PA153 - NLP

1. Formální gramatiky

- soubor pravidel, jež generují či rozpoznávají věty
  (gramatiky / automaty)
- terminální a neterminální soubory
  - S → NP VP , NP → A N
    (věta; jm. fráze; přísud. tr.; adj.; subst.)
- množina: T terminály (slova, abeceda), N neterminály (synt. kat.), podmn. kart. součinu (N ∪ T)* N, počáteční symbol S gramatiky

Jazykové inženýrství

- všeobecně NLP – na FI tvorba slovníků (DEBdict, ...), korpusy, ...

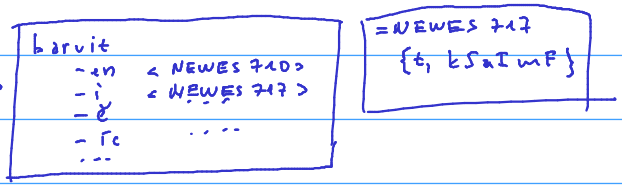
2. Morfologická rovina popisu jazyka

- nejnižší rovina
- cílem dostat: 1/ lemma, 2/ gramatický význam
- problémy:
  - jak popisovat (pražský vs. brněnský popis)
  - do jaké míry lemmatisovat - otcova → otcův/otec ?
  - co s dubletami (myslet/myslit)
  - manuál pro ruční anotaci musí být jednoduchý a jasný (ne jako pražský 300str. man.)
    - když se na tom neshodnou lidi, stroj to nemůže udělat líp

Anotační systémy

- pozicní (prg): význam zn. určen pozicí: NNIS4-----A---- – Hajič, Hlaváčková, Hladká
- atributový (brno): dvojice atribut-hodnota: k1gInSc4 (druh, rod, čís., pád) – Sedláček, Čmejrek
- heterogenní (brat.): komb. atr. a pos, pozice kódují str. i hodnotu
  - rychle vyčerp. značky
- BNC použil pevnou množinu hotových značek pro ang (AJC - adjective, comparative)

AJKA

- organizuje data do lemma:vzor → 

- pro cs asi 1800 vzorů
- odtrháváním a připojováním intersegmentů a koncovek → generování tvarů
- db analysatoru: množ. koncovek, vzory + interseg., přípustné kombinace

- obecně: úkoly morf analysatoru:
  - vytvořit všechny verze (tvary)
  - vybrat nejpravděpodobnější (tagger)
  - zanalysovat neznámá slova (guesser)

AJKA vs. MAJKA https://nlp.fi.muni.cz/czech-morphology-analyser/majka.html

- A-- už nevyvíjení, blbě napsaná, velká, nepřehledná
  - chyby v práci se složeninami
  - "ajka dělá nějakou magii se skládáním slov, blíže jsem nezkoumal"
  - špatný handling záporných tvarů "být"
  - podivnosti s vidy

2.1. Reprezentace významu

- sémantika nezávisí na kontextu (užití, situaci), ale pragmatika může význam změnit
- repr. predikátovou logikou (omezená vyjadřov. síla), TIL (složitá)
- nutno zpracovat deixi, anafory, katafory, ...
  - Koupil jsem si auto. [To auto] je bílé.
- pro analýzu nutno mít KB (DB znalostí) - ontologie: klasifikace/vztahy objektů universa (eg. WordNet)
- TIL pracuje i s možnými světy, časem, ...
- Levin - klasifikace sloves (pohybu, emocí, změny, ...)
- víceznačnost - problém vybrat správného významu - Lesk ův alg. (společná slova v kontextu zkoumaného a v def. jednotlivých významů zkoumá.)

3. Morf. analýza

- viz #2

3.1 Sémantické role a pády

- role: agens, patiens
  - valenční rámce (třeba u sloves v lexikonu); nutno ověřovat na datech v korpusu
  - FI: Vebalex (Pala)
- pády: oblique cases - nesyntaktické pády

4. Algoritm. popis deklinace, konj., lemmatizace

- vzory, kmeny - viz #2

4.1. Sloves. val. rámce a třídy

- rámce: E. Rosch: teorie prototypů - obecný zástupce třídy + specifikátory
  - "židle je víc nábytek než sporák" taky nábytek, ale víc se liší od klasického nábytku

5. Lexikální analýza

viz # 2.1.

5.1. Syntaktická analýza, koncepce

- gramatiky (viz #1)

6. Elektronické slovníky

- Brown Corpus, Longman, COBUILD, TEI (guide, jak dělat slovníky - struktura)
- v XML, DTD popisuje strukturu; XSLT - transformace do jiných formátů
- dict writing systems - app na tvorbu slovníků (propojení s korpusy, morfoanalyzátory, ...)
- komerční vs. open - DEB (dict edit & browse)
  - DEBdict, DEBvisDict, JSP
- praktická vs. teoretická lexikografie
  - praktická: analýza slov. zás., teorie tvorby slov.,
- lexikální databáze - detailní DB s valencemi, gramatikou, styl. zařazením, ...
  - PRALED, DANTE (eng)
    - Pražská Lexikální Databáze
    - Database of ANalysed Texts of English
- macrostructure - struktura většího objektu (asi celého záznamu vč. příkladů, valencí, etc.)
- microstructure - struktura jednoho malého záznamu

6.1. Analýza promluvy

- odkazy v textu: anafora (vzad), katafora (vpřed), endofora (uvnitř textu), exofora (vně textu),
  koreference (různé výrazy označující totéž (Klaus, bývalý prezident)
- anaphora resolution - rozpoznávání a. - k čemu referují se impl. pomocí zásobníku (naivní přístup)
  - umělost míť znalosti o světě - KB (knowledge base)
    - SUMO-MILO, WordNet, ConceptNet

7. Víceznačnost v NL, desamb., evaluace výsledků desamb.

- WSD - Word Sense Disambiguation - na základě kontextu
- $f_{WSD}(w, c) \to s$, kde w = množina slov, c = množ. kontextů, s = množ. kontextů
- Leskův algoritmus - porovná výskyt slov v kontextu zkoum. slova se slovy v definicích významů zkoum. slov, vybere nejvyšší shodu (naivní)
- všechny WSD závisí na nějaké DB významů; ty se liší (SSJC x SSČx ...)
- komponentová analýza (+HUMAN, +FEMALE) → sémantické třídy (příbuzná slova) - sém. sítě → odvozování
  - WordNet - významy uloženy podobně jako v mozku (asi?) = ontologie (konceptualisace)
    - hyperonymie, hyponymie, tropon, meron, ...
- rozliš. lex. významu (pro stroje): vzdálenost mezi koncepty

7.1. Tagsety

- prg x brno - viz #2.

8. Notace pro synt. analýzu, parciální analyzátory

- viz #1
- parciální synt. anal. - nepracuje s úplnými texty, s lexikálním významem
  - užití pro rozpoznávání klíčových frází - TopPicks

8.1. Logická analýza významu věty

- TIL - množina možných světů, pravdy závisí na světě a na čase
  - intense
  - extense ?

9. Desambiguace pro NL

- morfologická (tagger), lexikální (WSD)
- M. součástí analysátorů (zřejmě?) – Majka (cz), Treetagger (enTenTen), CLAWS (BNC)
  - vybírá nejpravděpodobnější kategorii podle kontextu, příp. dle frekvencí výskytu kandidátů u trénovacích datech
  - když není, guesser – dost random výběr
- WSD (viz #7)

9.1. Princip kompozicionality (log./sém. anal., TIL)

- p. kompozicionality = význam skup. slov je složeninou významů jednotlivých slov
  - neplatí u pevných kolokací = MWE (multiword expressions) – nutno detekovat např. pro MT
  - MWE by měly mít vlastní slov. hesla (asi?)
  - pevné (zakopaný pes) vs. vzory (vzít <někoho> na hůl)
- log. anal., TIL, PL1 – ???

10. Sémantické značkování, desambiguace

- WSD viz #7
- sém. značkování: sém. sítě jako WordNet užívají XML a rekursivní reference u jednotl. synsetů (synonym. řada) ke svým hypo-, hyper-, mero-, ...-nymům

10.1. Pragmatická analýza

- p. jevy jsou mimojazykové
- externí prag.: komunikační situace (mluvčí, posluchač, čas, ...)
- interní prag.: vztah mezi mluvčími a významem
  - oznámení, rozkaz, ...
- implikatury – intence mluvčího
  - lokuce (co bylo řečeno), ilokuce (co bylo interpretováno), perlokuce (důsledky, činy)
- Griceovy maximy (kvality, kvantity, ...)

11. Lexikální DB a nástroje

- viz #6

11.1. Interní a externí pragmatika

- viz # 10.1

12. DC gramatiky a využití Prologu v synt. anal. NL

- DCG = gramatiky vymezených klausulí
- umožňují zachovat nekontextovou podobu pravidel a získat kontextovou citlivost
- pravidlo: neterm. → predikát s argumenty ($S \to N_p V_p \sim s(s(N_p, V_p))$)
- rekurse!
- Prolog: zápis predikátů, inference – logické vyvozování faktů

12.1. Form. representace významu, PL1, TIL

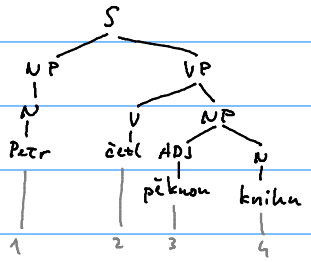
FOPL

- PL1 – omezená, TIL složitý...
- ne všechno je predikát

• • •

15. Reprezentace pro syntakt. analýzu

- složkové – zobrazení složek věty (a jejich závislostí)



- závislostní – zobrazení jen synt. závislostí slov

! největší srando je u stroj. analysátorů poznat, který z výstupních stromů je správný, bo oni to nepoznají