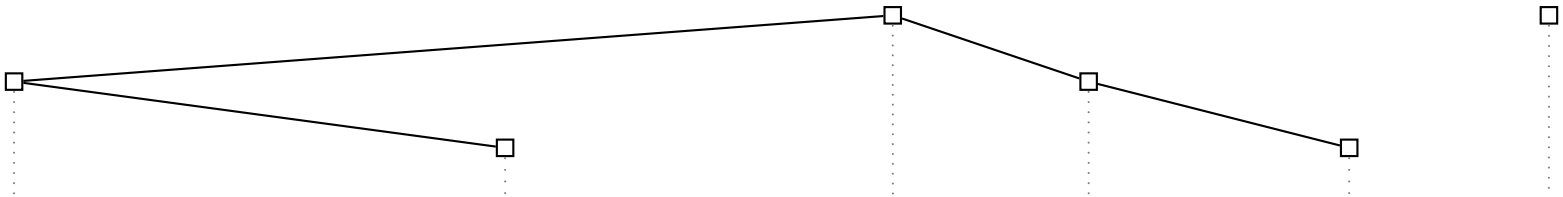
Syntactic Analysis: A Tough Cookie

- Task: String of words $\rightarrow$ tree.
- There are $O(n^n)$ configurations theoretically possible.
  - Ribarov[1]: sentences of 20 words have 300 structures in PDT, not 15.6 billion structures.
  - Bojar [2004]: Allow observed local configurations and you'll get 9^n possible solutions. ($9^{20} \approx 10^{19}$)
- Statistical approaches simply cheat: they provide most common (frequent) analyses and therefore commonly seem to perform well.

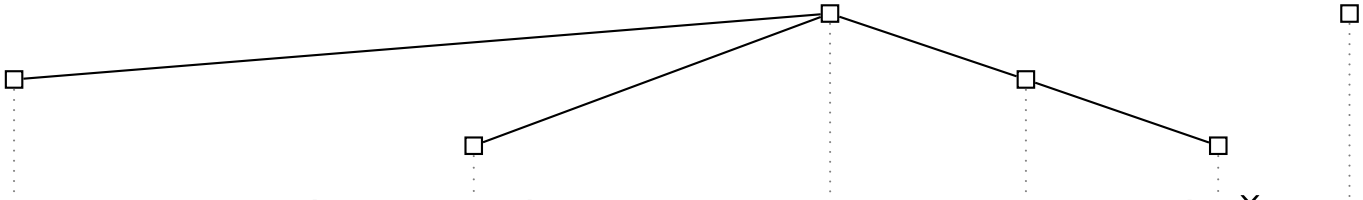
[1] Thesis in preparation

What a Lexicon Might Tell

(1)



(2)



The managers started to like the hypermarket.

The Two Goals

- Extending monolingual syntactic lexicons.
  - Current lexicons are incomplete.
  - Building lexicons is a demanding task.
  - Eventually, human decisions are necessary.

  $\Rightarrow$ An automatic preprocessing of corpus examples will help.

- Providing translation dictionaries with syntactic information.
  - No formalized syntactic information in current dictionaries.

  $\Rightarrow$ Several semi-automatic steps proposed.

Extending Monolingual Syntactic Lexicons (I)

- Treebank data are not sufficient:

In PDT, after having observed	20,000	75,000	training sentences
a new lemma (i.e. word) comes every	1.6	1.8	test sentences
a new full morphological tag comes every	110	290	test sentences
a new simplified tag[2] comes every	280	870	test sentences

  - PDT covers only 5,407 of 22,276 Czech verbs observed in Czech National Corpus.
  - PDT contains more than 50 occurences per verb only for a few hundreds of verbs.

[2] The simplified tag comprises only POS, SUBPOS, CASE, NUMBER and GENDER information.

Extending Monolingual Syntactic Lexicons (II)

- Simple scheme:
  - (Get more texts, e.g. from the Czech National Corpus or the Internet.)
  - Morphologically annotate 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: (55% verb frames observed correctly)

  $\Rightarrow$ Select "nice examples" only.
  - Keep sentences containing the observed phenomenon.
  - Filter out all the sentences where the phenomenon is "hidden" and/or interferes with other phenomena.

Extending Monolingual Syntactic Lexicons (III)

- Bojar [2002] designs a scripting language to select nice examples.
- A sample script to select nice examples of verbal frames helps current parsers:
  - 5–10% improvement in correct dependencies.
    (Reached 88% dependencies correct)
  - 10–15% improvement in correctly observed verb frames.
    (Reached 65% frames correct.)
- Different scripts should be used when extracting different kinds of data.

Providing Translation Dictionaries with Syntactic Information

Accessible Czech-English dictionaries are for humans only.

- Necessary steps:
  - Add morphological information.
  - Add syntactic structure and agreement constraints.
  - Provide entries with examples.
  - Estimate monolingual and parallel frequencies.
- Benefits:
  - Better machine translation, better monolingual syntactic lexicons.
  - Possibility to align deep syntactic lexicons.

Steps Proposed

- Manual disambiguation necessary. (We shall see.)
- Automatic morphological analysis and grouping saves a lot of work.
- Corpus data as an additional source:

Data Annotation Level		Possible Tasks to Augment Entries
Entries	**Corpora**	
no annotation	any annotation	no useful task
morphology	morphology	possibly annotate internal agreement
morphology	trees	find syntactic structure
trees	morphology	find examples, estimate frequency
trees	trees	confirm structure

Morphological Disambiguation: Manually!

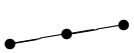
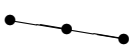
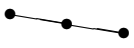
Noun and Noun/Adjective	Correct Interpretation	English Translation
husa divoká	Noun Adjective	grey goose
kniha účetní†	Noun Adjective	account book
napětí dovolené†	Noun Adjective	permissible stress
chyba měření	Noun Noun	measurement error
plán prací†	Noun Noun	schedule of operation
rozsah měření	Noun Noun	range of measurement
Numeral/Verb and Noun	**Correct Interpretation**	**English Translation**
tři prdele†	Numeral Noun	shitloads
pět švestek	Numeral Noun	one's duds*
pět chválu	Verb Noun	sing someone's praises

† These expressions allow for another interpretation, too, mostly kind of funny.

* Part of the idiom *pick up one's duds*.

Adding Syntactic Information

- Manual annotation plausible and preferable for important groups.

Noun Adjective Noun	Syntactic Structure	English Translation
Komise Evropské unie		Commission of the European Community
náhrada způsobené škody		dilapidation
látkou potažené sedadlo		fabric-covered seat
poruchy způsobené přijímačem		set noise
nevolnost způsobená pohybem		kinesia

- Once the syntactic information is present:
  - Adding agreement constraints automatically, per structure.
  - Searching for examples more precise.
  - Searching allows to find more examples (modified multi-word entries).

Searching for Examples, Estimating Frequencies

- In treebanks:
  - Finding a subtree in a forest is easy.
  - Only most prominent examples expected in PDT.
  - Possibility to search in automatic trees of "nice sentences".

- In plain corpora:
  - Use agreement constraints to reject random collocations.
  - Use syntactic structure to allow valid reordering and extra modifications present at specific places.
  - Relevant source to estimate frequencies.

Aligning Deep Syntactic Lexicons

- Czech and English deep syntactic lexicons under development. (VALLEX and FrameNet).
- Annotation schemata not equivalent but comparable.
- No common parallel corpus annotated with VALLEX on the Czech side and FrameNet on the English side.

$\Rightarrow$ Use surface syntactic translation dictionary as a bridging link.

Conclusion and Further Research

- We need syntactic lexicons.
- Automatic preprocessing saves effort (but does not solve the problem) for the two tasks:
  - Extending monolingual syntactic lexicons.
  - Providing translation dictionaries with syntactic information.
- Further goals:
  - Build a syntactic Czech-English dictionary.
  - Evaluate the utility of syntactic dictionaries.

References

Ondřej Bojar. Automatická extrakce lexikálně-syntaktických údajů z korpusu (Automatic extraction of lexico-syntactic information from corpora). Master's thesis, ÚFAL, MFF UK, Prague, Czech Republic, 2002. In Czech.

Ondřej Bojar. Czech Syntactic Analysis Constraint-Based, XDG: One Possible Start. *Prague Bulletin of Mathematical Linguistics*, 81:43–54, 2004. ISSN 0032-6585.