OBRADA PRIRODNOG JEZIKA (NLP)

Lingvističke osnove NLP-a

25. rujna 2009.

Vrste riječi

- Imenice
- Pridjevi
- Zamjenice
- Glagoli
- Prilozi
- Prijedlozi
- Veznici

... i još mnogo toga (brojevi, vlastita imena, kratice, akronimi itd.)

Morfologija i POS

- Riječi podliježu morfološkim radnjama kao što su dekliniranje (imenice, pridjevi, zamjenice), konjugiranje (glagoli), tvorbi složenica, i drugima
- Part-of-Speech (POS) je učestali (kratki) slijed riječi u rečenici (npr. New York, Ustavni sud) koje funkcioniraju kao cjelina i u kojemu svi ili samo neki dijelovi podliježu morfološkim radnjama
- POS može biti i sama riječ, pa čak i "prazna" riječ
- Zadatak NLP-a je označiti (annotation, labeling, tagging) POS-ove, ali njihove članove

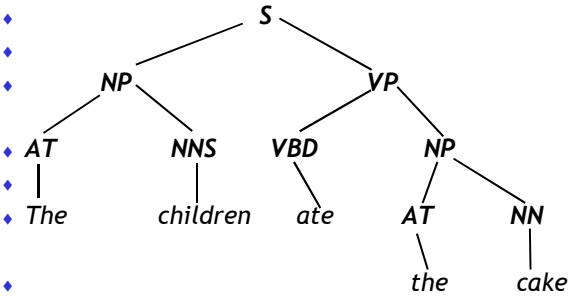
Označavanje POS-ova

- Children (NOUN) eat (VERB) sweet(ADJECTIVE) candy(NOUN)
- The(ARTICLE) children(NOUN) ate(VERB) the(ARTICLE) cake(NOUN)
- The(ARTICLE) news(NOUN) has(AUXILIARY) been(MAIN VERB) quite(ADVERB) sad(ADJECTIVE) in(PREPOSITION) fact(NOUN) .(PERIOD)

Označavanje u hrvatskome

- Radi složenosti hrvatske morfologije označavanje u hrvatskome izuzetno je složeno
- Hrvatski morfološki leksikon http://www.hnk.ffzg.hr/hml/ koristi MSD sustav označavanja http://nl.ijs.si/ME/V2/msd/html/node15.html#SECTION038100000000000000000 koji je razvijen za neke istočnoeuropske jezike + engleski http://nl.ijs.si/ME/V2/msd/html/
- Zato ćemo se zadržati na engleskim primjerima i njihovom konvencionalnom označavanja

Rečenično stablo (parsing)



CFG (Context Free Grammars)

- Sastoje se od nezaključnih (non-terminal) i zaključnih (terminal) simbola
- Nezaključni simboli povezuju se slijednim pravilima, s time da je prvi nezaključni simbol takozvana "majka" (mother)
- Slijedno pravilo određuje odnos između jednog nezaključnog simbola i njemu "podređenih" nezaključnih simbola, s time da to pravilo ne smije ovisiti o prethodnim nezaključnim simbolima, niti o simbolima koji slijede nakon "podređenih"

Primjer jednostavne CF gramatike

- Gramatička pravila
- S -> NP V
- NP -> N
- S: rečenica, NP: imenska fraza, V:glagol, N: imenica
- S, NP, V i N su neterminalni simboli
- Rječnik
- N -> John, Tom, ……
- V -> walks, talks, eats, went …..
- Na kraju se neterminalni simboli povezuju sa terminalnima, moji tvore sadržaj rječnika

Sustavi označavanja

- U hrvatskome se, kako je već rečeno, koristi, MSD sustav, ali samo za označavanje neterminalnih simbola na najnižoj razini rečeničkog stabla; kompletna CF gramatika za hrvatski nije napravljena
- U engleskome se koristi nekoliko standardnih sustava označavanja: Brown Tag Set, Penn Treebank Tag Set
- Primjeri tagginga: *NP:* vlastito ime (alternativa je Noun Phrase), *NN:* imenica u jednini, NNS: imenica u množini, *AT*: član (a, an ili the) itd.

Stohastičke gramatike

- Dobivaju se dodavanjem vjerojatnosti na "algebarsku" gramatičku strukturu, npr. na CF gramatiku
- Stohastički dodatak gramatici izomorfan je njezinoj algebarskoj strukturi
- Ovo znači da je svako slijedno pravilo iz CF gramatike opisano i svojom vjerojatnošću
- Pored apriornih vjerojatnosti pridruženih pravilima mogu se dodati i uvjetne vjerojatnosti koje se odnose na nizove pravila

Svojstva stohastičkih gramatika

- Robusnost: ulazni podatci često su opterećeni "šumom", npr. pravopisnim i tipografskim pogreškama, nepredvidivim sintaksnim konstrukcijama i sličnim. Stohastičke gramatike znaju raditi s takvim šumom
- Prenosivost: stohastičke gramatike se mogu prenositi (kao modeli) s jezika na jezik, putem "učenja" na tekstovnim korpusima
- Sposobnost poopćavanja: mogu raditi i "zaključivati" nad podatcima koje nikada nisu prije vidjele, tj. koji nisu bili obuhvaćeni korpusom za učenje

Primjer stohastičke CF gramatike

- London walks
- 1. S -> NP V .7
- 2. S -> NP .3
- 3. NP -> N .6
- 4. NP -> N N .2
- 5. N -> London .6
- 6. N -> walks .4
- 7. V -> walks 1.0
- Ako se rečenica tretira kao "London šeta" primjenjuju se pravila 1, 3, 5, 7 i dobiva se "težina" (.7)(.8)(.6)(1.0)=.336
- Ako se tretira ka imenska fraza (Londonske šetnice) pravila su 2, 4, 5, 6, tako da je težina(.3)(.2)(.6)(.4) = 0.0144

Prednosti stohastičkih gramatika

- Vjerojatnosti se pridružuju CFG pravilima, što znači da možemo pojednostaviti CFG strukturu, a da kroz vjerojatnosti dobijemo puno uporabljivih informacija o jeziku
- Stohastičke CF gramatike omogućuju npr. modeliranje učenja jezika, modeliranje promjena u jeziku, modeliranje pojava pogrešaka u pisanju i njihovo ispravljanje, a u slučaju paralelnih dvojezičnih korpusa temeljni su alat za stohastičko strojno prevođenje

Semantika

- Semantika se bavi značenjima riječi, POS-ova i iskaza
- U NLP-u semantika se dijeli na leksičku semantiku i kombinacijsku semantiku
- Leksička semantika se bavi hiperonimijom, hiponimijom, antonimijom, meronimijom, holonimijom, sinonimijom, homonimijom, polisemijom i homofonijom
- Kombinacijska se semantika bavi značenjima cjeline i dijelova s naglaskom na one leksičke kombinacije koje su značenjski pomaknute od "prostog zbroja" značenja njezinih djelova

Leksička semantika

- Hiperonimija/hiponimija: nadređeni i podređeni pojmovi (motorno vozilo, automobil)
- Antonimija: suprotni pojmovi (naprijed, natrag)
- Meronimija/holonimija: odnos dijela i cjeline (automobil, automobilski motor)
- Sinonimija: istoznačni pojmovi (automobil, auto)
- Homonimija: raznoznačni pojmovi (kosa na glavi, kosa za kositi travu)
- Polisemija: višeznačni pojmovi (matica)
- Homofonija: pojmovi koji se isto izgovaraju a različito pišu (*knight*, *night*; rijetka u hrvatskome)

WordNet

- WordNet je računalno pohranjena leksičko-semantička mreža koja je razvijena za engleski jezik na Sveučilištu Princeton u Sjedinjenim Državama prije 20-tak godina http://wordnet.princeton.edu/
- EuroWordNet je projekt EU s ciljem da s slične mreže naprave za sve europske jezike http://www.illc.uva.nl/EuroWordNet/
- Hrvatski WordNet je tek u začetku http://rmjt.ffzg.hr/p3.html

Semantičko parsiranje rečenice

- Rečenica se može predstaviti i putem svojih semantičkih dijelova (agenta, pacijenta, instrumenta, cilja itd.)
- *Dječak* (AGENT) *nas* (PACIJENT) je pogodio *s loptom* (INSTRUMENT)
- Semantički dijelovi često se preklapaju sa sintaksnim dijelovima rečenice (subjekt, objekt ...)
- Međutim, nastupaju komplikacije radi postojanja izravnog i neizravnog objekta, zatim u slučaju aktivne i pasivne rečenice, itd.

Pragmatika

- Pragmatika nadilazi istraživanje značenja rečenice i istražuje što je govornik cjelinom htio da iskaže
- Bavi se kako se iste rečenice koriste u različitim situacijama
- Predmeti istraživanja: prirodnost, prihvatljivost, primjernost, određenost, izravnost itd.
- Bavi se i anaforama radi izdvajanja informacije:

  I nema *sestre ni brata*
  I nema *oca ni majke*
  I nema *drage ni druga*

(Tin Ujević: „Svakidašnja jadikovka")