Parsing of Czech: Between Rules and Statistics

Miloš Jakubíček

NLP Centre
Faculty of Informatics, Masaryk University
jak@fi.muni.cz

IA161 Advanced Techniques of Natural Language Processing

Outline

1 What & Why

2 Formalisms

3 How?

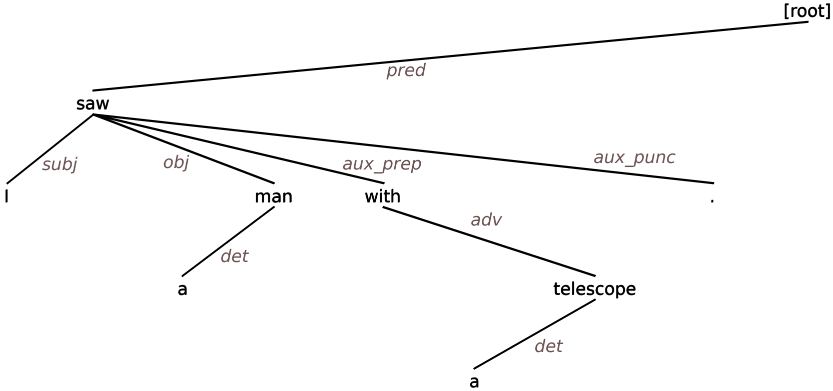
Natural Language Parsing

- What?
  - recovering surface structure of a sentence
  - a base point for further language analysis
- Why?
  - any advanced language processing
  - e.g. relations between words, phrase extraction, . . .

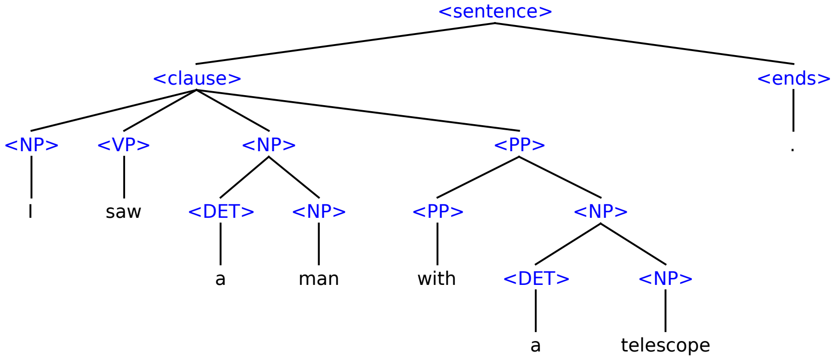
Formalisms

- Dependency syntax
- Phrase-structure (constituent) syntax
- Partial analysis/chunking
- . . . ...very many advanced formalisms

Example of a dependency tree



Example of a phrase-structure tree



How to do automatic parsing

- Key issues
  - ambiguity: sometimes a problem even for humans
  - „Karel mluvil o sexu s Britney Spears."
  - „I saw a man with a telescope."
  - low agreement
  - very hard to evaluate
  - ill-defined task?
- How to analyze syntax
  - rule-based systems
  - statistical based systems (induced grammars, machine learning methods)

Rules vs. statistics – which is better

- . . . it depends
- NLP field has mostly focused on statistical systems
  - very tempting from computer science point of view
  - outperforming rule-based systems on "standardized datasets"
  - . . . but: problems with overfitting, low flexibility of output
- goes back to: what is the task?

Examples of systems for Czech

- Synt
    - constituent system with dependency graph output
    - statistics: trained grammar rule probabilities
    - rules: grammar, rule levels
- SET
    - rules: grammar
- Czech word sketches
    - rules: sketch grammar (regexps over PoS tags)
    - statistics: association measures

Challenges

– in the interplay of rules and statistics:

- which syntactic phaenomena should be handled by rules/stats?
- which formalisms should be used for rules and stats?
- how to make the combination well-engineered?

⇒ parsing as a task-driven process

analytics of particular sentences
vs.
mining of general language knowledge