Syntaktická analýza

Dana Hlaváčková

Základní pojmy

- syntaktická analýza – parsing
- syntaktický analyzátor – parser
- věta – sentence S, clause
- nominální fráze – NP
- verbální fráze – VP
- závislostní stromy (větve, listy) – tree
- závislostní korpus, stromová banka – treebank

Syntaktická analýza

- předpoklad
  - tokenizace – tokeny (listy)
  - morfologicky označkovaný (a desambiguovaný) korpus
  - rozpoznání hranice věty – segmenter (statistický, pravidlový)
    - *...nechutnalo nám.*
    - *...Masarykovo nám. č. 13.*

Syntaktická analýza

- formální gramatika
- syntaktický analyzátor
- datové struktury – závislostní/složkové stromy
- treebank (závislostní korpus, stromová banka)
- automatická, poloautomatická, ruční anotace

K čemu to potřebujeme?

- další rovina popisu jazyka v NLP
- treebank – referenční data pro automatické nástroje
- synchronní (i diachronní) studie, vazba na slovesnou valenci a sémantickou rovinu
- frekvenční studie
- strojový překlad
- navazující aplikace, např. vývoj korektoru interpunkce, aktuální členění věty, koreferenční vztahy, dialogové systémy
- čeština – jeden z nejobtížnějších jazyků – flexe a volný slovosled

Syntaktická analýza – Praha

- historické pozadí
- lingvistický strukturalismus, Pražská škola
- Pražský lingvistický kroužek (1926, Mathesius, Jakobson, Trnka)
- funkčně generativní popis (FGD, Sgall, 60. léta)
  - závislostní syntax
  - hloubková (tektogramatická) struktura
  - formální popis aktuálního členění věty

Syntaktická analýza – Praha

- ÚFAL MFF UK http://ufal.mff.cuni.cz/pdt2.0
- PDT 2.0 (*Prague Dependency Treebank, Pražský zavislostní korpus*)
- ruční anotace
- rovina anotace:
  - slovní
  - morfologická (2 mil. slovních jednotek)
  - syntaktická (analytická; 1,5 mil. slovních jednotek)
  - sémantická (tektogramatická; 0,8 mil. slovních jednotek)
  - aktuální členění věty, koreferenční vztahy
- teoreticky závislý, určen pro strojové učení

Syntaktická analýza – Brno

- FI MU – CZPJ
- syntaktický analyzátor klara, synt, set
- morfologicky značkovaný korpus
- formální popis gramatiky
- anotace na úrovni částí věty (crowdsourcing)
  - sentence (věta)
  - clause (věty v souvětí)
  - nominální fráze (NP)
  - verbální fráze (VP)
  - koordinace (COORD)

Odkazy

- synt
- http://nlp.fi.muni.cz/projekty/wwwsynt/
- set

http://nlp.fi.muni.cz/projekty/set/wwwset.cgi/first_page

- Penn Treebank Project (University of Pennsylvania)
- http://www.cis.upenn.edu/~treebank/