Uvod u obradu prirodnog jezika

12.1. Označavanje vrste riječi (Part-of-speech tagging)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Vrste riječi

- Vjerojatno od Aristotela (384-322 pne) postojala je ideja o vrstama riječi
  - tj. leksičkim kategorijama, POS
- Dioysius Thrax (100 pne) kaže da ima 8 vrsta riječi
  - Thrax: imenica, glagol, član, prilog, prijedlog, veznik, čestica, zamjenica
  - Školska gramatika: imenica, glagol, pridjev, prilog, prijedlog, veznik, zamjenica, usklik, broj, čestica

Vrste riječi

Promjenjive

Imenice

- Vlastite: IBM, Italija
- Opće: mačka/mačke, snijeg

Glagoli

- Glavni: vidi, stajale
- Pomoćni: će/ćemo/ćete

Pridjevi

dobro/bolje/najbolje

Zamjenice

on ona onaj

Brojevi

jedan drugoj trima

Nepromjenjive

- Prilozi: kada gdje kamo zašto
- Prijedlozi: od iza po
- Veznici: i ili ali
- Uzvici: ah oh eh
- Čestice: ne zar evo

Označavanje vrste riječi (POS označavanje)

- Riječi često mogu imati više vrsta: dan
  - Danas je dobar dan. = imenica
  - On je bogom dan suprug. = pridjev
  - Poklon je dan njemu. = glagol

- Problem označavanja vrste riječi je određivanje POS oznake za određen primjerak riječi

POS označavanje

- Ulaz: Igra dobro s drugima
- Višesmislenost: N/V N/A/R S A
- Izlaz: Igra/V dobro/R s/S drugima/A

- Korištenje:
  - Text-u-govor (kako se izgovara "luk")
  - Možemo napraviti regularne izraze kao A* N+ kako bi dobili fraze
  - Kao ulaz za ubrzavanje potpunog parsiranja
  - Ako se zna oznaka, možemo se vratiti na nju kasnije radi nekih drugih zadataka

Performanse POS označavanja

- Koliko je dobro označenih riječi (točnost):
  - oko 97%
  - ali osnova je već oko 90%
    - osnovno POS označavanje je označavanje na "najjednostavniji" mogući način
      - označi riječ s njenom najfrekventnijom oznakom
      - označi nepoznate riječi kao imenice
  - djelomično lako jer
    - mnoge riječi nisu višesmislene
- I ljudi ponekad imaju problema s određivanjem vrste riječi.

Koliko je teško POS označavanje

- Oko 11% riječi u Brown korpusu su višeznačne obzirom na POS označavanje
- Ali su većina njih učestale riječi: npr. that
  - I know that he is honest = IN
  - Yes, that play was nice = DT
  - You can't go that far = RB
- 40% oblika riječi su višeznačne

Uvod u obradu prirodnog jezika

12.2. Modeli sekvenci kod POS označavanja (Sequence models in POS tagging)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Izvori informacija

- Koji su glavni izvori informacija za POS označavanje?
  - Znanje o susjednim riječima
    - Bill saw that man yesterday
    - NNP NN DT NN NN
    - VB VB(D) IN VB NN
  - Znanje o vjerojatnosti riječi
    - man se rijetko koristi kao glagol...
- Znanje o vjerojatnosti riječi se pokazuje najkorisnijim, ali znanje o susjednim riječima također pomaže

Više i bolje osobine → Tager temeljen na osobinama

- Mogu biti iznenađujuće dobri ako se gleda sama riječ:
  - riječ — ili: ili → C
  - riječ s malim slovima — nad: nad → S
  - prefiksi — reprodukcija: re- → Nc
  - sufiksi — nositi: -iti → V
  - riječ s prvim velikim slovom — Meridian: CAP → Np
  - oblik riječi — 35-ta: d-x → A
- Onda napraviti maxent (ili kakav god) model za predviđanje oznake
  - Maxent P(t|w): 93.7% ukupno / 82.6% za nepoznate

Točnosti POS označavanja

- Približne točnosti
  - Najveća frekvencija ~90% / ~50%
  - Trigram HMM ~95% / ~55%
  - Maxent P(t|w) ~93.7% / ~82.6%
  - Tnt(HMM++) ~96.2% / ~86.0%
  - MEMM ~96.9% / ~86.9%
  - Dvosmjerne ovisnosti ~97.2% / ~90.0%
  - Gornja granica: ~98% (ljudski)

Kako poboljšati nadzirane rezultate?

- Izgraditi bolje osobine!

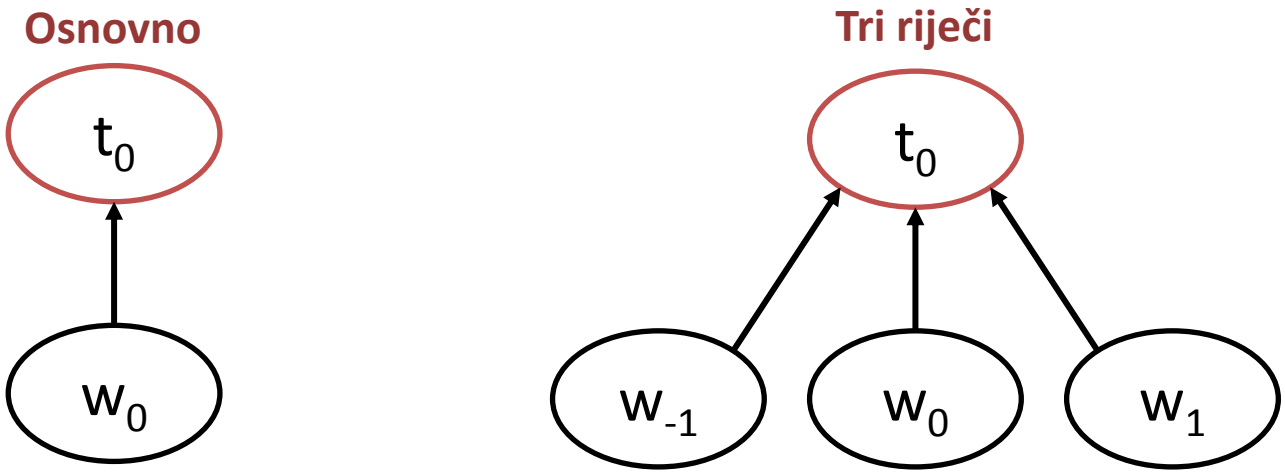
		S	
V	P	N	A
Vrti	se	oko	sunca

  - Ovo smo mogli popraviti izgradnjom osobine koja gleda sljedeću riječ

A			
NP	N	V	A
Dobri	ljudi	ostaju	dobrim

  - Ovo smo mogli popraviti izgradnjom osobine koja vezuje riječi s prvim velikim slovom i riječi s malim slovima

Označavanje bez informacija o sekvenci



Model	Osobine	Tokeni	Nepoznato	Rečenica
Osnovno	56805	**93.69%**	82.61%	26.74%
3 riječi	239767	**96.57%**	86.78%	48.27%

Korištenje samo riječi kod izravnog klasifikatora radi dobro kao osnovni (HMM ili diskriminativni) model sequence!!!

Rezime POS označavanja

- Za POS označavanje, promjena iz generativnog u diskriminativni model **ne daje** značajna poboljšanja
- Jedna od dobiti su **preklapajuće osobine opservacije**.
- MEMM dozvoljava integraciju bogatih osobina opservacija, ali može patiti zbog nekorištenja sljedećih opservacija;
  Ovaj efekt se može ublažiti dodavanjem ovisnosti o sljedećim riječima
- Ova dodatna snaga (MEMM, CRF, Perceptron) modela pokazuje poboljšanja u točnosti
- Što je **veća točnost** diskriminativnog modela to ga potrebno **duže trenirati**.