

Uvod u obradu prirodnog jezika

Uvod

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Što je obrada prirodnog jezika?

- **Odgovaranje na pitanja (Question answering)**

- IBM-ov Watson je pobijedio u kvizu Jeopardy

Voditelj

Fox show featuring characters named Itchy and Scratchy



Watson

Who are the Simpsons?

Voditelj

William Wilkinson's "An Account of the Principalities of Wallachia and Moldavia" inspired this author's most famous novel



Watson

Who is Bram Stoker?



Što je obrada prirodnog jezika?

- **Ekstrakcija informacija (Information extraction)**

Subject: Sastanak zavoda

Date: 8. veljače 2012.

To: Branko Žitko

Sastanak zavoda će se održati sutra u sobi 30 od 12:00 do 13:30.

Pozdrav



Događaj: Sastanak zavoda

Datum: 09.02.2012.

Početak: 12:00

Kraj: 13:30

Mjesto: Soba 30

Što je obrada prirodnog jezika?

- **Ekstrakcija informacija i sentimentalna analiza**



Veličina i težina

- ✓ Lijepo i kompaktno za nosi
- ✓ Dok je kamera mala i lagarene teške, glomazne, profesionalne
- ✗ Kamera izgleda veoma kruščastoga morate biti veoma oprezni

Već na prvi pogled, kvaliteta izrade ostavlja dobar dojam. Tijelo je čvrsto i odaje pozitivan **dojam** kojeg samo malo kvare pojedine tipke sa "gnjecavim" osjećajem pri pritisku. Kotačić za odabir načina rada se vrti relativno lako te ga je moguće nehotice zakrenuti pri vađenju ili vraćanju aparata u džep ili torbicu, te je potrebno ponekad obratiti pozornost na njega prije fotografiranja. Zakretni LCD zaslon je obješen na potpuno metalnu konstrukciju koja je vrlo čvrsta i djeluje kao da može izdržati godine svakodnevnog zakretanja i navlačenja u svim smjerovima. Ponekad nas je smetao **nedostatak većeg gripa** tako da će E-PL5 djelovati nespretno u većim rukama, no to je subjektivno. U redakciji se ipak svi slažemo da najbolje ležanje u ruci pružaju Sony NEX-5 i NEX-6 koji imaju izbočeni grip. Glavni **izbornik** je lagan za korištenje, a mnogobrojne opcije logično raspoređene. Najbitnije opcije poput ISO vrijednosti, balansa bijele i sličnih postavki se podešavaju putem brzog izbornika koji se poziva "OK" tipkom.

Iako PEN E-PL5 daje odlične fotografije već u automatskom modu i prikidan je za korištenje i kompletnim amaterima, svoju punu vrijednost pokazuje u rukama naprednih korisnika.

Olympus naime nikad nije škrtario sa opcijama pa tako E-PL5 ima niz detalja koje nedostaju mnogim profesionalnim DSLR aparatima, kao što je bežična kontrola do tri skupine bljeskalica, podešavanje brzine kontinuiranog okidanja od 1 do 8 fotografija u sekundi, pixel mapping, mogućost potpune preraspodjele kontrola na tipkama, direktni odabir fokusnih točaka i njihove veličine, mijenjanje smjera rotacije kontrolnog kotačića, čak 6 vrsta bracketinga (ISO, WB, Flash, Art filteri, HDR, AE), promjena EV i ISO koraka, antivibracijska odgoda okidanja sa podesivim vremenom, zasebna korekcija sva tri načina mjerjenja svjetla... opcijama nema kraja!

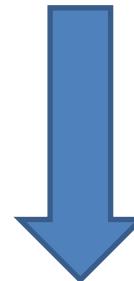
Što je obrada prirodnog jezika?

- **Strojno prevodenje (Machine translation)**

- Google Translate

这不过是一个时间的问题

potpuno automatski



kao pomoć
ljudskim prevoditeljima

This however is a matter of time

To, **ipak**, je pitanje vremena

- međutim
- ma kako
- kako god
- ma koji

Jezične tehnologije

Uglavnom riješene

Detekcija SPAM-a
(SPAM detection)

napravi dijagram klasa.



buy V1AGRA ...



Označavanje dijelova teksta
(Part of speech tagging)

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Prepoznavanje imenovanih entiteta
(Named entity recognition)

PERSON

Po mišljenju **Orsata Frankovića**, direktora studija
Laboratorium, kad **Hrvatska** postane...

ORG

LOC

Jezične tehnologije

Dobro napreduju

Sentimentalna analiza

Najbolji restoran u Splitu!



Konobar nas je ignorirao 20 minuta!

Određivanje smisla riječi (Word sense disambiguation)

Trebam nove baterije za mog **miša**.



Strojno prevođenje (Machine translation)

今天是美好的一天

Danas je lijep dan.

Rješavanje odnosa (Coreference resolution)

On je ulovio muhu u letu.

Parsanje (Parsing)



Mogu vidjeti Šoltu s prozora!

Ekstrakcija informacija (Information extraction)

Pozvani ste na zabavu
sljedeći Petak 27 Svibnja
u 8:30



Zabava
27.05.2012
20:30

Jezične tehnologije

Još uvijek jako teško

Odgovaranje na pitanja
(Question answering)

Kolika je učinkovit ibuprofena u smanjenju povišene temperature kod bolesnika sa akutnom bolešću?

Parafraziranje

XYZ je stekao ABC jučer.
ABC je preuzeo XYZ

Sažimanje
(Summarization)

Fondovi idu prema gore
Dionice su skočile
Cijene nekretnina rastu



Ekonomija je dobra

Dijalog



Gdje se prikazuje
Gospodar prstenova?

U kinu Central u
19:30. Želite li
rezervirati kartu?



Obrada prirodnog jezika

- Višeznačnost čini obradu prirodnog jezika teškom

Brat mi je slomio ruku

Mnogi naučnici **sumnjaju** da se u pećini kriju značajna arheološka blaga

Iskopao je crno zlato u dvorištu

Popila je kavu i vruću čokoladu sa **šlagom**

Uvod u obradu prirodnog jezika

2.1. Regularni izrazi

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Što su regularni izrazi?

- **Formalni jezik za specificiranje textualnih nizova.**
 - Prepostavimo da u tekstu moramo pronaći riječ *sustav*
 - Ona se može izraziti na nekoliko načina
 - *sustav*
 - *sustavi*
 - *Sustav*
 - *Sustavi*
 - *sustavima*
 - *Sustavima*

Pisanje regularnog izraza?

- Regularni izraz je oblika */uzorak/zastavice*
- Regularni izrazi bez zastavica

Izraz	Primjer	Objašnjenje
/a/	Anamarija	prvo pojavljivanje znaka <i>a</i>
/A/	Anamarija	prvo pojavljivanje znaka <i>A</i>
/am/	Anamarija	prvo pojavljivanje niza <i>am</i>

- Zastavica *i* ignorira velike i male znakove

Izraz	Primjer	Objašnjenje
/a/i	Anamarija	prvo pojavljivanje malog ili velikog znaka <i>a</i>
/Am/i	Anamarija	prvo pojavljivanje niza <i>am</i> bez obzira na veličinu znaka, odnosno <i>am</i> , <i>AM</i> , <i>aM</i> ili <i>Am</i>

Pisanje regularnog izraza?

- Zastavica **g** nastavlja globalnu pretragu

Izraz	Primjer	Objašnjenje
/a/g	Anamarija	sva pojavljivanja znaka a
/a/ig	Anamarija	sva pojavljivanja malog ili velikog znaka A

- Zastavica **i** ignorira velike i male znakove

Izraz	Primjer	Objašnjenje
/a/i	Anamarija	prvo pojavljivanje malog ili velikog znaka a
/Am/i	Anamarija	prvo pojavljivanje niza am bez obzira na veličinu znaka, odnosno am , AM , aM ili Am

Regularni izrazi: Skup znakova

- skup znakova

Uzorak	Primjer	Objašnjenje
[sS]ustav	Sustavni sustavi	nizovi znakova koji počinju sa znakom s ili S i iza njega slijedi niz znakova ustav
[aeiou]	samoglasnici	svi samoglasnici a e i o u

- negirani skup znakova

znak ^ na početku označava negaciju samo kada je na prvom mjestu iza uglatih zagrada

Uzorak	Objašnjenje
[^aeiou]	Danas u 3 sata.
[^0123456789]	Danas u 3 sata.

Regularni izrazi: Raspon znakova

- **raspon znakova**

znak – skup znamenaka u rasponu između dva znaka

Uzorak	Primjer	Objašnjenje
[A-Z]	Broj Pi = 3.14.	sva velika slova (bez "naših" znakova)
[a-z]	Broj Pi = 3.14.	sva mala slova (bez "naših" znakova)
[0-9]	Broj Pi = 3.14	sve brojčane znamenke ekvivalentno [0123456789]
[m-s]	Broj Pi = 3.14	sva mala slova od m do s

Regуларни изрази: Specijalni znakovi

- rezervirani znakovi

znakovi `+ * ? ^ $ \ . [] { } () | /` imaju specijalno značenje i potrebno je staviti `\` ispred njih

kod raspona znakova potrebno je staviti `\` ispred `\ -]`

Uzorak	Primjer	Objašnjenje
<code>\+</code>	$1 + 1 = 2$	znak <code>+</code>
<code>[+\-]</code>	$3 + 2 - 1 = 4$	znak <code>+</code> ili znak <code>-</code>

- specijalni znakovi

Uzorak	Objašnjenje
<code>\t</code>	TAB znak (ASCII 9)
<code>\n</code>	LINE FEED (ASCII 10)
<code>\r</code>	CARRIAGE RETURN (ASCII 13)

Regуларни изрази: Клase znakova

- točka

Uzorak	Primjer	Objašnjenje
.	Broj Pi = 3.14.	bilo koji znak osim novog reda ekvivalentno [^\n\r]

- klase znakova

Uzorak	Primjer	Objašnjenje
\w	Broj Pi = 3.14.	alfanumerički znakovi i _ ekvivalentno [A-Za-z0-9_]
\W	Broj Pi = 3.14.	negacija od \w [^A-Za-z0-9_]
\d	Broj Pi = 3.14.	svi numerički znakovi ekvivalentno [0-9]
\D	Broj Pi = 3.14.	negacija od \d [^0-9]
\s	Broj Pi = 3.14.	prazni znakovi (razmak, novi red, tabulator)
\S	Broj Pi = 3.14.	negacija od \w

Regularni izrazi: Kvantifikatori

- **Kvantifikatori**

- + 1 ili više pojavljivanja
- * 0 ili više pojavljivanja
- ? 0 ili jedno pojavljivanje

Uzorak	Primjer	Objašnjenje
e+	b be bee	jedno ili više pojavljivanja znaka e
r\w+	riba ribi grize rep	nizovi kojima je prvo slovo r i iza njega 1 ili više alfanumeričkih znakova
e*	b be bee	nula ili više pojavljivanja znaka e NAPOMENA: uključuje i prazne znakove
r\w*	riba ribi grize rep	nizovi kojima je prvo slovo r i iza njega 0 ili više alfanumeričkih znakova
past?i	pasi će pasti travu	nula ili jedno pojavljivanje znaka t

Regularni izrazi: Grupiranje i alternacija

- Grupiranje i reference

Uzorak	Primjer	Objašnjenje
(ha)+	hahaha haa hah!	ha je grupa koji se ponavlja 1 ili više puta
(\w)a\1	pad mam sam gag	\1 se referencira na prvu grupu \w

- Alternacija |

Uzorak	Primjer	Objašnjenje
p(a e u)t	pat pet pit pot put	znak p iza kojeg može biti a e ili u i na kraju t
p(ame i)t	pametno piti	znak p iza kojeg može biti ame ili i i na kraju t

Regularni izrazi: Sidra

- Početak i kraj

- početak linije
- kraj linije

Izraz	Primjer	Objašnjenje
<code>/^\w+/.gm</code>	Jedan dva. Tri četiri.	alfanumerički znakovi na početku linije
<code>/\w+\.\$/gm</code>	Jedan dva. Tri četiri.	alfanumerički znakovi i točka na kraju linije

- granice riječi `\b`

granica između alfanumeričkog znaka i nealfanumeričkog znaka

Uzorak	Primjer	Objašnjenje
<code>\br</code>	riba ribi grize rep	r je na početku niza alfanumeričkih znakova
<code>[aeiou]\b</code>	riba ribi grize rep	a e i o u je na kraju niza alfanumeričkih znakova

Primjeri

- Pronađite u tekstu sve instance riječi "on".

on

On je onda u ponoć otišao kući. Sutradan je on ponudio bonus.

[oO]n

On je onda u ponoć otišao kući. Sutradan je on ponudio bonus.

\b[oO]n\b

On je onda u ponoć otišao kući. Sutradan je on ponudio bonus.

Regularni izrazi: Pogreške

proces koji smo upravo prošli temelji se na utvrđivanju dvije vrste pogrešaka

- lažno pozitivni (TIP I)
 - Odgovarajući nizovi koji se ne bi trebali podudarati.
(onda, ponoć, bonus)
- lažno negativni (TIP II)
 - Ne označavanje nizova koji bi se trebali označiti.
(On)

on

On je **onda** u **ponoć** otišao kući. Sutradan je **on** **ponudio bonus**.

Regularni izrazi: Pogreške

- Obrada prirodnog jezika uvijek se bavi sljedećim pogreškama
- Smanjenje stupnja pogreške za aplikacije često uključuje dva pristupa rješavanja pogrešaka:
 - povećanje **točnosti** ili **preciznosti**
(smanjenje lažno pozitivnih)
 - povećanje **pokrivenosti** ili **odziva**
(smanjenje lažno negativnih)

Regularni izrazi: Sažetak

- Regularni izrazi igraju iznenađujuće veliku ulogu
 - sofisticirani nizovi regularnih izraza često predstavljaju prvi model za bilo koju obradu teksta
- Za mnogo teže zadatke koriste se klasifikatori strojnog učenja
 - regularni izrazi se koriste kao obilježja u klasifikatorima
 - mogu biti vrlo korisni za obuhvaćanje općenitosti

Uvod u obradu prirodnog jezika

2.2. Tokenizacija (opojavničenje) riječi i korpusi (Word tokenization and corpuses)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Koliko ima riječi u rečenici?

- "Ja radim ovaaj uglav- uglavnom aaa obradu poslovnih podataka"
 - fragmenti, ispunjeni pauzom
- "**mačka** u šeširu je drugačija od drugih **mačaka!**"
 - **Lema:** isti korijen, dio govora, smisao grube riječi **mačka** i **mačaka** = ista lema
 - **Oblik riječi:** puni utjecaj oblika riječi **mačka** i **mačaka**= različita forma riječi

Koliko ima riječi u rečenice?

"Oni leže na livadi u Dugom ratu i gledaju na druge livade"

- **Tip riječi:** jedinstveni element rječnika
- **Pojavnica (Token):** primjerak leme u tekstu
- Koliko ima riječi u rečenici
 - 12 pojavnica
(ili 11 – Dugom ratu – jedna pojavnica)
 - 11 tipova
(ili 9 – Dugom ratu – jedna pojavnica)
(ili 8 – livadi i livade – ista lema)

Koliko ima riječi u korpusu?

N = broj pojavnica

V = rječnik = skup tipova riječi

|V| = broj tipova riječi u rječniku

Heapsov zakon = Herdanov zakon

$$|V| = kN^{\beta} \quad 0.67 < \beta < 0.75$$

	Pojavnice = N	Tip = V
Centrala telefonskih razgovora	2.4 milijuna	20 tisuća
Shakespeare	884 000	31 tisuća
Google N-grams	1 trilijun	13 milijuna

Korpus

- Riječi se ne pojavljuju od nigdje!
- Tekst nastaje
 - kod određenog pisca (pisaca),
 - u određeno vrijeme,
 - u određenoj varijaciji,
 - na određenom jeziku,
 - zbog određene funkcije.

Korpusi variraju po dimenzijama

- **Jezik:** 7097 jezika na svijetu
- **Varijante:** čakavica, kajkavica, ...
- **Žanr:** novine, fikcija, znanstvi radovi, Wikipedia...
- **Demografija autora:** dob, spol, etična skupina...

Izgradnja korpusa

- **Motivacija**
 - Zašto je korpus napravljen?
 - Tko ga je napravio?
 - Tko je financirao izgradnju?
- **Situacija**
 - Radi čega je tekst napisan?
- **Proces skupljanja**
 - Ako je podsampliran, kako je podsampliran?
 - Je li bilo koncenzusa?
 - Predobrada?
- **Proces anotacije, varijante jezika, demografija**

Normalizacija teksta

- **Svaki zadatak obrade prirodnog jezika uključuje normalizaciju teksta:**
 1. Tokenizacija/segmentacija riječi u aktivnom tekstu
 2. Normalizacija formata riječi
 3. Segmentacija rečenica u aktivnom tekstu

Tokenizacija po praznom znaku

- Jednostavan način tokenizacije
 - za jezike koji koriste razmak između riječi arapski, grčki, cirilični, latinski... sustav pisanja
 - segmentiranje pojavnice između dva prazna znaka
- Unix alati za tokenizaciju po praznom znaku
 - "tr" naredba
 - za danu tekstualnu datoteku, ispiše pojavnice i njihove frekvencije

Jednostavna tokenizacija u UNIX-u

- Za danu tekstualnu datoteku vraća pojavnice i njihove frekvencije

```
tr -sc "A-Za-zŠĐĆĆŽšđććž0-9" "\n" < alan_ford.txt | sort | uniq -c
```

5 će
10 ćemo
1 ćete
4 ćeš
17 ću
1 Čeka
1 Čekaj
1 Čemu
10 Čini
1 Čitava
1 Čuj
1 Čujmo
2 Čuo
1 čahure
...

Zamjena svih
nealfanumeričkih
znakova s novim
redom

Sortiranje

Prebrojavanje
jedinstvenih

Jednostavna tokenizacija u UNIX-u

- Za danu tekstualnu datoteku vraća pojavnice i njihove frekvencije

```
tr -sc "A-Za-zŠĐĆĆŽšđććž0-9" "\n" < alan_ford.txt | sort | uniq -c | sort -n -r
```

109 je

101 da

97 se

55 sam

47 u

46 i

40 na

35 za

34 to

31 ne

27 li

22 A

20 mi

...

Sortiranje po
frekvenciji

Problemi tokenizacije

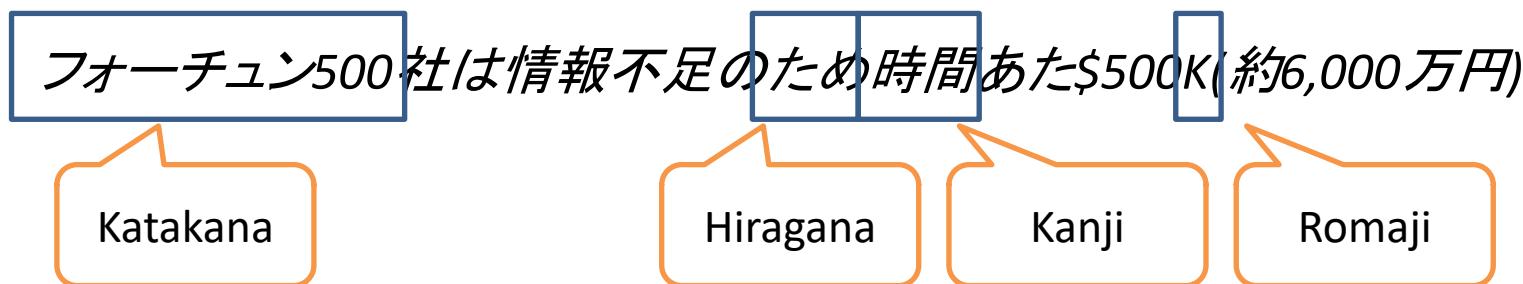
- Ne smiju se izbaciti svi interpunkcijski znakovi
 - km/s, dr. sc
 - cijene (\$59.99)
 - datumi (01.10.2021)
 - URL (<https://www.pmfst.hr>)
 - hashtag (#nlp)
 - email (netko@pmfst.hr)
- Klitike: riječi koje ne stoje same uz sebe
 - are u we're
- Kada višerječni izrazi postaju tokeni?
 - New York, bed & breakfast

Tokenizacija kod jezika bez razmaka

- Francuski
 - L'**ensemble** jedna pojavnica ili dvije?
 - L ? L' ? Le ?
 - Težnja je da se l'**ensemble** opojavniči kao **ensemble**
- Njemačke imeničke složenice nisu segmentirane
 - **Lebensversicherungsgesellschaftsangestellter**
 - 'zaposlenik tvrtke za životno osiguranje'
 - pronalaženje informacija (information retrieval) u Njemačkom zahtjeva *razdvajanje složenica*

Problemi kod jezika bez razmaka

- Kineski i japanski nemaju razmake između riječi
 - 伊万尼塞維奇现在居住在美国东南部的美國加州。
 - 伊万尼塞維奇 现在 居住 在 美国 东南部 的 美國加州
 - Ivanišević danas živi u US jugoistočnoj Kaliforniji
- U japanskom jeziku se pojavljuju riječi pisane drugim abecedama



- Korisnik može sve izraziti u Hiragana abecedi

Tokenizacija riječi u kineskom

- je zapravo segmentacija riječi (Word segmentation)
- Kineske riječi se tvore od znakova
 - znakovi se tvore najčešće od jednog sloga i jednog morfema
 - prosječna riječ je duga 2.4 znaka
- Standardni algoritam za segmentaciju:
 - Maksimalno podudaranje – pohlepni algoritam (Maximum matching – greedy)

Maksimalno podudaranje

Algoritam za segmentaciju riječi

- za danu listu riječi i za niz znakova
 1. stavi pokazivač na početak niza.
 2. pronađi najdulju riječ u rječniku koja odgovara niz s početkom u pokazivaču.
 3. pomaknite pokazivač preko riječi u nizu.
 4. idи на 2.

Maksimalno podudaranje

Thecatinthehat

the cat in the hat

Thetabledownthere

the table down there

theta bled own there

- Nije primjenjivo za engleski jezik
- ali odlično radi za kineski
 - 莎拉波娃现在居住在美国东南部的佛罗里达。
 - 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
- Moderni probabilistički segmentacijski algoritmi još i bolje

Uvod u obradu prirodnog jezika

2.3. Kodiranje uparivanjem byte-ova (Byte pair encoding)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Druga opcija za tokenizaciju

Umjesto

- segmentacije po praznom znaku
- segmentacija po jednom znaku

mogu se iskoristiti **podaci** da nam "kažu" kako tokenizirati

Segmentacija podriječi

(jer pojavnice mogu biti dio riječi, kao i sama riječ)

Tokenizacija podriječi

Tri uobičajena algoritma:

- Kodiranje parova byte-ova (Byte-Pair Encoding (BPE))
(Sennrich et al., 2016)
- Unigram language modeling tokenization (Kudo, 2018)
- WordPiece (Schuster and Nakajima, 2012)

Svi imaju dva dijela:

- Učenje tokena koji uzima korpus za treniranje i inducira rječnik
- Segmentiranje tokena koji uzima sirove testne rečenice i tokenizira ih po rječniku

BPE učenje tokena

U početku je rječnik skup individualnih znakova

$$= \{A, B, C, D, \dots, a, b, c, d, \dots\}$$

Ponavljam:

- izaberi dva najfrekventnija susjedna znaka u korpusu za treniranje (recimo 'A' i 'B')
- dodaj novi spojeni simbol 'AB' u rječnik
- zamijeni svaki susjedni 'A' 'B' u korpusu s 'AB' (spajanje) dok se ne napravi k spajanja.

BPE dodatak

Većina algoritama podriječi se izvode nad razmakom odvojenim tokenima

Stoga se prvo dodaje specijalni znak za kraj riječi '_' prije razmaka u korpusu za treniranje

Slijedi separacija u znakove

BPE dodatak

Većina algoritama podriječi se izvode nad razmakom odvojenim tokenima

Stoga se prvo dodaje specijalni znak za kraj riječi '_' prije razmaka u korpusu za treniranje

Slijedi separacija u znakove

BPE primer

low low low low lowest lowest newer newer newer
newer newer newer wider wider wider new new

BPE primjer

low low low low low lowest lowest newer newer newer
newer newer newer wider wider wider new new

korpus

5	l	o	w	_			
2	l	o	w	e	s	t	_
6	n	e	w	e	r	_	
3	w	i	d	e	r	_	
2	n	e	w	_			

rječnik

_	d	e	i	l	n	o	r	s	t	w
---	---	---	---	---	---	---	---	---	---	---

BPE primjer

spoji e r s er

korpus

5 l o w _
2 l o w e s t _
6 n e w e r _
3 w i d e r _
2 n e w _

rječnik

_ d e i l n o r s t w e r

spoji er _ u er _

korpus

5 l o w _
2 l o w e s t _
6 n e w e r _
3 w i d e r _
2 n e w _

rječnik

_ d e i l n o r s t w e r er _

BPE primjer

spoji n e u ne

korpus

5 l o w _
2 l o w e s t _
6 ne w er_
3 w i d er_
2 ne w _

rječnik

_ d e i l n o r s t w e r e r_ ne

BPE primjer

Sljedeća spajanja su

Spajanje	Rječnik
(ne, w)	_ d e i l n o r s t w e r e r_ ne new
(l, o)	_ d e i l n o r s t w e r e r_ ne new lo
(lo, w)	_ d e i l n o r s t w e r e r_ ne new lo low
(new, er_)	_ d e i l n o r s t w e r e r_ ne new lo low newer_
(low, _)	_ d e i l n o r s t w e r e r_ ne new lo low newer_ low_

BPE primjer

BPE segmentiranje

Na testnim podacima, pokreni svako spajanje naučeno nad trening podacima

- pohlepno
- u redoslijedu kako se učilo

Stoga, spoji svaki `er` u `er`, onda svaki `er_` u `er_`, itd.

Rezultat

- testni skup "n e w e r_" će se tokenizirati kao puna riječ
- testni skup "l o w e r_" će se tokenizirati kao "low er_"

BPE svojstva

Obično uključuje frekventne riječi
i frekventne podriječi

- koje su često morfemi kao –est ili –er

Morfem je najmanja smislena jedinica jezika

Uvod u obradu prirodnog jezika

2.4. Normalizacija i izvlačenje korijena riječi (Word normalization and stemming)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Normalizacija riječi

- postavljanje riječi/tokena u standardni format
 - U.S.A. ili USA
 - uhhuh ili uh-huh
 - Fed ili fed
 - sam, smo, ste

Promjena veličine slova (Case folding)

- Prilikom pronalaženja informacija često se velika slova prebacuju u mala
 - jer korisnici teže upotrebi malih slova
 - Mogući izuzeci: Veliko slovo u sredini rečenice?
 - npr., Srednja Dalmacija
 - CARNet
 - Za sentimentnu analizu, strojno učenje, ekstrakciju informacija
 - promjena veličine slova pomaže

Lematizacija

- Smanjivanje infleksija ili varijanta oblika riječi na osnovni oblik (lema)
 - leti, lete, letim, leteći → letjeti
 - ptica, ptice, pticama, ptici → ptica
- *One ptice su visoko letjele* → *Onaj ptica biti visok letjeti*
- Lematizacija: traženje ispravnog oblika glavne riječi u rječniku

Morfologija

- **Morfemi**
 - mali smisleni dijelovi riječi
 - **korijen riječi** (stem): temeljni dio
 - **afiksi**: dijelovi koji se dodaju korijenu riječi
 - često imaju gramatičke funkcije

Izvlačenje korijena riječi (stemming)

- Smanjivanje oblika riječi na njegov korijen u pronalaženju informacija
- Korijen riječi se dobiva grubim cijepanjem afiksa
 - ovisno o jeziku
 - npr. **automati**, **automatski**, **automatizacija** se svodi na **automat**

Splitska Kinoteka priredila je odlican program do kraja tjedna. Uz poznata filmska ostvarenja tu je i jedna manje razvikan filmska poslastica.



Splitsk Kinotek priredi je odlican progra do kraj tjedn Uz poznat filmsk ostvarenj tu je i jedn manj razvikan filmsk poslastic

Porterov algoritam

- Najčešći alat za izvlačenje korijena riječi u Engleskom jeziku

Korak 1a

sses → ss	caresses → caress
ies → i	ponies → poni
ss → ss	caress → caress
s → Ø	cats → cat

Korak 1b

(*v*)ing → Ø	walking → walk
	sing → sing
(*v*)ed → Ø	plastered → plaster
...	

Korak 2 (za duge korijene)

ational → ate	relational → relate
izer → ize	digitizer → digitize
ator → ate	operator → operate
...	

Korak 3 (za duže korijene)

al → Ø	revival → reviv
able → Ø	adjustable → adjust
ate → Ø	activate → activ
...	

Uvod u obradu prirodnog jezika

2.4. Segmentacija rečenice i stabla odluke (Sentence segmentation and decision trees)

Branko Žitko

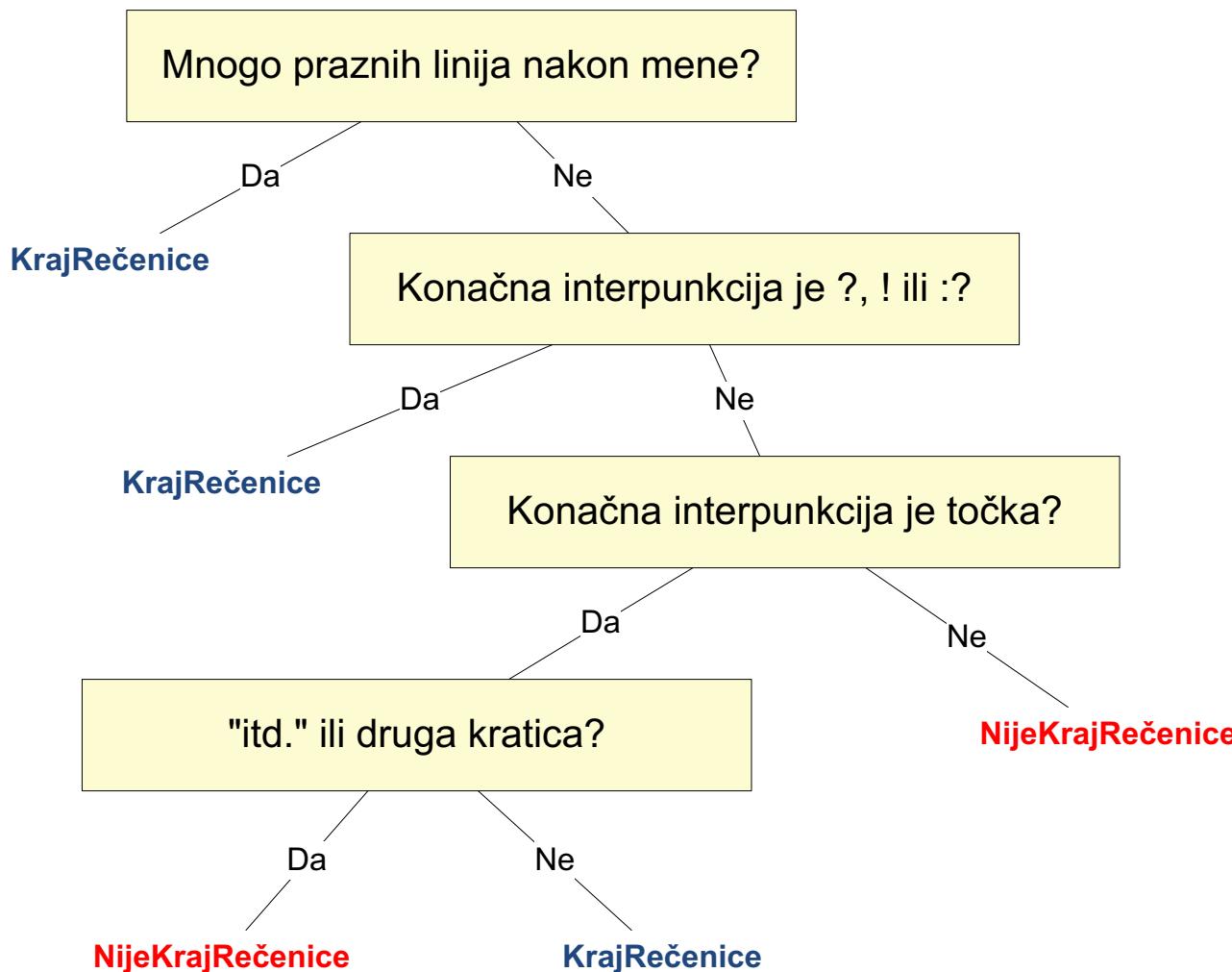
prevedeno od: Dan Jurafsky, Chris Manning

Segmentacija rečenice

- !, ? – uglavnom jednoznačni
- "." – više značna i može označavati
 - kraj rečenice
 - kratice poput dr. itd.
 - brojeve poput .02% 4.3
- Sagraditi binarni klasifikator
 - koji traži "."
 - odlučuje jeli KrajRečenice/NijeKrajRečenice
 - klasifikatori: ručno pisana pravila, regularni izrazi ili strojno učenje

Stablo odluke

- Odluka je li pojavnica predstavlja kraj rečenice.



Profinjenje stabla odluke

- riječi s točkom:
 - mala slova, velika slova, prvo veliko slovo, broj
- riječi nakon točke:
 - mala slova, velika slova, prvo veliko slovo, broj
- Numeričke osobine:
 - duljina riječi s točkom
 - vjerojatnost (rijec s točkom se pojavljuje na kraju rečenice)
 - vjerojatnost (rijec nakon točke se pojavljuje na početku rečenice)

Uvod u obradu prirodnog jezika

3.1. Minimalna udaljenost dva niza znakova (Minimum edit distance)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Koliko su slična dva niza znakova?

- Ispravak pravopisnih grešaka
 - Korisnik je unio
 - "žrafa"
 - Što bi bila ispravka
 - grafa
 - rafa
 - žirafa
- Računalna biologija
 - poravnaj nizove nukleotida

```
AGGCTATCACCTGACCTCCAGGCCGATGCC  
TAGCTATCACGACCAGCGGGTCGATTGCCGAC
```
 - poravnanje

```
-AGGCTATCACCTGACCTCCAAGGCCGA--TGCCC---  
TAG-CTATCAC--GACCAGC--GGTCGA TT TGCCC GAC
```
- Koristi se kod strojnog prevodenja, ekstrakcije informacija, prepoznavanja govora...

Udaljenost nizova znakova

- Minimalna udaljenost između dva niza znakova
- je minimalni broj operacija
 - ubacivanja
 - brisanja
 - zamjene
- potrebnih da se jedan niz znakova transformira u drugi

Rezultat povratnog praćenja

- Dva niza znakova i njihovo poravnanje

T	R	A	D	*	I	C	I	J	A
*	D	E	D	U	K	C	I	J	A

Minimalna udaljenost nizova znakova

- Dva niza znakova i njihovo poravnanje
 - b – brisanje
 - z – zamjena
 - u – ubacivanje

T	R	A	D	*	I	C	I	J	A
*	D	E	D	U	K	C	I	J	A
b	z	z		u	z				

- Ako svaka operacija vrijedi 1 bod
 - onda je udaljenost 5
- ako zamjena iznosi 2 boda (Levenshtein)
 - onda je udaljenost 8

Poravnanje u računalnoj biologiji

- Za dani niz dušikovih baza

AGGCTATCACCTGACCTCCAGGCCGATGCC

TAGCTATCACGACC CGGGT CGATTGCCCGAC

- Poravnanje

-AGGCTATCACCTGACCTCCA GGCGA -- TGCCC ---

TAG-CTATCAC -- GACC GC -- GGT CGA TT TGCCCC GAC

- Za dva dana niza, poravnaj svaki znak u drugi znak ili razmak

Druge upotrebe udaljenosti nizova znakova

- Evaluacija strojnog prevodenja i prepoznavanja govora

Govornik je potvrdio da je utakmica počela.

Govornik kaže kako je utakmica počela maloprije.

B

Z

Z

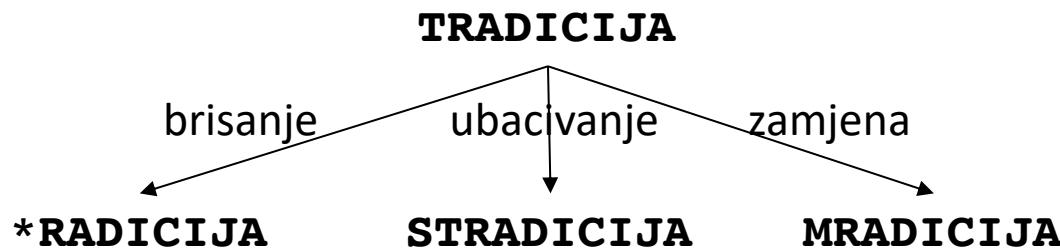
U

- Ekstrakcija imenovanih entiteta i njihove koreference

- Podravka d.o.o. je najavila danas
- Podravka je profitirala
- Predsjednik Podravke Ante Antić je jučer najavio
- za predsjednika uprave Podravke Antu Antića

Kako odrediti minimalnu udaljenost?

- Potraga za putanjom (nizom operacija) koja transformira početnu riječ u krajnju riječ
 - inicijalno stanje: riječ koja se transformira
 - operacije: ubaci, izbriši, zamjeni
 - ciljno stanje: riječ u koju se transformira
 - vrijednost putanje: ono što želimo minimizirati: broj operacija



Minimalna udaljenost kao traženje putanje

- Prostor svih putanja je ogroman!
 - ne možemo naivno pristupiti problemu pretrage
 - mnogo različitih putanja se generira od nekog stanja
 - ne moramo pratiti svaku od njih
 - samo najkraću putanju iz svih ponovno posjećenih stanja

Definicija minimalne udaljenosti

- Za dva niza znakova
 - S veličine m
 - T veličine n
- Definiramo $D[i, j]$
 - udaljenost između $S[1 \dots i]$ i $T[1 \dots j]$
 - npr. prvih i znakova od S i prvih j znakova od T
 - udaljenost između S i T je onda $D[m, n]$

Uvod u obradu prirodnog jezika

3.2. Izračun minimalne udaljenosti (Computing minimum edit distance)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Dinamičko programiranje za minimalnu udaljenost

- **Dinamičko programiranje:** tablični izračun od $D[m, n]$
- Rješavanje problema kombiniranjem rješenja podproblema
- Pristup odozdo prema dolje (Bottom-up)
 - Izračuna se $D[i, j]$ za male i, j
 - Izračuna se veliki $D[i, j]$ temeljem prethodno izračunatih manjih vrijednosti
 - npr. izračuna se $D[i, j]$ za sve
 - $i (0 < i < m)$ i
 - $j (0 < j < n)$

Definicija minimalne udaljenosti (Levensthein)

- Inicijalizacija:

$$D[i, 0] = i$$

$$D[0, j] = j$$

- Relacija povratka:

za svaki $i = 1 \dots m$

za svaki $j = 1 \dots n$

$$D[i, j] = \min \begin{cases} D[i - 1, j] + 1 \\ D[i, j - 1] + 1 \\ D[i - 1, j - 1] + \begin{cases} 2: \text{ako } S[i] \neq T[j] \\ 0: \text{ako } S[i] = T[j] \end{cases} \end{cases}$$

- Zaustavljanje:

$D[m, n]$ je udaljenost

Tablica minimalne udaljenosti

$$D[i,j] = \min \begin{cases} D[i-1,j] + 1 \\ D[i,j-1] + 1 \\ D[i-1,j-1] + \begin{cases} 2 & \text{ako } S[i] \neq T[j] \\ 0 & \text{ako } S[i] = T[j] \end{cases} \end{cases}$$

Tablica minimalne udaljenosti

$$D[i, j] = \min \begin{cases} D[i - 1, j] + 1 \\ D[i, j - 1] + 1 \\ D[i - 1, j - 1] + \begin{cases} 2 & \text{ako } S[i] \neq T[j] \\ 0 & \text{ako } S[i] = T[j] \end{cases} \end{cases}$$

S \ T	#	D	E	D	U	K	C	I	J	A
#	0	1	2	3	4	5	6	7	8	9
T	1	2	3	4	5	6	7	8	9	10
R	2	3	4	5	6	7	8	9	10	11
A	3	4	5	6	7	8	9	10	11	10
D	4	3	4	5	6	7	8	9	10	11
I	5	4	5	6	7	8	9	8	9	10
C	6	5	6	7	8	9	8	9	10	11
I	7	6	7	8	9	10	9	8	9	10
J	8	7	8	9	10	11	10	9	8	9
A	9	8	9	10	11	12	11	10	9	8

Uvod u obradu prirodnog jezika

3.3. Povratno praćenje za izračun poravnavanja (Backtrace in computing alignment)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Izračun poravnavanja

- Minimalna udaljenost nije dovoljna
- Često je potrebno **uskladiti** svaki znak od dva niza znakova
- Ovo se vrši korištenjem **povratnim praćenjem**
- Svakim ulaskom u čeliju tablice trebamo znati odakle smo došli
- Kada dođemo do kraja
 - pratimo trag od gornjeg desnog kuta kako bi pročitali poravnanje

Dodavanje povratnog praćenja

- Inicijalizacija:

$$D[i, 0] = i$$

$$D[0, j] = j$$

- Relacija povratka:

za svaki $i = 1 \dots m$

za svaki $j = 1 \dots n$

$$D[i, j] = \min \begin{cases} D[i - 1, j] + 1 & \text{brisanje} \\ D[i, j - 1] + 1 & \text{ubacivanje} \\ D[i - 1, j - 1] + \begin{cases} 2: \text{ako } S[i] \neq T[j] \\ 0: \text{ako } S[i] = T[j] \end{cases} & \text{zamjena} \end{cases}$$

$$Ptr[i, j] = \begin{cases} \uparrow & \text{brisanje} \\ \leftarrow & \text{ubacivanje} \\ \nwarrow & \text{zamjena} \end{cases}$$

- Zaustavljanje:

$D[m, n]$ je udaljenost

Tablica minimalne udaljenosti

$$D[i, j] = \min \begin{cases} D[i - 1, j] + 1 & \text{brisanje} \\ D[i, j - 1] + 1 & \text{ubacivanje} \\ D[i - 1, j - 1] + \begin{cases} 2: \text{ako } S[i] \neq T[j] \\ 0: \text{ako } S[i] = T[j] \end{cases} & \text{zamjena} \end{cases}$$

S \ T	#	D	E	D	U	K	C	I	J	A
#	0	1	2	3	4	5	6	7	8	9
T	1	2	3	4	5	6	7	8	9	10
R	2	3	4	5	6	7	8	9	10	11
A	3	4	5	6	7	8	9	10	11	10
D	4	3	4	5	6	7	8	9	10	11
I	5	4	5	6	7	8	9	8	9	10
C	6	5	6	7	8	9	8	9	10	11
I	7	6	7	8	9	10	9	8	9	10
J	8	7	8	9	10	11	10	9	8	9
A	9	8	9	10	11	12	11	10	9	8

Tablica minimalne udaljenosti

$$Ptr[i.j] = \begin{cases} \uparrow & \text{brisanje} \\ \leftarrow & \text{ubacivanje} \\ \nwarrow & \text{zamjena} \end{cases}$$

S\T	#	D	E	D	U	K	C	I	J	A
#	0	\leftarrow 1	\leftarrow 2	\leftarrow 3	\leftarrow 4	\leftarrow 5	\leftarrow 6	\leftarrow 7	\leftarrow 8	\leftarrow 9
T	↑ 1	$\nwarrow \uparrow$ \leftarrow 2	$\nwarrow \uparrow$ \leftarrow 3	$\nwarrow \uparrow$ \leftarrow 4	$\nwarrow \uparrow$ \leftarrow 5	$\nwarrow \uparrow$ \leftarrow 6	$\nwarrow \uparrow$ \leftarrow 7	$\nwarrow \uparrow$ \leftarrow 8	$\nwarrow \uparrow$ \leftarrow 9	$\nwarrow \uparrow$ \leftarrow 10
R	↑ 2	$\nwarrow \uparrow$ \leftarrow 3	$\nwarrow \uparrow$ \leftarrow 4	$\nwarrow \uparrow$ \leftarrow 5	$\nwarrow \uparrow$ \leftarrow 6	$\nwarrow \uparrow$ \leftarrow 7	$\nwarrow \uparrow$ \leftarrow 8	$\nwarrow \uparrow$ \leftarrow 9	$\nwarrow \uparrow$ \leftarrow 10	$\nwarrow \uparrow$ \leftarrow 11
A	↑ 3	$\nwarrow \uparrow$ \leftarrow 4	$\nwarrow \uparrow$ \leftarrow 5	$\nwarrow \uparrow$ \leftarrow 6	$\nwarrow \uparrow$ \leftarrow 7	$\nwarrow \uparrow$ \leftarrow 8	$\nwarrow \uparrow$ \leftarrow 9	$\nwarrow \uparrow$ \leftarrow 10	$\nwarrow \uparrow$ \leftarrow 11	\nwarrow 10
D	↑ 4	\nwarrow 3	\nwarrow 4	\nwarrow 5	\nwarrow 6	\nwarrow 7	\nwarrow 8	\nwarrow 9	\nwarrow 10	\uparrow 11
I	↑ 5	\uparrow 4	$\nwarrow \uparrow$ \leftarrow 5	$\nwarrow \uparrow$ \leftarrow 6	$\nwarrow \uparrow$ \leftarrow 7	$\nwarrow \uparrow$ \leftarrow 8	$\nwarrow \uparrow$ \leftarrow 9	\nwarrow 8	\leftarrow 9	\leftarrow 10
C	↑ 6	\uparrow 5	$\nwarrow \uparrow$ \leftarrow 6	$\nwarrow \uparrow$ \leftarrow 7	$\nwarrow \uparrow$ \leftarrow 8	$\nwarrow \uparrow$ \leftarrow 9	\nwarrow 8	\uparrow 9	$\nwarrow \uparrow$ \leftarrow 10	$\nwarrow \uparrow$ \leftarrow 11
I	↑ 7	\uparrow 6	$\nwarrow \uparrow$ \leftarrow 7	$\nwarrow \uparrow$ \leftarrow 8	$\nwarrow \uparrow$ \leftarrow 9	$\nwarrow \uparrow$ \leftarrow 10	\uparrow 9	\nwarrow 8	\leftarrow 9	\leftarrow 10
J	↑ 8	\uparrow 7	$\nwarrow \uparrow$ \leftarrow 8	$\nwarrow \uparrow$ \leftarrow 9	$\nwarrow \uparrow$ \leftarrow 10	$\nwarrow \uparrow$ \leftarrow 11	\uparrow 10	\uparrow 9	\nwarrow 8	\leftarrow 9
A	↑ 9	\uparrow 8	$\nwarrow \uparrow$ \leftarrow 9	$\nwarrow \uparrow$ \leftarrow 10	$\nwarrow \uparrow$ \leftarrow 11	$\nwarrow \uparrow$ \leftarrow 12	\uparrow 11	\uparrow 10	\uparrow 9	\nwarrow 8

Tablica minimalne udaljenosti

TRAD*ICIJA
 *DEDUKCIJA
 bzz uz

S\T	#	D	E	D	U	K	C	I	J	A
#	0	← 1	← 2	← 3	← 4	← 5	← 6	← 7	← 8	← 9
T	↑ 1	↖↑ ← 2	↖↑ ← 3	↖↑ ← 4	↖↑ ← 5	↖↑ ← 6	↖↑ ← 7	↖↑ ← 8	↖↑ ← 9	↖↑ ← 10
R	↑ 2	↖↑ ↖ 3	↖↑ ← 4	↖↑ ← 5	↖↑ ← 6	↖↑ ← 7	↖↑ ← 8	↖↑ ← 9	↖↑ ← 10	↖↑ ← 11
A	↑ 3	↖↑ ← 4	↖↑ ↖ 5	↖↑ ← 6	↖↑ ← 7	↖↑ ← 8	↖↑ ← 9	↖↑ ← 10	↖↑ ← 11	↖ 10
D	↑ 4	↖ 3	↖ 4	↖ 5	↖ 6	↖ 7	↖ 8	↖ 9	↖ 10	↖ 11
I	↑ 5	↑ 4	↖↑ ← 5	↖↑ ← 6	↖↑ ← 7	↖↑ ↖ 8	↖↑ ← 9	↖ 8	↖ 9	↖ 10
C	↑ 6	↑ 5	↖↑ ← 6	↖↑ ← 7	↖↑ ← 8	↖↑ ← 9	↖ 8	↑ 9	↖↑ ← 10	↖↑ ← 11
I	↑ 7	↑ 6	↖↑ ↖ 7	↖↑ ← 8	↖↑ ← 9	↖↑ ← 10	↑ 9	↖ 8	↖ 9	↖ 10
J	↑ 8	↑ 7	↖↑ ← 8	↖↑ ← 9	↖↑ ← 10	↖↑ ← 11	↑ 10	↑ 9	↖ 8	↖ 9
A	↑ 9	↑ 8	↖↑ ↖ 9	↖↑ ← 10	↖↑ ← 11	↖↑ ← 12	↑ 11	↑ 10	↑ 9	↖ 8

Tablica minimalne udaljenosti

TRADICIJA**
***DED*UKCIJA**
bzz buu

S\T	#	D	E	D	U	K	C	I	J	A
#	0	← 1	← 2	← 3	← 4	← 5	← 6	← 7	← 8	← 9
T	↑ 1	↖↑ ← 2	↖↑ ← 3	↖↑ ← 4	↖↑ ← 5	↖↑ ← 6	↖↑ ← 7	↖↑ ← 8	↖↑ ← 9	↖↑ ← 10
R	↑ 2	↖↑ ↖ 3	↖↑ ← 4	↖↑ ← 5	↖↑ ← 6	↖↑ ← 7	↖↑ ← 8	↖↑ ← 9	↖↑ ← 10	↖↑ ← 11
A	↑ 3	↖↑ ← 4	↖↑ ↖ 5	↖↑ ← 6	↖↑ ← 7	↖↑ ← 8	↖↑ ← 9	↖↑ ← 10	↖↑ ← 11	↖ 10
D	↑ 4	↖ 3	↖ 4	↖ 5	↖ 6	↖ 7	↖ 8	↖ 9	↖ 10	↖ 11
I	↑ 5	↑ 4	↖↑ ← 5	↖↑ ← 6	↖↑ ↖ 7	↖↑ ↖ 8	↖↑ ← 9	↖ 8	↖ 9	↖ 10
C	↑ 6	↑ 5	↖↑ ← 6	↖↑ ← 7	↖↑ ← 8	↖↑ ← 9	↖ 8	↑ 9	↖↑ ← 10	↖↑ ← 11
I	↑ 7	↑ 6	↖↑ ↖ 7	↖↑ ← 8	↖↑ ← 9	↖↑ ← 10	↑ 9	↖ 8	↖ 9	↖ 10
J	↑ 8	↑ 7	↖↑ ← 8	↖↑ ← 9	↖↑ ← 10	↖↑ ← 11	↑ 10	↑ 9	↖ 8	↖ 9
A	↑ 9	↑ 8	↖↑ ↖ 9	↖↑ ← 10	↖↑ ← 11	↖↑ ← 12	↑ 11	↑ 10	↑ 9	↖ 8

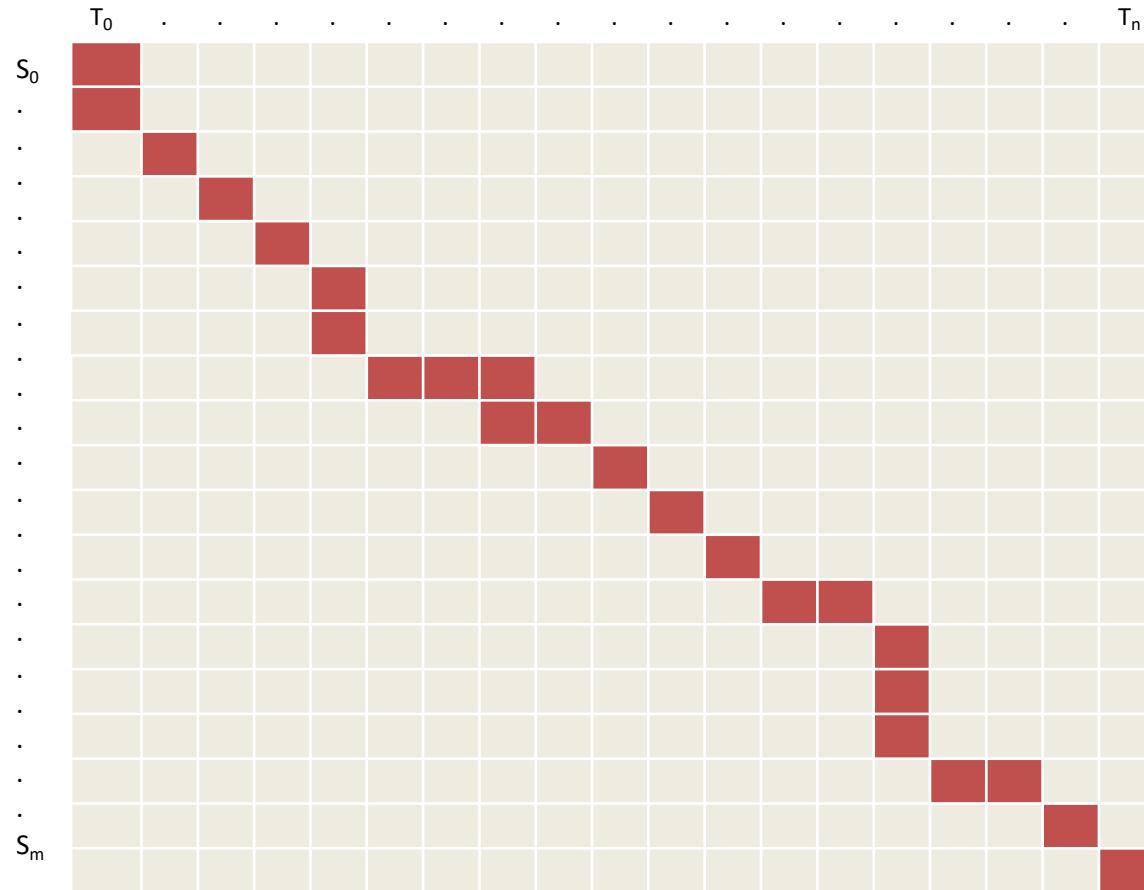
Tablica minimalne udaljenosti

TRADI**CIJA**
*****D*EDUKCIJA**
bbb buuuu

S\T	#	D	E	D	U	K	C	I	J	A
#	0	← 1	← 2	← 3	← 4	← 5	← 6	← 7	← 8	← 9
T	↑ 1	↖↑ ← 2	↖↑ ← 3	↖↑ ← 4	↖↑ ← 5	↖↑ ← 6	↖↑ ← 7	↖↑ ← 8	↖↑ ← 9	↖↑ ← 10
R	↑ 2	↖↑ ← 3	↖↑ ← 4	↖↑ ← 5	↖↑ ← 6	↖↑ ← 7	↖↑ ← 8	↖↑ ← 9	↖↑ ← 10	↖↑ ← 11
A	↑ 3	↖↑ ← 4	↖↑ ← 5	↖↑ ← 6	↖↑ ← 7	↖↑ ← 8	↖↑ ← 9	↖↑ ← 10	↖↑ ← 11	↖ 10
D	↑ 4	↖ 3	↖ 4	↖ 5	↖ 6	↖ 7	↖ 8	↖ 9	↖ 10	↑ 11
I	↑ 5	↑ 4	↖↑ ← 5	↖↑ ← 6	↖↑ ← 7	↖↑ ← 8	↖↑ ← 9	↖ 8	↖ 9	↖ 10
C	↑ 6	↑ 5	↖↑ ← 6	↖↑ ← 7	↖↑ ← 8	↖↑ ← 9	↖ 8	↑ 9	↖↑ ← 10	↖↑ ← 11
I	↑ 7	↑ 6	↖↑ ← 7	↖↑ ← 8	↖↑ ← 9	↖↑ ← 10	↑ 9	↖ 8	↖ 9	↖ 10
J	↑ 8	↑ 7	↖↑ ← 8	↖↑ ← 9	↖↑ ← 10	↖↑ ← 11	↑ 10	↑ 9	↖ 8	↖ 9
A	↑ 9	↑ 8	↖↑ ← 9	↖↑ ← 10	↖↑ ← 11	↖↑ ← 12	↑ 11	↑ 10	↑ 9	↖ 8

Matrica udaljenosti

- Svaka ne opadajuća putanja od (m,n) do $(0,0)$ odgovara poravnanju dvaju niza znakova
- Optimalno poravnanje se sastoji od optimalnih podporavnjanja



Minimalna udaljenost nizova znakova

- Dva niza znakova i njihovo poravnanje

T	R	A	D	*	I	C	I	J	A
*	D	E	D	U	K	C	I	J	A

Performanse

- vrijeme $O(mn)$
- prostor $O(mn)$
- povratno praćenje $O(m+n)$

Uvod u obradu prirodnog jezika

3.4. Težinska minimalna udaljenost (Weighted edit distance)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

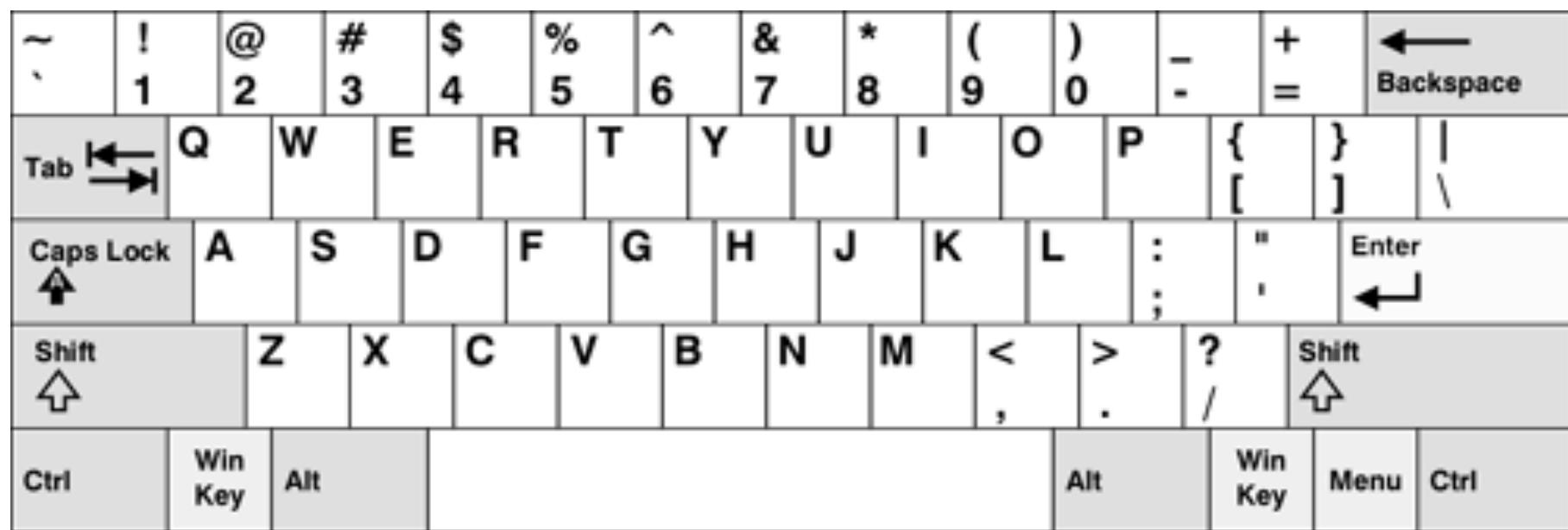
Zašto dodati težinski faktor?

- Ispravak pravopisnih grešaka
 - neka slova imaju veću vjerojatnost krivog unosa u odnosu na druga
- Biologija
 - određena brisanja i umetanja su vjerojatnija od drugih

Matrica konfuzije za pravopisne pogreške

zamjena[X,Y] = supstitucija od X(pogrešno) s Y(točno)

	Y(točno)																									
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
X	0	0	7	1342	0	0	2118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0		
a	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	14	0	1	0	18	
f	0	15	0	3	1	0	5	2	0	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	
p	0	11	1	2	0	6	5	0	2	0	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	3	0	



Težinska minimalna udaljenost

- Inicijalizacija:

$$D[0,0] = 0$$

$$D[i, 0] = D[i, 0] + \text{brisanje}(S[i]); \quad 1 \leq i \leq m$$

$$D[0, j] = D[0, j] + \text{ubacivanje}(T[j]); \quad 1 \leq j \leq n$$

- Relacija povratka:

za svaki $i = 1 \dots m$

za svaki $j = 1 \dots n$

$$D[i, j] = \min \begin{cases} D[i - 1, j] + \text{brisanje}(S[i]) \\ D[i, j - 1] + \text{ubacivanje}(T[j]) \\ D[i - 1, j - 1] + \text{zamjena}(S[i], T[j]) \end{cases}$$

- Zaustavljanje:

$D[m, n]$ je udaljenost

Uvod u obradu prirodnog jezika

4.1. Uvod u n-grame

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Probabilistički modeli jezika

- Cilj: pridružiti vjerojatnost rečenici

- Strojno prevodenje

$P(\text{snažan vjetar večeras}) = P(\text{veliki vjetar večeras})$

- Ispravljanje pravopisnih grešaka

Ured je oko petnaest **minueta** od moje kuće

$P(\text{oko petnaest minuta od}) > P(\text{oko petnaest minueta od})$

- Prepoznavanje govora

$P(\text{udio sam Ivanu}) \gg P(\text{idi osam Ivan u})$

- Sumarizacija, odgovaranje na pitanja, itd. itd.

Probabilističko modeliranje jezika

- Cilj: izračunati vjerojatnost rečenice ili niza riječi

$$P(W) = P(w_1, w_2, \dots, w_n)$$

- Slični zadaci: vjerojatnost sljedeće riječi

$$P(w_5 | w_1, w_2, w_3, w_4)$$

- Model koji izračunava

$P(W)$ ili $P(w_n | w_1, w_2, \dots, w_{n-1})$ se zove **model jezika**

- Bolji naziv bi bila **gramatika**, ali **model jezika MJ** je standard

Kako izračunati $P(W)$

- Kako izračunati ovu združenu vjerojatnost

$P(\text{njegova, voda, je, tako, čista, da})$

- Osloniti se na pravilo **lanca kod vjerojatnosti**

Pravilo lanca

- Definicija uvjetne vjerojatnosti

$$P(A|B) = \frac{P(A,B)}{P(B)} \text{ ili}$$

$$P(A|B)P(B) = P(A, B) \text{ ili}$$

$$P(A, B) = P(A|B)P(B)$$

- Više varijabli:

$$P(A, B, C, D) = P(A)P(B|A)P(C|A, B)P(D|A, B, C)$$

- Opće pravilo lanca

$$\begin{aligned} P(x_1 x_2 \dots x_n) &= P(x_1)P(x_2|x_1)P(x_3|x_1 x_2) \dots P(x_n|x_1 x_2 \dots x_{n-1}) \\ &= P(x_1)P(x_2|x_1)P(x_3|x_1^2) \dots P(x_n|x_1^{n-1}) \end{aligned}$$

Primjena lanca vjerojatnosti

$$P(w_1 w_2 \dots w_n) = P(w_1^n) = \prod_{k=1}^n P(w_k | w_1 \dots w_{k-1}) = \prod_{k=1}^n P(w_k | w_1^{k-1})$$

$P(\text{njegova voda je tako čista jer}) =$

$P(\text{njegova}) \times$

$P(\text{voda} | \text{njegova}) \times$

$P(\text{je} | \text{njegova voda}) \times$

$P(\text{tako} | \text{njegova voda je}) \times$

$P(\text{čista} | \text{njegova voda je tako}) \times$

$P(\text{jer} | \text{njegova voda je tako čista})$

Kako procijeniti vjerojatnosti

- Možemo li samo prebrojiti i podijeliti?

$$P(\text{je} | \text{njesto je tako čista jer}) = \frac{c(\text{njesto je tako čista jer})}{c(\text{njesto je tako čista jer})}$$

- Ne možemo! Jer ima previše mogućih rečenica, ali i previše rečenica koje se ne pojavljuju.

Markovljeva pretpostavka

- pojednostavljenje pretpostavke

$$P(\text{je|njegova voda je tako čista jer}) \approx P(\text{je|jer})$$

- ili možda

$$P(\text{je|njegova voda je tako čista jer}) \approx P(\text{je|čista jer})$$

Markovljeva pretpostavka

$$P(w_1 \dots w_n) \approx \prod_i P(w_i | w_{i-k}^{i-1})$$

- drugim riječima, aproksimiramo svaku vrijednost u produktu

$$P(w_i | w_1^{i-1}) \approx P(w_i | w_{i-k}^{i-1})$$

Najjednostavniji slučaj: Unigram

$$P(w_1 \dots w_n) \approx \prod_i P(w_i)$$

- neke generirane rečenice iz unigram modela

Hmm Sviđa pred čitava vama

Dakle vodom 315 momče pothvat lopova posljednji nisu manje

Da pucao zapijevajte koga

Šesta dobiti golubarnik ostane

Zapamti slobode

Moj kljuć nije izvjesne duboka odvedite

Nesretniče organizaciju mikrofilmom uspjeti zajedničkog ispričam

Otkuda igle kotač znam opasnost tanjurima

Vidiš stoji aviona ostao

Čuj evo kontakt ubijati

Reci ubijen čitavu mušterija sreće

bigram model

- Uvjet prethodne riječi

$$P(w_1 \dots w_n) \approx \prod_i P(w_i | w_{i-1})$$

- neke generirane rečenice iz bigram modela

Trebalo je vjerojatno smaragd od hizmara.

Jedi rižu i mikrofilm s tanjurima.

Nadam se agencija Ford reklamni crtež.

Ostatak hoću natrag.

Tako se okrenuti sklopiti ću na električnu stolicu.

Imaš li drugih mana.

Ulovio sam sretan da ga šefe ja putujem u posljednji čas Miss.

Jesi li vi bolji način silaženja s x zrakama.

Oni tipovi.

Lisičine nisu doprli ispušni plinovi.

n-gram modeli

- Možemo proširiti na trigram, 4-gram, 5-gram...
- Općenito, svi ovi modeli su nedovoljni
 - jer jezik ima **daleke ovisnosti**

"Računalo koje sam stavio u dnevnu sobu na petom katu se srušilo"

- Ali često se možemo zadovoljiti s n-gram modelom

Uvod u obradu prirodnog jezika

4.2. Procjena vjerojatnosti N-grama

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Procjena vjerojatnosti bigrama

- Procjena maksimalne izglednosti
(MLE maximum likelihood estimation)

$$P(w_i | w_{i-1}) = \frac{C(w_{i-1} w_i)}{\sum_{\textcolor{blue}{w}} C(w_{i-1} \textcolor{blue}{w})}$$

$$P(w_i | w_{i-1}) = \frac{C(w_{i-1} w_i)}{C(w_{i-1})}$$

Primjer

- Procjena maksimalne izglednosti

$$P(w_i | w_{i-1}) = \frac{C(w_{i-1} w_i)}{C(w_{i-1})}$$

< s > ja sam Ivan </ s >

< s > Ivan ja sam </ s >

< s > ja ne volim kuhana jaja i šunku </ s >

$$P(\text{ja} | \langle s \rangle) = \frac{2}{3} = 0.67 \quad P(\text{Ivan} | s) = \frac{1}{3} = 0.33 \quad P(\text{sam} | \text{ja}) = \frac{2}{3} = 0.67$$

$$P(\langle /s \rangle | \text{Ivan}) = \frac{1}{2} = 0.5 \quad P(\text{Ivan} | \text{sam}) = \frac{1}{2} = 0.55 \quad P(\text{ne} | \text{ja}) = \frac{1}{3} = 0.33$$

Primjer

< s > ja sam Ivan < /s >

< s > Ivan ja sam < /s >

< s > ja ne volim jesti < /s >

< s >	ja	sam	Ivan	< /s >	ne	volim	jesti
3	3	2	2	3	1	1	1

	< s >	ja	sam	Ivan	< /s >	ne	volim	jesti
< s >		2		1				
ja			2			1		
sam				1	1			
Ivan		1			1			
< /s >								
ne							1	
volim								1
jesti					1			

Primjer

< s > ja sam Ivan < /s >

< s > Ivan ja sam < /s >

< s > ja ne volim jesti < /s >

< s >	ja	sam	Ivan	< /s >	ne	volim	jesti
3/16	3/16	2/16	2/16	3/16	1/16	1/16	1/16

	< s >	ja	sam	Ivan	< /s >	ne	volim	jesti
< s >		2/3		1/3				
ja			2/3			1/3		
sam				1/2	1/2			
Ivan		1/2			1/2			
< /s >								
ne							1/1	
volim								1/1
jesti					1/1			

Još primjera

- neke rečenice iz Berkeley Restaurant Project (BeRP)
 - can you tell me about any good cantonese restaurants close by
 - mid priced thai food is what i'm looking for
 - tell me about chez panisse
 - can you give me a listing of the kinds of food that are available
 - i'm looking for a good place to eat breakfast
 - when is caffe venezia open during the day

Broj bigrama

- Od 9222 rečenica ($V = 1446$)

	ja	želim	da	jedem	kinesku	hranu	ručak	potrošim
ja	5	827	0	9	0	0	0	2
želim	2	0	608	1	6	6	5	1
da	2	0	4	686	2	0	6	211
jedem	0	0	2	0	16	2	42	0
kinesku	1	0	0	0	0	82	1	0
hranu	15	0	15	0	1	4	0	0
ručak	2	0	0	0	0	1	0	0
potrošim	1	0	1	0	0	0	0	0

Bigram vjerojatnosti

- Normalizacija po unigramu

ja	želim	da	jedem	kinesku	hranu	ručak	potošim
2533	927	2417	746	158	1093	341	278

- Rezultat

	ja	želim	da	jedem	kinesku	hranu	ručak	potošim
ja	0,001974	0,32649	0	0,003553	0	0	0	0,00079
želim	0,002157	0	0,655879	0,001079	0,006472	0,006472	0,005394	0,001079
da	0,000827	0	0,001655	0,283823	0,000827	0	0,002482	0,087298
jedem	0	0	0,002681	0	0,021448	0,002681	0,0563	0
kinesku	0,006329	0	0	0	0	0,518987	0,006329	0
hranu	0,013724	0	0,013724	0	0,000915	0,00366	0	0
ručak	0,005865	0	0	0	0	0,002933	0	0
potošim	0,003597	0	0,003597	0	0	0	0	0

bigram procjene vjerojatnosti rečenice

$$P(<\text{s}> \text{Ja želim domaću hranu} </\text{s}>) =$$

$$P(\text{Ja} | <\text{s}>)$$

$$\times P(\text{želim} | \text{Ja})$$

$$\times P(\text{domaću} | \text{želim})$$

$$\times P(\text{hranu} | \text{domaću})$$

$$\times P(</\text{s}> | \text{hranu})$$

$$= 0.000031$$

Koje vrste znanja?

$P(\text{domaću} | \text{želim}) = 0.0011$

$P(\text{kinesku} | \text{želim}) = 0.0065$

$P(\text{da} | \text{želim}) = 0.66$

$P(\text{jedem} | \text{da}) = 0.28$

$P(\text{hranu} | \text{da}) = 0$

$P(\text{želim} | \text{potrošim}) = 0$

$P(\text{ja} | <\text{s}>) = 0.25$

Praktični problemi

- Sve se radi u logaritamskom prostoru
 - izbjegavanje prelijeva ispod granice realnih brojeva
 - zbrajanje je brže nego množenje

$$p_1 \times p_2 \times p_3 \times p_4 = \exp(\log(p_1) + \log(p_2) + \log(p_3) + \log(p_4))$$

Uvod u obradu prirodnog jezika

4.3. Evaluacija i perpleksija

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Evaluacija: koliko je dobar naš model?

- Da li naš model jezika više preferira dobre rečenice ili one loše?
 - pridruživanje veće vjerojatnosti "realnim" ili "često promatranim" rečenicama
 - nego "ne gramatičkim" ili " rijetko promatranim" rečenicama
- Treniramo model na **trening skupu**
- Testiramo performanse nad neviđenim podacima
 - **Testni skup** je neviđen skup podataka različit od trening skupa
 - **Evaluacijska metrika** govori koliko je dobar model nad testnim skupom

Vanjska evaluacija n-gram modela

- Najbolja evaluacija za usporedbu modela A i B
 - Staviti svaki model da radi
 - ispravak pravopisnih grešaka,
 - prepoznavanje govora,
 - strojno prevodenje
 - Prikupiti rezultate preciznosti modela A i B
 - koliko je pogrešnih riječi ispravljeno u točne riječi
 - koliko je prepoznatih riječi
 - koliko riječi je točno prevedeno
 - Usporediti preciznost od modela A i B

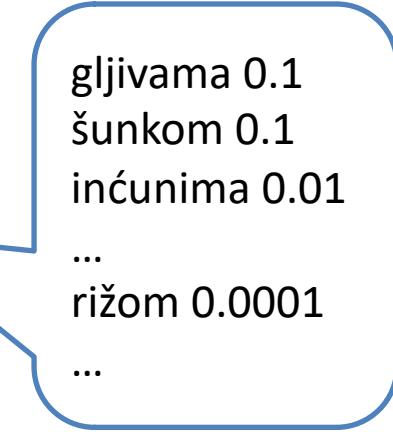
Teškoće vanjske evaluacije n-gram modela

- Vanjska (in-vivo) evaluacija
 - dugotrajna: može trajati danima, tjednima...
- Stoga
 - ponekad koristiti **unutarnju** evaluaciju: **perpleksija**
 - Loša aproksimacija
 - osim ako testni podaci izgledaju kao podaci za trening
 - Općenito korisna samo u pilot eksperimentima
 - Ali je dobra za razmišljanje

Intuicija perpleksije

- Shannonova igra:
 - Koliko dobro ćemo predvidjeti sljedeću riječ?

Uvijek naručujem picu sa sirom i _____



gljivama 0.1
šunkom 0.1
inćunima 0.01
...
rižom 0.0001
...

- Unigrami su očajni kod ove igre...
- Bolji model za tekst
 - je onaj koji pridružuje veću vjerojatnost riječi koja se zapravo pojavila
- Perpleksija je zapravo **težinski faktor grananja.**

Perpleksija

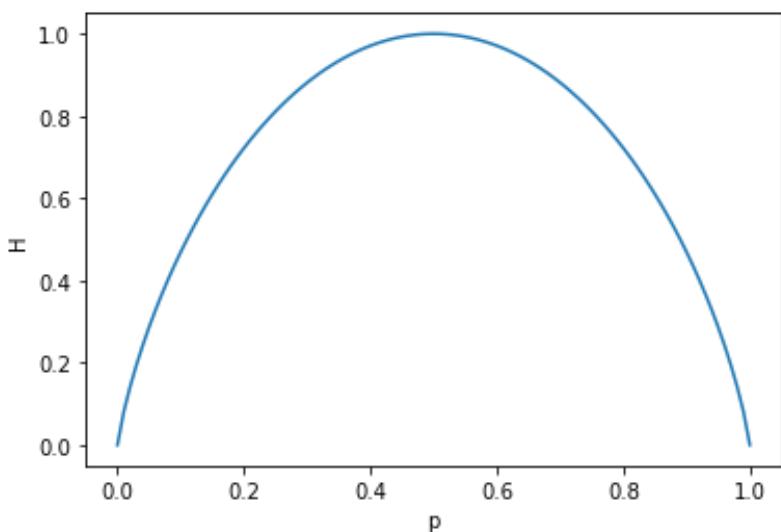
- Najbolji model jezika je onaj koji najbolje predviđa neviđeni testni skup
 - daje najveću vjerojatnost rečenici
- **Perpleksija** je inverzna vjerojatnost testnog skupa W normaliziranog po broju riječi

$$PP(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}}$$

- pravilo lanca $PP(W) = \sqrt[N]{\prod_{i=1}^N \frac{1}{P(w_i | w_1 \dots w_{i-1})}}$
- za bigram $PP(W) = \sqrt[N]{\prod_{i=1}^N \frac{1}{P(w_i | w_{i-1})}}$
- minimizacija perpleksije je isto što i maksimizacija uvjetne vjerojatnosti

Perpleksija - teorija informacija

- **Entropija** – mjera neizvjesnosti distribucije p
 - $H(p) = -\sum_i p_i \log_2 p_i$
 - mjera "bit"
-
- za distribuciju $p, 1 - p$



Perpleksija - teorija informacija

- **Unakrsna entropija** (cross entropy) – entropija distribucija p, q
- $H(p, q) = -\sum_i p_i \log_2 q_i$
- $H(p, q) = H(p) + D_{KL}(p||q)$
- **Kullback-Leibler divergencija** – udaljenost između distribucija p, q
- $D_{KL}(p||q) = -\sum_i p_i \log_2 \frac{p_i}{q_i}$
- Ako se ne zna p , prepostavi se da je uniformna distribucija (maksimalna entropija)
- $H(p, q) = -\frac{1}{N} \sum_{i=1}^N \log_2 q_i$

Perpleksija - teorija informacija

- Neka je T testni skup
- Unakrsna entropija modela p
- $H_p(T) = -\frac{1}{|T|} \sum_{i=1}^{|T|} \log_2 p(w_i)$
- Perpleksija modela p – mjera koliko dobro model p predviđa T
- $PP_p(T) = 2^{H_p(T)} = 2^{-\frac{1}{|T|} \sum_{i=1}^{|T|} \log_2 p(w_i)} = 2^{-\frac{1}{|T|} \log_2 \prod_{i=1}^{|T|} p(w_i)}$
$$= \left(2^{\log_2 \prod_{i=1}^{|T|} p(w_i)} \right)^{-\frac{1}{|T|}} = \left(\prod_{i=1}^{|T|} p(w_i) \right)^{-\frac{1}{|T|}} = \sqrt[|T|]{\frac{1}{\prod_{i=1}^{|T|} p(w_i)}}$$

Perpleksija kao faktor grananja

- Pretpostavimo da se rečenica sastoji od brojčanih znamenka
- Koja je perpleksija ovih rečenica prema modelu koji pridružuje
 $P = \frac{1}{10}$ svakoj znamenki?

$$PP(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} = \left(\frac{1}{10^N}\right)^{-\frac{1}{N}} = \frac{1}{10}^{-1} = 10$$

Perpleksija kao faktor grananja

- Pretpostavimo da je znamenka 0 frekventnija
- U trening skupu:
 - znamenka 0 se pojavljuje 91 put $P(0) = 0.91$
 - ostale znamenke se pojavljuju točno jedanput $P(w) = 0.01$

0 0 0 0 0 0 0 0 0 0

0 0 0 1 0 0 0 0 0 0

0 0 0 2 0 0 0 0 0 0

0 0 0 0 0 3 0 0 0 0

0 0 4 0 0 0 0 0 0 0

0 0 0 0 0 0 0 5 0 0

0 0 0 6 0 0 0 0 0 0

0 0 0 0 7 0 0 0 0 0

0 0 0 0 0 0 0 0 8 0

9 0 0 0 0 0 0 0 0 0

$$P(0) = \frac{91}{100} \quad P(w \neq 0) = \frac{1}{100}$$

$$P(0|0) = \frac{81}{91} \quad P(w \neq 0|0) = \frac{1}{91}$$

$$P(0|w \neq 0) = \frac{1}{1}$$

$$P(0|) = \frac{1}{1} \quad P(w \neq 0|) = \frac{0}{91}$$

Perpleksija kao faktor grananja

- Pretpostavimo da je znamenka 0 frekventnija
- U trening skupu:
 - znamenka 0 se pojavljuje 91 put
 - ostale znamenke se pojavljuju točno jedanput
- Testni skup je rečenica $W=0000030000$
- Kolika je perpleksija?
- Za unigram

$$PP(W) = P(0000030000)^{-\frac{1}{10}} = \sqrt[10]{0.91^{-9} * 0.01^{-1}} \approx 1.73$$

- Za bigram

$$\begin{aligned} PP(W) &= (P(0|)P(0|0)^7 P(0|3)P(3|0))^{-\frac{1}{10}} \\ &= \sqrt[10]{\frac{1^{-1}}{1} * \frac{81^{-7}}{91} * \frac{1^{-1}}{91} * \frac{1^{-1}}{1}} \approx 1.70 \end{aligned}$$

Manja perpleksija = bolji model

- Wall Street Journal
 - Trening od 38 miliona riječi
 - Test od 1.5 miliona riječi

N-gram	Unigram	Bigram	Trigram
perpleksija	962	170	109

- Što nam više informacija n-gram daje to je manja perplexija (veća izvjesnost)

Uvod u obradu prirodnog jezika

4.4. Generalizacija i nule

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Shannon-ova metoda vizualizacije

- Izaberi slučajni bigram ($< s >$, w) prema svojoj vjerojatnosti
- Sada izaberi slučajni bigram (w, x) prema svojoj vjerojatnosti
- I tako dalje sve dok se ne dođe do $</ s >$
- Zatim spoji riječi u rečenicu

```
<s> Ja
Ja želim
želim da
da jedem
jedem kinesku
kinesku hranu
hranu </s>
```

Ja želim da jedem kinesku hranu.

Aproksimacija Shakespeare-a

Unigram

swallowed confess hear both. Which. Of save on trail for are ay device and rote life have
Every enter now severally so, let
Hill he late speaks; or! a more to leg less first yon enter
Are where exeunt and sighs have rise excellency took of .. Sleep knave we. near; vile like

bigram

What means, sir. I confess she? then all sorts, he is trim, captain.
Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow.
What we, hath got so she that I rest and sent to scold and namre bankrupt, nor the first gentleman?

Trigram

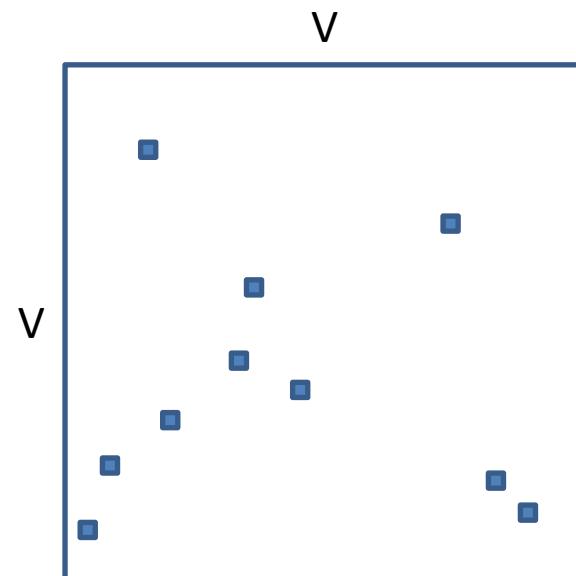
Sweet prince, Falstaff shall die. Hany of Monmouth's grave.
This shall forbid it should be branded, if renown made it empty.
Indeed the duke; and had a very good friend.
Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done.

4-gram

King Henry. What' I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv 'd
in;
Will yon not tell me who I am?
It cannot be but so.
Indeed the short and the long. Many, 'tis a noble Lepidns.

Shakespeare kao korpus

- $N=884647$ pojavnica, $V=29066$ riječi
- Shakespeare je producirao 300000 bigram tipova od $V^2= 844832356$ mogućih bigrama.
 - Stoga 99.96% mogućih bigrama nikad nisu viđeni (nule u tablici)
- 4-grami su još gori



Opasnosti prekoračenja

- n-grami su dobri za predviđanje riječi samo ako je testni korpus sličan korpusu treninga
 - U stvarnosti, to često nije slučaj
 - Potrebno je napraviti trening robustnih modela i zatim generalizirati!
 - Jedna vrsta generalizacije: Nule!
 - stvari koje se uopće ne pojavljuju u skupu za trening
 - ali se pojavljuju u testnom skupu

Nule

Skup za trening	Testni skup
... negirao je optužbe negirao je ponude ...
... negirao je izvještaje negirao je posudbe ...
... negirao je tvrdnje ...	
... negirao je zahtjeve ...	

$$P(\text{ponude} \mid \text{negirao je}) = 0$$

Bigrami s nultom vjerojatnošću

- Pridruživanje vjerojatnosti 0 na testnom skupu
- Stoga se ne može izračunati perpleksija! (dijeljenje s 0)

Nepoznate riječi

- **Otvoreni rječnik** protiv **zatvorenog rječnika**
- Ako unaprijed znamo sve riječi
 - rječnik V je fiksan
 - zadaci **zatvorenog rječnika**
- Često ne znamo sve riječi
 - riječi izvan rječnika
 - zadaci **otvorenog rječnika**
- Kreira se pojavnica za nepoznate riječi <UNK> - unknown (OOV out of vocabulary riječi)

Nepoznate riječi

- Treniranje vjerojatnosti OOV riječi kod otvorenog rječnika
 - kreiranje **fiksnog leksikona L** veličine V
 - prilikom normalizacije teksta, svaka riječ iz trening skupa koja nije u L se promijeni u $\langle\text{UNK}\rangle$
 - sada se njihove vjerojatnosti treniraju kao i za ostale riječi
- Prilikom dekodiranja
 - koristi se $\langle\text{UNK}\rangle$ za svaku riječ koja nije u podacima za trening
- Ako nemamo fiksni leksikon, onda se implicitno stvara leksikon mijenjanjem riječi iz trening skupa u $\langle\text{UNK}\rangle$ temeljem njihove frekvencije.
 - Možemo zamijeniti s $\langle\text{UNK}\rangle$ sve riječi iz skupa za treniranje koje se javljaju manje od n puta (n je neki mali broj)

Uvod u obradu prirodnog jezika

4.5. Izglađivanje: dodaj jedan (Laplaceovo izglađivanje)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Ideja izglađivanja (prema Dan Klein-u)

- Kada imamo spremnu statistiku

$P(w|negirao\ je)$

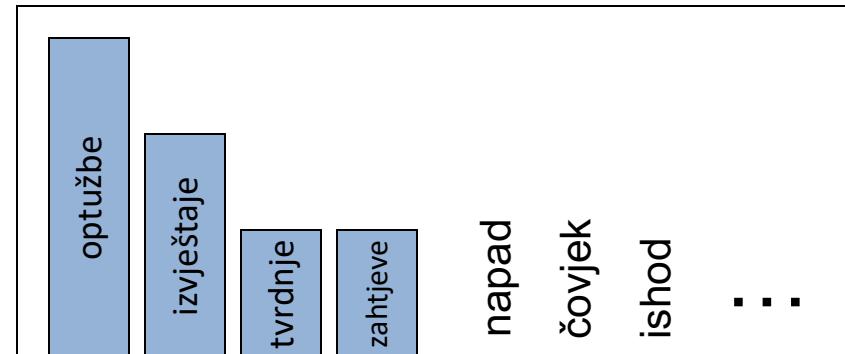
3 optužbe

2 izvještaje

1 tvrdnje

1 zahtjeve

UKUPNO: 7



- Uzmi masu vjerojatnosti radi bolje generalizacije

$P(w|negirao\ je)$

2.5 optužbe

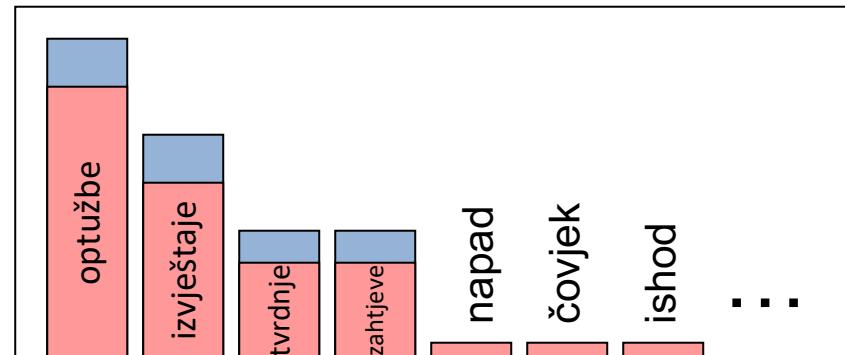
1.5 izvještaje

0.5 tvrdnje

0.5 zahtjeve

2 ostalo

UKUPNO: 7



Dodaj jedan - procjena

- ili Laplaceovo izglađivanje
- Pretvaramo se da smo svaku riječ vidjeli još jedan put
- Dodaj jedan kod svih prebrojavanja!
- Maksimalna izvjesnost (MLE)
 - unigram $P_{MLE}(w_i) = \frac{C(w_i)}{N}$
 - bigram $P_{MLE}(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})}$
- Dodaj 1 (Laplace)
 - unigram $P_{Laplace}(w_i) = \frac{c_i+1}{N+V}$
 - bigram $P_{Laplace}(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)+1}{\sum_w(C(w_{i-1}w)+1)} = \frac{C(w_{i-1}w_i)+1}{C(w_{i-1})+V}$

Dodaj jedan - procjena

- Prilagođeni brojač c^*

$$P_{Laplace}(w_i) = \frac{c_i + 1}{N + V} = \frac{c_i^*}{N} \Rightarrow c_i^* = (c_i + 1) \frac{N}{N + V}$$

$$P_{Laplace}(w_i | w_{i-1}) = \frac{C(w_{i-1}w_i) + 1}{C(w_{i-1}) + V} = \frac{c^*(w_{i-1}w_i)}{C(w_{i-1})} \Rightarrow$$

$$c^*(w_{i-1}w_i) = \frac{(C(w_{i-1}w_i) + 1) \times C(w_{i-1})}{C(w_{i-1}) + V}$$

Berkley restorant korpus

- s primjenjenim Laplaceovim izglađivanjem

	ja	želim	da	jedem	kinesku	hranu	ručak	potrošim
ja	6	828	1	10	1	1	1	3
želim	3	1	609	2	7	7	6	2
da	3	1	5	687	3	1	7	212
jedem	1	1	3	1	17	3	43	1
kinesku	2	1	1	1	1	83	2	1
hranu	16	1	16	1	2	5	1	1
ručak	3	1	1	1	1	2	1	1
potrošim	2	1	2	1	1	1	1	1

Laplaceovo izglađivanje

- Unigram

ja	želim	da	jedem	kinesku	hranu	ručak	potrošim
2533	927	2417	746	158	1093	341	278

- $V = 1446$

$$P_{Laplace}(w_n | w_{n-1}) = \frac{C(w_{n-1} w_n) + 1}{C(w_{n-1}) + V}$$

	ja	želim	da	jedem	kinesku	hranu	ručak	potrošim
ja	0,001505	0,207675	0,000251	0,002508	0,000251	0,000251	0,000251	0,000752
želim	0,00126	0,00042	0,255775	0,00084	0,00294	0,00294	0,00252	0,00084
da	0,000775	0,000258	0,001292	0,177474	0,000775	0,000258	0,001808	0,054766
jedem	0,000455	0,000455	0,001364	0,000455	0,007727	0,001364	0,019545	0,000455
kinesku	0,001241	0,00062	0,00062	0,00062	0,00062	0,051489	0,001241	0,00062
hranu	0,006282	0,000393	0,006282	0,000393	0,000785	0,001963	0,000393	0,000393
ručak	0,001671	0,000557	0,000557	0,000557	0,000557	0,001114	0,000557	0,000557
potrošim	0,001155	0,000577	0,001155	0,000577	0,000577	0,000577	0,000577	0,000577

Rekonstruirani brojač

$$c^*(w_{i-1}w_i) = \frac{(C(w_{i-1}w_i) + 1) \times C(w_{i-1})}{C(w_{i-1}) + V}$$

	ja	želim	da	jedem	kinesku	hranu	ručak	potrošim
ja	3,8	527	0,64	6,4	0,64	0,64	0,64	1,9
želim	1,2	0,39	238	0,78	2,7	2,7	2,3	0,78
da	1,9	0,63	3,1	430	1,9	0,63	4,4	133
jedem	0,34	0,34	1	0,34	5,8	1	15	0,34
kinesku	0,2	0,098	0,098	0,098	0,098	8,2	0,2	0,098
hranu	6,9	0,43	3,9	0,43	0,86	2,2	0,43	0,43
ručak	0,57	0,19	0,19	0,19	0,19	0,38	0,19	0,19
potrošim	0,32	0,16	0,32	0,16	0,16	0,16	0,16	0,16

Usporedba s početnim bigramom

	ja	želim	da	jedem	kinesku	hranu	ručak	potošim
ja	5	827	0	9	0	0	0	2
želim	2	0	608	1	6	6	5	1
da	2	0	4	686	2	0	6	211
jedem	0	0	2	0	16	2	42	0
kinesku	1	0	0	0	0	82	1	0
hranu	15	0	15	0	1	4	0	0
ručak	2	0	0	0	0	1	0	0
potošim	1	0	1	0	0	0	0	0

	ja	želim	da	jedem	kinesku	hranu	ručak	potošim
ja	3,8	527	0,64	6,4	0,64	0,64	0,64	1,9
želim	1,2	0,39	238	0,78	2,7	2,7	2,3	0,78
da	1,9	0,63	3,1	430	1,9	0,63	4,4	133
jedem	0,34	0,34	1	0,34	5,8	1	15	0,34
kinesku	0,2	0,098	0,098	0,098	0,098	8,2	0,2	0,098
hranu	6,9	0,43	3,9	0,43	0,86	2,2	0,43	0,43
ručak	0,57	0,19	0,19	0,19	0,19	0,38	0,19	0,19
potošim	0,32	0,16	0,32	0,16	0,16	0,16	0,16	0,16

Dodaj 1 procjena je tupi alat

- Dodaj 1 se praktički ne koristi kod n-grama:
 - vidjet ćemo bolje metode
- Međutim dodaj 1 se koristi kod izglađivanja u drugim modelima obrade prirodnog jezika
 - za klasifikaciju teksta
 - u domenama gdje je mali broj nula

Dodaj k izglađivanje

- Alternativa dodaj 1 izglađivanju je pomicanje manjeg dijela vjerojatnosti sa viđenih na neviđene događaje.
- Umjesto dodaj 1, dodat će se decimalni broj k npr. 0.5, 0.05, 0.1

$$P_{dodaj-k}^*(w_n | w_{n-1}) = \frac{C(w_{n-1} w_n) + k}{C(w_{n-1}) + kV}$$

- k se često odabire optimizacijom na razvojnom skupu podataka

Trening podaci

Razvojni podaci

Testni podaci

Uvod u obradu prirodnog jezika

4.6. Interpolacija i odustajanje

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Interpolacija i odustajanje

- Ponekad pomaže korištenje **manjeg** sadržaja
- Uvjeti nad manjim sadržajem za sadržaje o kojima se ne zna
- **Odustajanje (Backoff):**
 - koristi trigram ako imаш dobre dokaze, inače bigram, inače unigram
- **Interpolacija:**
 - pomiješaj unigram, bigram, trigram
- Interpolacija radi bolje

Linearna interpolacija

- jednostavna interpolacija

$$\begin{aligned}\hat{P}(w_n | w_{n-2} w_{n-1}) &= \lambda_3 P(w_n | w_{n-2} w_{n-1}) \\ &\quad + \lambda_2 P(w_n | w_{n-1}) \\ &\quad + \lambda_1 P(w_n)\end{aligned}$$

$$\sum_i \lambda_i = 1$$

- lambda je uvjetovan sadržajem

$$\begin{aligned}\hat{P}(w_n | w_{n-2} w_{n-1}) &= \lambda_3 (w_{n-2}^{n-1}) P(w_n | w_{n-2} w_{n-1}) \\ &\quad + \lambda_2 (w_{n-2}^{n-1}) P(w_n | w_{n-1}) \\ &\quad + \lambda_1 (w_{n-2}^{n-1}) P(w_n)\end{aligned}$$

Kako odrediti lambda?

- Koristeći **razvojni** korpus

Trening podaci

Razvojni
podaci

Testni podaci

- Izaberi lambde tako da maskimiziraš vjerojatnost razvojnog korpusa
 - namjesti n-gram vjerojatnosti na korpusu za trening
 - zatim odredi lambde tako da daje najveće vjerojatnosti nad razvojnim korpusom

$$\log P(w_1 \dots w_n | M(\lambda_1 \dots \lambda_k)) = \sum_i \log P_{M(\lambda_1 \dots \lambda_k)}(w_i | w_{i-1})$$

Kako odrediti lambda?

- Metodom maksimiziranja očekivanja
(Expectation Maximization)
- H – razvojni skup
- minimizacija $\frac{-1}{|H|} \sum_{i=1}^{|H|} \log_2 \hat{p}_\lambda(w_i | h_i)$ nad λ gdje je

$$\begin{aligned}\hat{p}_\lambda(w_i | h_i) &= \hat{p}_\lambda(w_i | w_{i-2} w_{i-1}) = \lambda_3 p_3(w_n | w_{n-2} w_{n-1}) \\ &\quad + \lambda_2 p_2(w_n | w_{n-1}) \\ &\quad + \lambda_1 p_1(w_i) \\ &\quad + \frac{\lambda_0}{|V|}\end{aligned}$$

- Izračun očekivanja za $j = 0 \dots 3$

$$c(\lambda_j) = \sum_{i=0}^{|H|} \frac{\lambda_j p_j(w_i | h_i)}{\hat{p}_\lambda(w_i | h_i)}$$

- sljedeći lambda za $j = 0 \dots 3$

$$\lambda_{j,next} = c(\lambda_j) / \sum_{k=0 \dots 3} c(\lambda_k)$$

Kako odrediti lambda?

- **Metoda maksimiziranja očekivanja**
- Sirova distribucija nad rječnikom $V=\{a, b, c, \dots, z\}$ $|V|=26$

$$p(a) = 0.25,$$

$$p(b) = 0.5,$$

$$p(w) = 1/64 \text{ za } w \in \{c, \dots, r\}$$

$$p(w) = 0 \text{ za } w \in \{s, \dots, z\}$$

$$\begin{aligned} p_{\lambda}(w_i|h_i) &= \hat{p}_{\lambda}(w_i|w_{i-2}w_{i-1}) = \lambda_3 p_3(w_n|w_{n-2}w_{n-1}) \\ &\quad + \lambda_2 p_2(w_n|w_{n-1}) \\ &\quad + \lambda_1 p_1(w_i) \\ &\quad + \frac{\lambda_0}{|V|} \end{aligned}$$

- Razvojni skup **H = babu** (λ_1 unigram, λ_0 uniformno)

- Počni s $\lambda_1 = 0.5$ i $\lambda_0 = 0.5$

$$p'_{\lambda}(b) = 0.5 * 0.5 + 0.5 / 26 = 0.27$$

$$p'_{\lambda}(a) = 0.5 * 0.25 + 0.5 / 26 = 0.145$$

$$p'_{\lambda}(u) = 0.5 * 0 + 0.5 / 26 = 0.02$$

$$c(\lambda_j) = \sum_{i=0}^{|H|} \frac{\lambda_j p_j(w_i|h_i)}{\hat{p}_{\lambda}(w_i|h_i)}$$

$$c(\lambda_1) = 0.5 * 0.5 / 0.27 + 0.5 * 0.25 / 0.145 + 0.5 * 0.5 / 0.27 + 0.5 * 0 / 0.02 = 2.72$$

$$\begin{aligned} c(\lambda_0) &= 0.5 * 0.04 / 0.27 + 0.5 * 0.04 / 0.145 + 0.5 * 0.04 / 0.27 + 0.5 * 0.04 / 0.02 \\ &= 1.28 \end{aligned}$$

- Normaliziraj $\lambda_1=0.68$, $\lambda_0=0.32$

- Ponavljam sve dok nove lambde se malo razlikuju od prethodnih lambdi (npr. < 0.01)

Veliki n-grami na Web-u

- Kako se postaviti prema velikim n-gram korpusima
 - Obrezivanje (Pruning)
 - spremiti samo n-grame koji imaju broj pojavljivanja > prag
 - ukloniti jedinke od n-grama višeg reda
 - obrezivanje temeljeno entropijom
 - Povećanje efikasnosti
 - efikasne podatkovne strukture kao "prefiksna stabla" (tries)
 - Bloom-ovi filteri: aproksimativni modeli jezika
 - spremanje riječi preko indeksa, a ne stringova
 - Huffmanovo kodiranje za spremanje velikog broja riječi u 2 bajta
 - kvantiziranje vjerojatnosti (4-8 bitova umjesto broja s pomičnim zarezom od 8 bajtova)

Izglađivanje kod velikih n-grama

- Glupo odustajanje (Stupid backoff)
- Nema popuštanja, ali se koriste relativne frekvencije

$$S(w_i | w_{i-k+1}^{i-1}) = \begin{cases} \frac{c(w_{i-k+1}^i)}{c(w_{i-k+1}^{i-1})} & \text{ako } c(w_{i-k+1}^i) > 0 \\ 0.4 \times S(w_i | w_{i-k+2}^{i-1}) & \text{u suprotnom} \end{cases}$$

$$S(w_i) = \frac{c(w_i)}{N}$$

Uglađivanje n-grama

- Dodaj jedan:
 - Dobro kod kategorizacije teksta
 - Nije dobro za modeliranje jezika
- Najčešće korištena metoda:
 - Proširenje Kneser-Ney interpolacije
- Za velike n-grame:
 - glupo odustajanje

Napredno modeliranje jezika

- Diskriminativni modeli:
 - izbor težina n-grama radi poboljšanja zadatka, a ne radi prilagođavanju skupu za trening
- Modeli temeljeni na parsiranju
- Modeli s predpohranom (Caching models)
 - nedavno korištene riječi imaju veću vjerojatnost da se pojave

$$P_{CACHE}(w|history) = \lambda P(w_i|w_{i-2}w_{i-1}) + (1 - \lambda) \frac{c(w \in history)}{|history|}$$

- ali se pokazala jako lošom metodom kod prepoznavanja govora

Uvod u obradu prirodnog jezika

4.7. Good-Turing izglađivanje

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Napredni algoritmi za izglađivanje

- Ideja većine algoritama izglađivanja
 - Good-Turing
 - Kneser-Ney
 - Witten-Bell
- Upotrijebi broj "stvari" koje smo **jednom vidjeli**
- kako bi procijenili broj "stvari" koje **nikad nismo vidjeli**

Notacija: N_c = frekvencija od frekvencije c

- N_c = broj riječi koje smo vidjeli c puta

Ivan sam ja sam ja Ivan ja nikad ne jedem

ja	3
sam	2
Ivan	2
nikad	1
ne	1
jedem	1

$$N_3 = 1$$

$$N_2 = 2$$

$$N_1 = 3$$

Ideja Good-Turing izglađivanja

- Loviš ribu i ulovio si:
 - 10 srdela, 3 oslića, 2 bukve, **1 zubaca, 1 gofu, 1 komarču** = 18 riba
- Koliko je vjerojatno da će sljedeća riba biti zubatac?
1/18
- Koliko je vjerojatno da će sljedeća riba biti neka nova?
3/18 (jer $N_1 = 3$)
- S obzirom na to, koja je vjerojatnost da će sljedeća riba biti zubatac?
 - Mora biti manja od 1/18
 - Kako to procijeniti

Good Turing proračun

$$P_{GT}^*(\text{riječi s frekvencijom } 0) = \frac{N_1}{N}$$

$$c^* = \frac{(c+1)N_{c+1}}{N_c}$$

- Nije viđeno (npr. Salpa)

$$c = 0$$

$$P_{MLE}(\text{Salpa}) = \frac{0}{18} = 0$$

$$P_{GT}^*(\text{Salpa}) = \frac{N_1}{N} = \frac{3}{18}$$

- Viđeno jednom (npr. Zubatac)

$$c = 1$$

$$P_{MLE}(\text{Zubatac}) = \frac{1}{18} = 0.055$$

$$c^*(\text{Zubatac}) = \frac{(1+1)*N_{1+1}}{N_1} = \frac{2*N_2}{N_1} = \frac{2*1}{3} = \frac{2}{3}$$

$$P_{GT}^*(\text{Zubatac}) = \frac{c^*}{N} = \frac{\frac{2}{3}}{18} = \frac{1}{27}$$

Good-Turing brojevi

- Brojevi iz Church & Gale (1991)
- 22×10^6 riječi iz AP Newswire

$$c^* = \frac{(c + 1)N_{c+1}}{N_c}$$

- relativni popust

$$d_c = \frac{c^*}{c}$$

- absolutni popust

$$d_c = |c^* - c| \approx 0.75 \text{ za } c > 1$$

c	c^*
0	0.0000270
1	0.446
2	1.26
3	2.24
4	3.24
5	4.22
6	5.19
7	6.21
8	7.24
9	8.25

Uvod u obradu prirodnog jezika

4.7. Kneser-Ney izglađivanje

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Apsolutni popust

- Uštedimo na vremenu i jednostavno oduzmimo 0.75 (ili neki d)

$$P_{\text{ApsolutniPopust}}(w_i | w_{i-1}) = \frac{c(w_{i-1}w_i) - d}{c(w_{i-1})} + \lambda(w_{i-1})P(w_i)$$

bigram sa popustom
težina interpolacije

unigram

- Možda koristiti ekstra vrijednosti od d za bigrame koji se pojavljuju 1 put)
- ali hoćemo li uopće koristiti unigram vjerojatnosti $P(w)$?

Kneser-Ney izglađivanje I

- Bolja procjena vjerojatnosti za unigrame nižeg reda!
 - Shannon-ova igra: *Loše vidim bez svojih čitačih _____?*
 - "Zeland" je češći od "naočala,"
 - ... ali "Zeland" uvijek slijedi iza "Novi"
- Unigram je koristan ako nismo prije vidjeli bigram!
- Umjesto $P(w)$: "Koliko vjerojatno je w "
- $P_{NASTAVAK}(w)$: "Koliko je vjerojatno da se w pojavi kao novi nastavak"
 - za svaku riječ w prebroji sve bigrame koje upotpunjaju

Kneser-Ney izglađivanje II

- Koliko puta se w pojavljuje kao novi nastavak:

$$P_{NASTAVAK}(w) \propto |\{w_{i-1} : c(w_{i-1}w) > 0\}|$$

- Normaliziran ukupnim brojem bigram tipova

$$|\{(w_{j-1}w_j) : c(w_{j-1}w_j) > 0\}|$$

$$P_{NASTAVAK}(w) = \frac{|\{w_{i-1} : c(w_{i-1}w) > 0\}|}{|\{(w_{j-1}w_j) : c(w_{j-1}w_j) > 0\}|}$$

Kneser-Ney izglađivanje III

- Broj tipova riječi viđenih da prethode w

$$|\{w_{i-1} : c(w_{i-1}w) > 0\}|$$

- Normaliziran brojem riječi koje prethode svim riječima:

$$P_{NASTAVAK}(w) = \frac{|\{w_{i-1} : c(w_{i-1}w) > 0\}|}{\sum_{w'} |\{w'_{i-1} : c(w'_{i-1}w') > 0\}|}$$

- Česta riječ "Zeland" koja se često nalazi iza "Novi" će imati nisku vjerojatnost nastavka

Kneser-Ney izglađivanje IV

$$P_{KN}(w_i | w_{i-1}) = \frac{\max(c(w_{i-1}w_i) - d, 0)}{c(w_{i-1})} + \lambda(w_{i-1}) P_{NASTAVAK}(w_i)$$

- λ je normalizacijska konstanta; količina vjerojatnosti koju smo izbacili

$$\lambda(w_{i-1}) = \frac{d}{c(w_{i-1})} |\{w : c(w_{i-1}w) > 0\}|$$

normalizirani popust

broj tipova riječi koji mogu slijediti iza w_{i-1}
= # tipova riječi koje smo izbacili
= # puta koliko smo primijenili normalizirani popust

Kneser-Ney izglađivanje: rekurzivni oblik

$$P_{KN}(w_i | w_{i-n+1}^{i-1}) = \frac{\max(c_{KN}(w_{i-n+1}^i) - d, 0)}{c_{KN}(w_{i-n+1}^{i-1})} + \lambda(w_{i-n+1}^{i-1}) P_{KN}(w_i | w_{i-n+2}^{i-1})$$

$$c_{KN} = \begin{cases} \text{broj(*) za veći red} \\ \text{brojnastavka(*) za manji red} \end{cases}$$

- *brojnastavka* = broj jedinstvenih sadržaja s jednom riječju od *

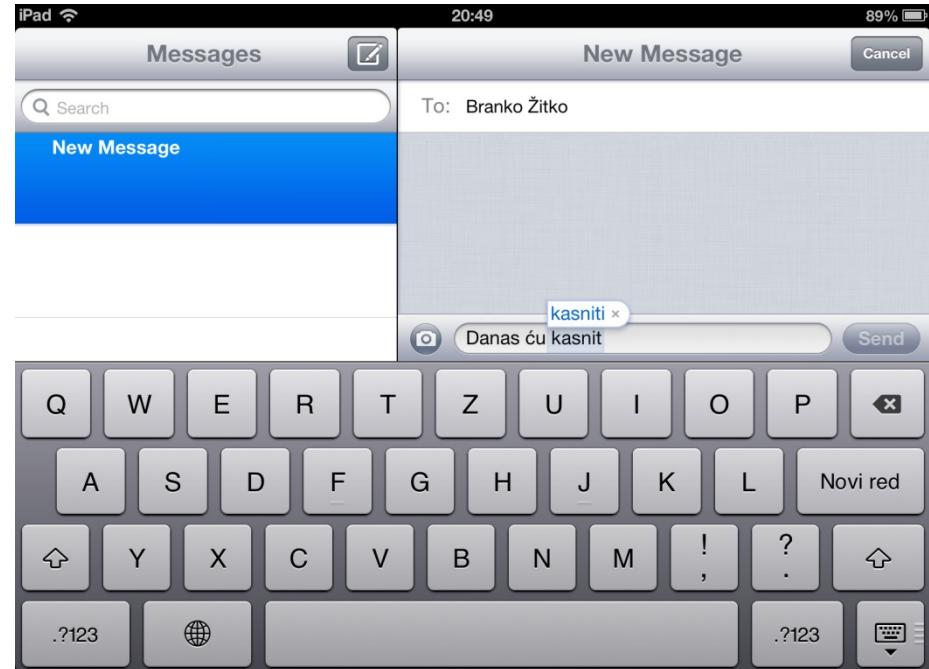
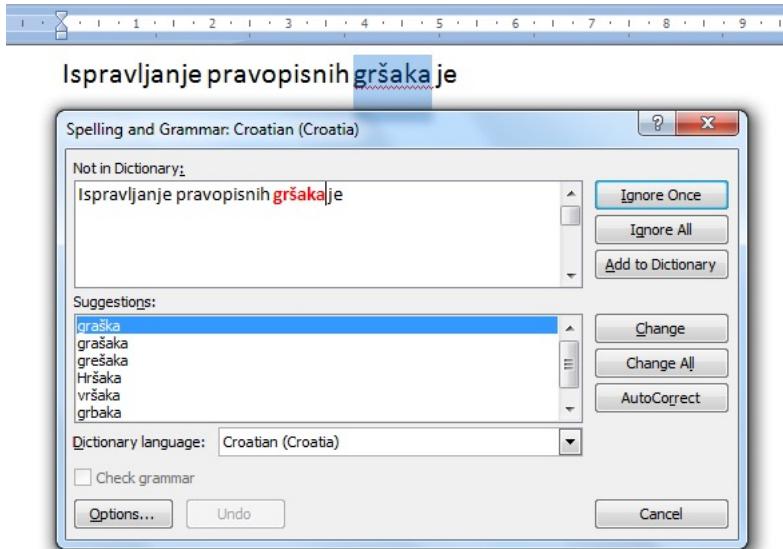
Uvod u obradu prirodnog jezika

5.1. Ispravljanje pravopisnih grešaka (spelling correction)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Primjena ispravljanja pravopisnih grešaka



Oko 174.000 rezultata (0,18 sek)

Prikazuju se rezultati za [obrada prirodnog jezika](#)

Umjesto toga pretražite [obrada ptirodnog jezika](#)

Pravopisni zadaci

- Detekcija pravopisnih grešaka
- Ispravljanje pravopisnih grešaka
 - automatsko ispravljanje
 - sma -> sam
 - predlaganje ispravke
 - lista prijedloga

Tipovi pravopisnih grešaka

- Greške ne-riječi (non-word errors)
 - žrafa -> žirafa
- Greške stvarnih riječi (real-word errors)
 - Tipografske greške
 - staklo -> stablo
 - Kognitivne greške (homonimi, homografi, homofoni)
 - pas -> pas,
 - luk -> luk
 - gore -> gore
 - knight -> night

Stope pravopisnih grešaka

26% Web upiti

Wang et al. 2003

13% Prekucavanje bez brisanja

Whitelaw et al. English&German

7% Riječi ispravljene prekucavanjem na malim uređajima (mobitel)

2% Neispravljene riječi na malim uređajima

Soukoreff & MacKenzie 2003

1-2% Prekucavanje

Kane & Wobbrock 2007, Gruden et al. 1983

Pravopisne greške ne-riječi

- Detekcija pravopisnih grešaka ne-riječi
 - svaka riječ koja nije u **rječniku** je greška
 - bolje je imati veliki rječnik
- Ispravljanje pravopisnih grešaka ne-riječi
 - generiranje **kandidata**: stvarne riječi koje su slične pogrešnoj riječi
 - izbor najboljeg kandidata:
 - najkraća težinska udaljenost riječi
 - najveća vjerojatnost kanala sa šumom

Pravopisne greške stvarnih riječi

- Za svaku riječ w generira se skup kandidata
 - pronađi kandidate sa sličnim **izgovorom**
 - pronađi kandidate sa sličnim **pravopisom**
 - uključi w u skup kandidata
- Izbor najboljeg kandidata
 - kanal sa šumom (noisy channel)
 - klasifikator

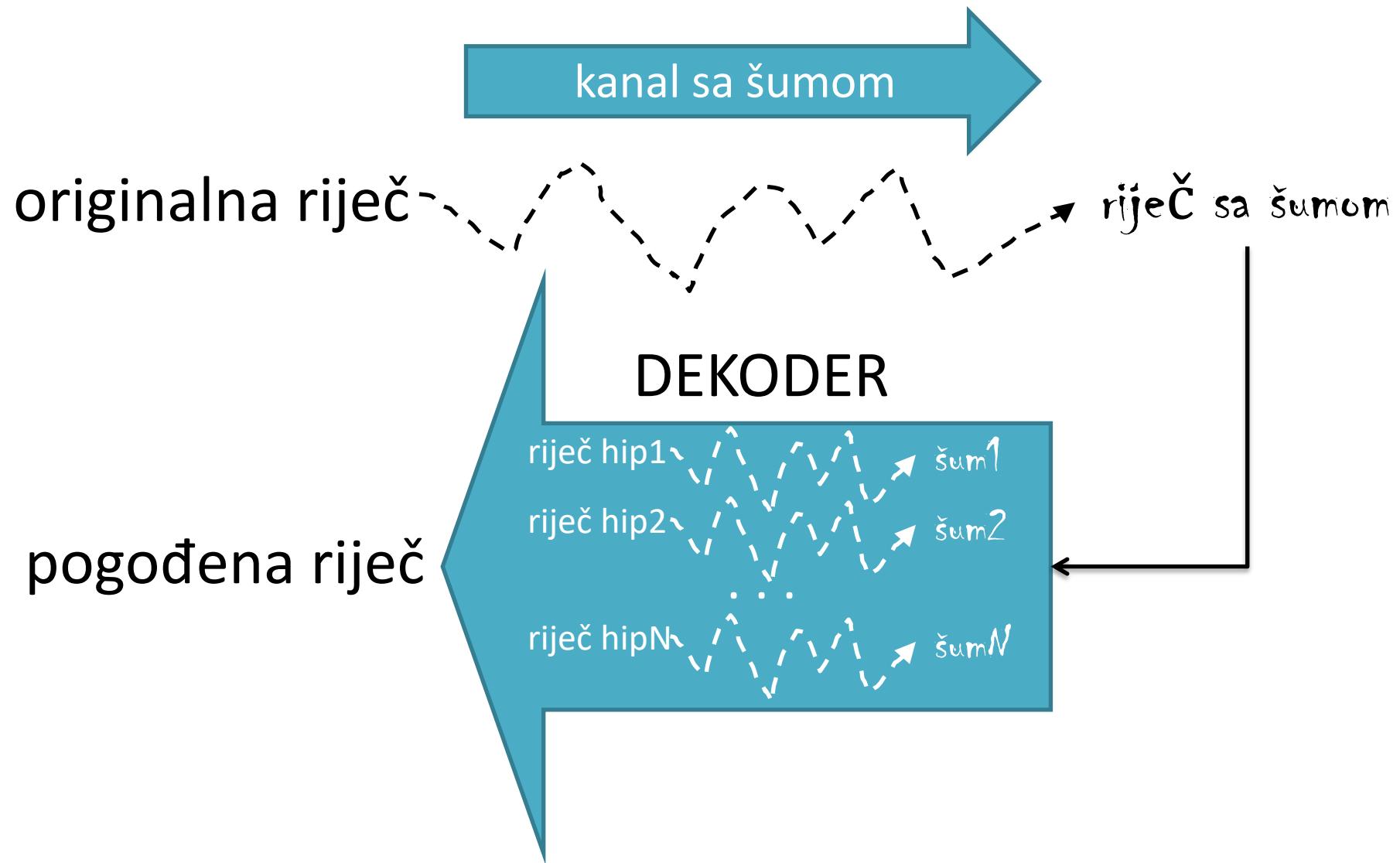
Uvod u obradu prirodnog jezika

5.2. Model ispravljanja pravopisnih grešaka korištenjem kanala sa šumom (the noisy channel model of spelling)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Ideja kanala sa šumom



Kanal sa šumom

- za danu pogrešno napisanu riječ x
- pronađi točnu riječ w

$$\hat{w} = \operatorname{argmax}_{w \in V} P(w|x)$$

$$= \operatorname{argmax}_{w \in V} \frac{P(x|w)P(w)}{P(x)}$$

$$= \operatorname{argmax}_{w \in V} P(x|w)P(w)$$

Izglednost
(model kanala)

prior

Povijest kanala sa šumom

- **IBM**

Mays, Eric, Fred J. Damerau and Robert L. Mercer. 1991.
Context based spelling correction. *Information Processing
and Management*, 23(5), 517–522

- **AT&T Bell Labs**

Kernighan, Mark D., Kenneth W. Church, and William A.
Gale. 1990. A spelling correction program based on a noisy
channel model. *Proceedings of COLING 1990*, 205-210

Primjer ispravljanja greške ne-riječi

acress

Generiranje kandidata

- Riječi sa sličnim pravopisom
 - mala udaljenost između riječi i pogreške
- Riječi sa sličnim izgovorom
 - mala udaljenost između izgovora riječi i pogreške

Damerau-Levenshtein udaljenost

- Minimalna udaljenost dva niza znakova koja uključuje sljedeće operacije:
 - Ubacivanje
 - Brisanje
 - Zamjena
 - Transpozicija dva susjedna znaka
- Damerau-Levenshtein udaljenost

Riječi udaljene za 1 od acress

Greška	Kandidat	Točni znak	Pogrešni znak	Operacija
acress	actress	t	-	brisanje
acress	cress	-	a	ubacivanje
acress	caress	ca	ac	transpozicija
acress	access	c	r	zamjena
acress	across	o	e	zamjena
acress	acres	-	s	ubacivanje
acress	acres	-	s	ubacivanje

Generiranje kandidata

- 80% grešaka su udaljene za 1
- Skoro sve greške su udaljene za 2
- Dozvoliti ubacivanje **razmaka** i **crtice**
 - ovaideja -> ova ideja
 - splitskodalmatinska -> splitsko-dalmatinska

Model jezika

- Koristiti bilo koji algoritam za modeliranje jezika
 - unigram, bigram, trigram
- Ispravljanje pravopisnih grešaka za velike sadržaje (Web)
 - glupo odustajanje (Stupid backoff)

Vjerojatnost za prior kod unigrama

- od 404253213 riječi u Corpus of Contemporary English (COCA)

rijec	frekvencija rijeci	P(rijec)
actress	9321	0.0000230573
cress	220	0.0000005442
caress	686	0.0000016969
access	37038	0.0000916207
across	120844	0.0002989314
acres	12874	0.0000318463

Vjerojatnost kanala sa šumom

- Vjerojatnost modela pogreške
 - Kernighan, Church, Gale 1990
- Pogrešno napisana riječ $x = x_1 x_2 x_3 \dots x_m$
- Točna riječ $w = w_1 w_2 w_3 \dots w_n$
- $P(x|w) =$ vjerojatnost operacije
 - (brisanje/ubacivanje/zamjena/transpozicija)

Izračun vjerojatnosti pogreške: matrica konfuzije

brisanje[x, y]:	broj(xy unesenih kao x)
ubacivanje[x, y]:	broj(x unesenog kao xy)
zamjena[x, y]:	broj(x unesenog kao y)
transpozicija[x, y]:	broj(xy unesenih kao yx)

Ubacivanje i brisanje ovisi o prethodnom znaku

Matrica konfuzije za pravopisne pogreške

zamjena[X,Y] = supstitucija od X(pogrešno) s Y(točno)

	Y(točno)																									
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
X	a	0	0	7	1342	0	0	2118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0	
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	14	0	1	0	18	
f	0	15	0	3	1	0	5	2	0	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	
p	0	11	1	2	0	6	5	0	2	0	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	3	0	

Generiranje matrice konfuzije

Roger Mitton-ova lista pogrešaka

<https://www.dcs.bbk.ac.uk/~ROGER/corpora.html>

Peter Norvig-ova lista pogrešaka

<http://norvig.com/ngrams/spell-errors.txt>

Model kanala sa šumom

$$P(x|w) = \begin{cases} \frac{\text{brisanje}[x_{i-1}, w_i]}{\text{broj}[x_{i-1}w_i]} & \text{ako je brisanje} \\ \frac{\text{ubacivanje}[x_{i-1}, w_i]}{\text{broj}[x_{i-1}w_{i-1}]} & \text{ako je ubacivanje.} \\ \frac{\text{zamjena}[x_i, w_i]}{\text{broj}[w_i]} & \text{ako je zamjena.} \\ \frac{\text{transpozicija}[w_i, w_{i+1}]}{\text{broj}[w_i w_{i+1}]} & \text{ako je transpozicija} \end{cases}$$

Model kanala za acres

Kandidat	Točni znak	Pogrešni znak	x w	P(x w)
actress	t	-	c ct	0.000117
cress	-	a	a #	0.00000144
caress	ca	ac	ac ca	0.00000164
access	c	r	r c	0.000000209
across	o	e	e o	0.0000093
acres	-	s	es e	0.0000321
acres	-	s	ss s	0.0000342

Vjerojatnost kanala sa šumom za acress

Kandidat	Točni znak	Pogrešni znak	$x w$	$P(x w)$	$P(w)$	$10^9 * P(x w)P(w)$
actress	t	-	c ct	0.000117	0.0000231	2.7
cress	-	a	a #	0.00000144	0.000000544	0.00078
caress	ca	ac	ac ca	0.00000164	0.00000170	0.0028
access	c	r	r c	0.000000209	0.0000916	0.019
across	o	e	e o	0.0000093	0.000299	2.8
acres	-	s	es e	0.0000321	0.0000318	1.0
acres	-	s	ss s	0.0000342	0.0000318	1.0

Vjerojatnost kanala sa šumom za acress

Kandidat	Točni znak	Pogrešni znak	$x w$	$P(x w)$	$P(w)$	$10^9 * P(x w)P(w)$
actress	t	-	c ct	0.000117	0.0000231	2.7
cress	-	a	a #	0.00000144	0.000000544	0.00078
caress	ca	ac	ac ca	0.00000164	0.00000170	0.0028
access	c	r	r c	0.000000209	0.0000916	0.019
across	o	e	e o	0.0000093	0.000299	2.8
acres	-	s	es e	0.0000321	0.0000318	1.0
acres	-	s	ss s	0.0000342	0.0000318	1.0

Korištenje digram modela jezika

- "a stellar and **versatile actress** whose combination of sass and glamour..."
- Frekvencije iz Corpus of Contemporary American English s dodaj 1 izglađivanjem

$$P(\text{actress} \mid \text{versatile}) = 0.000021 \quad P(\text{whose} \mid \text{actress}) = 0.0010$$

$$P(\text{across} \mid \text{versatile}) = 0.000021 \quad P(\text{whose} \mid \text{across}) = 0.000006$$

$$P(\text{"versatile actress whose"}) = 0.000021 * 0.0010 = 210 \times 10^{-10}$$

$$P(\text{"versatile across whose"}) = 0.000021 * 0.000006 = 1 \times 10^{-10}$$

Korištenje digram modela jezika

- "a stellar and **versatile actress** whose combination of sass and glamour..."
- Frekvencije iz Corpus of Contemporary American English s dodaj 1 izglađivanjem

$$P(\text{actress} \mid \text{versatile}) = 0.000021 \quad P(\text{whose} \mid \text{actress}) = 0.0010$$

$$P(\text{across} \mid \text{versatile}) = 0.000021 \quad P(\text{whose} \mid \text{across}) = 0.000006$$

$$P(\text{"versatile actress whose"}) = 0.000021 * 0.0010 = 210 \times 10^{-10}$$

$$P(\text{"versatile across whose"}) = 0.000021 * 0.000006 = 1 \times 10^{-10}$$

Evaluacija

- Neki skupovi za testiranje pravopisnih pogrešaka (Engleski)
 - [Wikipedia's list of common English misspelling](#)
 - [Aspell filtered version of that list](#)
 - [Birkbeck spelling error corpus](#)
 - [Peter Norvig's list of errors \(includes Wikipedia and Birkbeck, for training or testing\)](#)
- Neki Hrvatski rječnici za pravopisne pogreške
 - [Croatian Dictionary \(Hrvatski Rjecnik\) for Mozilla Firefox, Thunderbird and SeaMonkey](#)
 - [Croatian dictionary and hyphenation patterns](#)

Uvod u obradu prirodnog jezika

5.3. Ispravljanje pravopisnih pogrešaka stvarnih riječi (real-word spelling correction)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Pravopisne greške stvarnih riječi

- ...odlazim za petnaest **minueta** kako bi stigao na vlak.
 - Razbio sam vazu **trome** mužu.
 - Možeš li me **nosivi**?
 - Računalo ima procesor **a** memoriju.
-
- 25-40% pravopisnih pogrešaka su stvarne riječi Kukich 1992

Ispravljanje pravopisnih grešaka stvarnih riječi

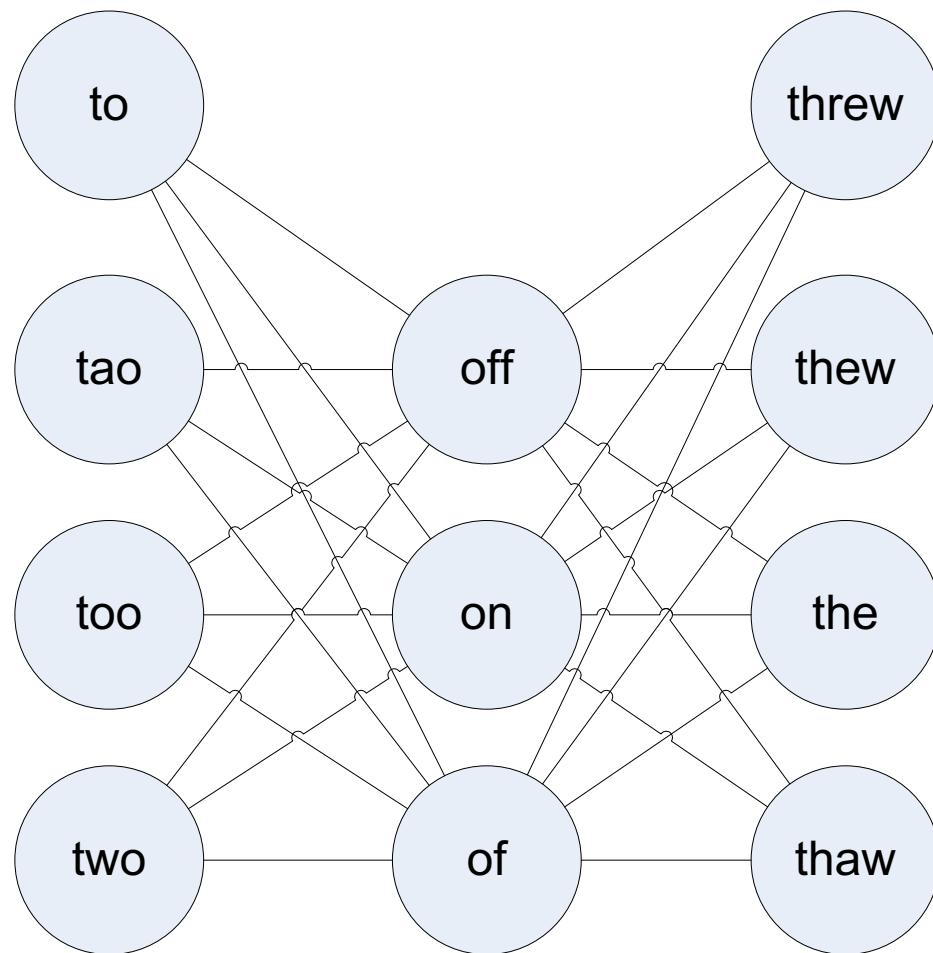
- Za svaku riječ u rečenici
 - generiraj skup kandidata
 - sama riječ
 - sve riječi iz rječnika koje su nastale provedbom jedne operacije (ubacivanje, brisanje, zamjena, transpozicija)
 - riječi koje su homonimi
- Izaberi najboljeg kandidata
 - model kanala sa šumom
 - specijalizirani klasifikator

Kanal sa šumom za stvarne riječi

- Za danu rečenicu $w_1, w_2, w_3 \dots w_n$
- Generiraj skup kandidata za svaku riječ w_i
 - Kandidat(w_1) = $\{w_1, w'_1, w''_1, w'''_1, \dots\}$
 - Kandidat(w_2) = $\{w_2, w'_2, w''_2, w'''_2, \dots\}$
 - Kandidat(w_n) = $\{w_n, w'_n, w''_n, w'''_n, \dots\}$
- Izaberi slijed riječi W koji maksimizira $P(W)$

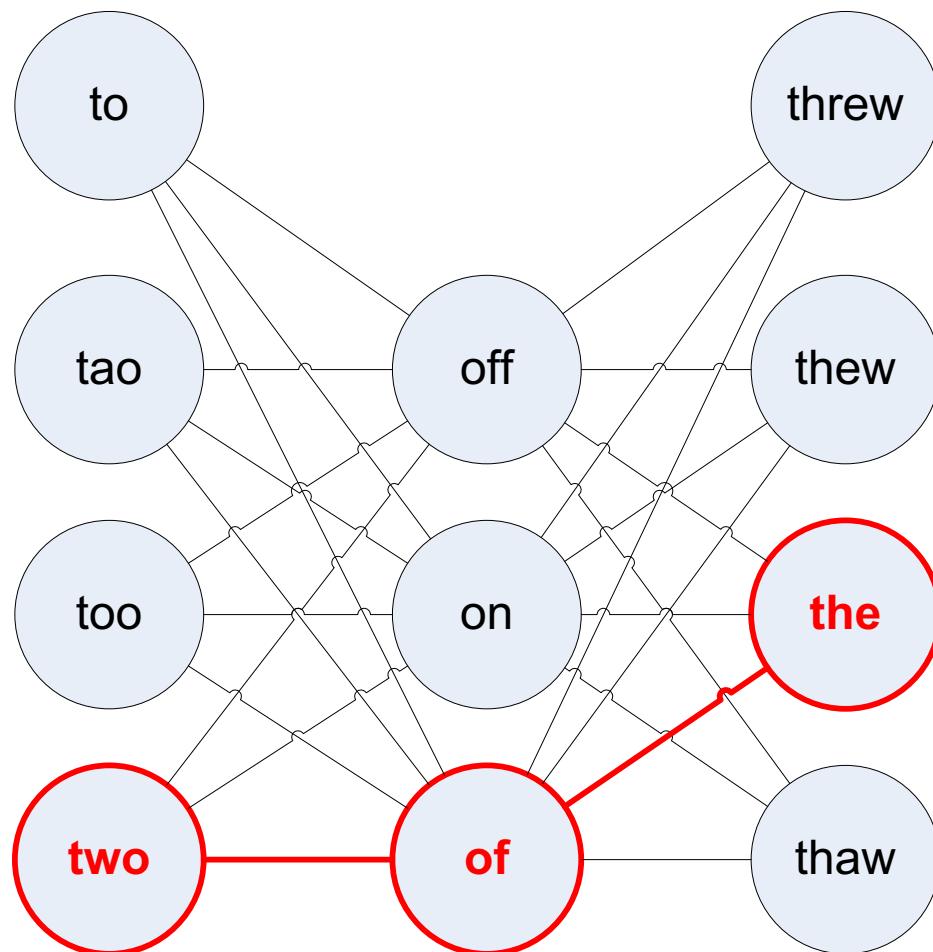
Kanal sa šumom za stvarne riječi

two of them ...



Kanal sa šumom za stvarne riječi

two of them ...



Pojednostavljenje: Jedna greška po rečenici

- Od svih mogućih rečenica s jednom pogrešnom riječi
 - w_1, w''_2, w_3, w_4 two off thew
 - w_1, w_2, w'_3, w_4 two of the
 - w'''_1, w_2, w_3, w_4 too of thew
 - ...
- Izaberi slijed W koji maksimizira $P(W)$

Kako izračunati vjerojatnosti

- Model jezika
 - unigram
 - bigram
 - ...
- Model kanala
 - Isto kao i za greške ne-riječi
 - dodatno, potrebno je izračunati vjerojatnost $P(w|w)$

$$\hat{W} = \operatorname{argmax}_{W \in C(X)} P(X|W)P(W)$$

...
only two of thew apples
oily two of thew apples
only too of thew apples
only to of thew apples
only tao of the apples
only two on thew apples
only two off thew apples
only two of the apples
only two of threw apples
only two of thew applies
only two of thew dapples
...

Vjerojatnost da nije greška

- Koja je vjerojatnost kanala točno upisane riječi?
- $P(\text{"the"} | \text{"the"})$
- Prepostavimo $P(w|w) = \alpha$
- Onda jednostavan model kanala je

$$p(x|w) = \begin{cases} \alpha & \text{ako } x = w \\ \frac{1 - \alpha}{|C(x)|} & \text{ako } x \in C(x) \\ 0 & \text{inače} \end{cases}$$

- Biranje α , ovisno o aplikaciji
 - 0.90 (1 greška od 10 riječi)
 - 0.95 (1 greška od 20 riječi)
 - 0.99 (1 greška od 100 riječi)
 - 0.995 (1 greška od 200 riječi)

Peter Norvig primjer s "thew"

Složeniji model kanal koristit će matrice konfuzije

x	w	x w	P(x w)	P(w)	$10^9 P(x w)P(w)$
thew	the	ew e	0.000007	0.02	144
thew	thew		0.95	0.0000009	90
thew	thaw	e a	0.001	0.0000007	0.7
thew	threw	h hr	0.000008	0.000004	0.03
thew	thwe	ew we	0.000003	0.0000004	0.0001

Uvod u obradu prirodnog jezika

5.4. Najsuvremeniji sustavi (state-of-the-art systems)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Treba li riječ ispravljati?

- Model kanala je sklon ispravljanju točnih riječi (rijetka imena) s frekventnijom riječi
- Proširenje modela kanala:
 - riječ potrebno ispravljati ili ne
- Primjeri:
 - korištenjem crne liste
 - zabrana promjene određenih pojavnica (brojevi, interpunkcija, riječi s jednim znakom)
 - razlika između vjerojatnosti promjene ili ne
 $\log P(w|x) - \log P(x|x) > \Theta$

HCI pitanja kod pravopisnih grešaka

- Ako smo veoma uvjereni kod ispravljana
 - automatsko ispravljanje
- Manje uvjereni
 - daj najbolju ispravku
- Još manje uvjereni
 - daj listu ispravaka
- Nismo uvjereni
 - samo označi kao grešku

Odluka najčešće ovisi o klasifikatoru

Suvremeni kanal sa šumom

- Nikad se ne množi model jezika s modelom kanala
- Pretpostavka nezavisnosti → vjerojatnost nije primjerena
- Umjesto toga ih izmjeri i skaliraj

$$\hat{w} = \operatorname{argmax}_{w \in V} P(x | w) P(w)^\lambda$$

- Nauči λ iz razvojnog skupa

Fonetski model pogreške

- Metafon, kod GNU aspell
 - Konvertiraj pogrešku u metafonski izgovor
 - Izbaci duple susjedne znakove, osim C
 - Ako riječ počinje s KN, GN, PN, AE, WR; izbaci prvi znak
 - Izbaci B ako je iza M i ako je na kraju riječi
 - ...
 - Pronađi riječi čiji je izgovor udaljen za 1-2 od pogrešne riječi
 - boduj rezultate
 - težinska udaljenost između kandidata i pogreške
 - udaljenost između izgovora kandidata i izgovora pogreške

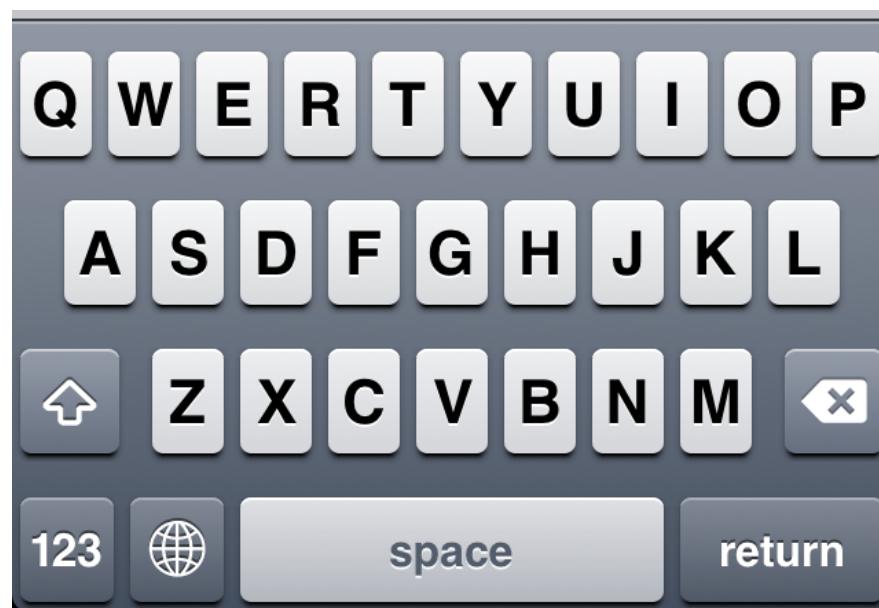
Poboljšanja na kanalu za šum

- Dozvoli bogatije operacije (Brill & Moore 2000)
 - ent -> ant
 - ph -> f
 - le -> al
- Integriranje izgovora u kanal (Toutanova & Moore 2002)

Model kanala

- Faktori koji mogu utjecati na $P(\text{pogreška} \mid \text{word})$
 - početni znak
 - krajnji znak
 - Okolni znakovi
 - pozicija znaka u riječi
 - susjedne tipke na tipkovnici
 - izgovor
 - transformacije sličnih morfema
 - ...

Susjedne tipke



Uvod u obradu prirodnog jezika

6.1. Zadaci klasifikacije teksta

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Je li ovo SPAM?

Subject: **Važna obavijest!**

From: PMF Split info@pmfst.hr

Date: 16. Svibanj, 2013 12:34:56

To: undisclosed-recipients

Sjajne vijesti!

Možete pristupiti najnovijim vijestima koristeći donji link za prijavu na forum Prirodoslovno-matematičkog fakulteta

<http://www.kontakt-forum.hr/forum/form-pmf-split.html>

Kliknite na gornji link da dobijete više informacija o ovom novom forumu. Također možete kopirati gornji link i prenijeti ga u Web preglednik i prijaviti se kako bi saznali više o ovoj novoj usluzi.

© Prirodoslovno-matematički fakultet

Pozitivna ili negativna kritika filma?

- nevjerljivo razočarenje 
- pun otkačenih likova i bogato primijenjena satira, s nekim velikim zapletima radnje 
- ovo je najveća ekscentrična komedija ikad snimljena 
- ovo je jadno, najgori dio je definitivno scena boksa 

Koja je tema ovog članka?

MEDLINE članak



MeSH - hijerarhija kategorija subjekta

- kemija
- krvotok
- terapija lijekovima
- embriologija
- epidemiologija
- ...

Klasifikacija teksta

- Pridruživanje kategorije, naslova ili žanra nekoj temi
- Detekcija spam-a
- Identifikacija autora
- Identifikacija dobi/spola
- Identifikacija jezika
- Analiza sentimenta
- ...

Klasifikacija teksta: definicija

- Ulaz:
 - dokument d
 - fiksni skup klasa $C = \{c_1, c_2, \dots, c_K\}$
- Izlaz:
 - predviđena klasa $c \in C$

Metode klasifikacije: ručno pisana pravila

- Pravila temeljena na kombinacijama riječi i drugih osobina
 - spam: crna-lista-adresa ILI ("\$" I "izabrani ste")
- Preciznost može biti velika
 - ako su pravila brižno pisana od strane eksperta
- Ali izgradnja i održavanje pravila je skupo

Metode klasifikacije: nadzirano strojno učenje

- Ulaz:
 - dokument d
 - fiksni skup klasa $C = \{c_1, c_2, \dots, c_K\}$
 - skup za trening N ručno označenih dokumenata $(d_1, c_1), \dots, (d_N, c_N)$
- Izlaz:
 - naučeni klasifikator $\gamma: d \rightarrow c$

Metode klasifikacije: nadzirano strojno učenje

- Bilo koja vrsta klasifikatora
 - Naivni Bayes
 - Logistička regresija
 - Stroj s potpornim vektorima
 - k-najbližih susjeda
 - ...

Uvod u obradu prirodnog jezika

6.2. Naivni Bayes (naive bayes)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Ideja naivnog Bayesa

- Jednostavna (naivna) metoda klasifikacije temeljena na Bayesovom pravilu
- Oslanja se na jednostavnu reprezentaciju teksta
 - vreća riječi (bag of words)

Reprezentacija vreće riječi

Y()=C

Volim ovaj film! Sladak je, ali sa satiričnim humorom. Dijalog je super i pustolovne scene su zabavne... Uspijeva biti hirovit i romantičan, iako ismijava konvencije žanra bajke. Ja bih ga preporučio bilo kome. Vidio sam ga nekoliko puta i uvek se radujem vidjeti ga ponovno kad god imam prijatelja koji ga još nije video.



Reprezentacija vreće riječi

Y(Volim ovaj film! **Sladak** je, ali sa **satiričnim** humorom. Dijalog je **super** i pustolovne scene su **zabavne**... Uspijeva biti **hirovit** i **romantičan**, iako **ismijava** konvencije žanra bajke. Ja bih ga **preporučio** bilo kome. Vidio sam ga **nekoliko** puta i uvijek se radujem vidjeti ga **ponovno** kad god imam prijatelja koji ga još nije video.)=C



Reprezentacija vreće riječi

Y(Volim ----- Sladak -----
satiričnim -----
super -----
zabavne ----- hirovit -
romantičan ----- ismijava -----

preporučio -----
nekoliko -----
----- ponovno -----

)=C



Reprezentacija vreće riječi

Y(

volim	2
sladak	2
preporučio	1
ismijava	1
super	1
...	...

)=C



Uvod u obradu prirodnog jezika

6.3. Formalizacija naivnog Bayesovog klasifikatora

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Bayesovo pravilo primijenjeno na dokumente i klase

- Za dokument d i klasu c

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Naivni Bayesov klasifikator

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(c|d)$$

MAP
=
Maximum a posteriori
=
najvjerojatnija klasa

$$= \operatorname{argmax}_{c \in C} \frac{P(d|c)P(c)}{P(d)}$$

Bayesovo
pravilo

$$= \operatorname{argmax}_{c \in C} P(d|c)P(c)$$

Izbacivanje
nazivnika

Naivni Bayesov klasifikator

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(d|c)P(c)$$

izglednost

priori

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(f_1, f_2, \dots, f_n | c)P(c)$$

Dokument d prikazan
kao skup osobina
 f_1, f_2, \dots, f_n

Naivni Bayesov klasifikator

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(f_1, f_2, \dots, f_n | c) P(c)$$

izglednost

prior

- Koliko često se klasa c pojavljuje
 - možemo izračunati relativne frekvencije u korpusu
- Kako odrediti izglednost od d i osobina f_1, f_2, \dots, f_n
 - procjena svih mogućih kombinacija osobina bi zahtijevala veliki broj parametara i ogromni skupove za treniranje
 - npr. svi mogući skupovi riječi i pozicija tih riječi
- Stoga koriste se dvije pojednostavljajuće pretpostavke

Naivna Bayesova pretpostavka nezavisnosti

$$P(f_1, f_2, \dots, f_n | c)$$

- **Pretpostavka vreće riječi**
 - pozicija nije važna
 - stoga f_1, f_2, \dots, f_n kodira identitet riječi, a ne njen položaj
- **Uvjetna nezavisnost**
 - vjerojatnost osobina $P(f_i | c_j)$ su nezavisne za danu klasu c

$$P(f_1, f_2, \dots, f_n | c) = P(f_1 | c) \times P(f_2 | c) \times \dots \times P(f_n | c)$$

Naivni Bayesov klasifikator

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(f_1, f_2, \dots, f_n | c) P(c)$$

$$c_{NB} = \operatorname{argmax}_{c \in C} P(c) \prod_{f \in F} P(f | c)$$

Primjena naivnog Bayesovog klasifikatora

pozicije \leftarrow sve pozicije riječi u testnom dokumentu

$$c_{NB} = \operatorname{argmax}_{c \in C} P(c) \prod_{i \in \text{pozicije}} P(w_i | c)$$

$$c_{NB} = \operatorname{argmax}_{c \in C} \left(\log P(c) + \sum_{i \in \text{pozicije}} \log P(w_i | c) \right)$$

Uvod u obradu prirodnog jezika

6.4. Učenje naivnog Bayesa

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Učenje naivnog Bayesovog modela

- Kako naučiti vjerojatnosti $P(c)$ i $P(f_i|c)$?
- Prvi pokušaj: procjena maksimalne izglednosti (MLE)
 - jednostavno koristi frekvencije podataka

$$\hat{P}(c) = \frac{N_c}{N_{doc}}$$

Broj dokumenata klase c

Broj svih dokumenata

$$\hat{P}(w_i|c) = \frac{\text{broj}(w_i, c)}{\sum_{w \in V} \text{broj}(w, c)}$$

Broj pojavljivanja riječi w_i u
svim dokumentima klase c

Broj pojavljivanja riječi w u
svim dokumentima klase c

V - sve riječi iz
svih dokumenata

Procjena parametara

- Koliko puta se riječ w_i pojavljuje među svim riječima u dokumentu klase c

$$\hat{P}(w_i|c) = \frac{\text{broj}(w_i, c)}{\sum_{w \in V} \text{broj}(w, c)}$$

- Kreira se mega-dokument za klasu c tako što se povežu svi dokumenti klase c
 - koristi se frekvencija riječi w iz mega-dokumenta

Problem kod maksimalne izglednosti

- Što ako nemamo niti jedan dokument za treniranje s riječju "fantastično" koja je klasificirana za klasu pozitivno?

$$\hat{P}(\text{"fantastično"}|\text{poz}) = \frac{\text{broj("fantastično", poz)}}{\sum_{w \in V} \text{broj}(w, \text{poz})} = 0$$

- Nulte vjerojatnosti se ne mogu izbjegći

$$c_{NB} = \operatorname{argmax}_{c \in \{\text{poz}, \text{neg}\}} P(\text{poz}) \prod_{i \in \text{pozicije}} P(w_i | c)$$

Laplace (dodaj 1) izglađivanje za naivnog Bayesa

$$\begin{aligned}\hat{P}(w_i|c) &= \frac{\text{broj}(w_i, c) + 1}{\sum_{w \in V} (\text{broj}(w, c) + 1)} \\ &= \frac{\text{broj}(w_i, c) + 1}{\sum_{w \in V} \text{broj}(w, c) + |V|}\end{aligned}$$

Laplace (dodaj 1) izglađivanje: nepoznate riječi

$$\begin{aligned}\hat{P}(w_u | c) &= \frac{\text{broj}(w_u, c) + 1}{\sum_{w \in V} \text{broj}(w, c) + |V| + 1} \\ &= \frac{1}{\text{broj}(w, c) + |V| + 1}\end{aligned}$$

- Nepoznate riječi se mogu ignorirati
 - Ako riječ w_u iz testnog skupa nije u rječniku V onda se w_u ignorira

Naivni Bayes: Učenje

UČENJE(D, C)

za svaku klasu $c \in C$

$N_{doc} \leftarrow$ broj dokumenata iz D

$N_c \leftarrow$ broj dokumenata iz D klase c

$prior[c] \leftarrow \frac{N_c}{N_{doc}}$

$V \leftarrow$ riječnik dokumenata iz D

$megadoc[c]$ proširi s d za $d \in D$ koji su klase c

za svaku riječ $w \in V$

$broj[w, c] \leftarrow$ broj pojavljivanja od w u $megadoc(c)$

$izvjesnost[w, c] \leftarrow \frac{broj[w, c] + 1}{\sum_{w' \in V} (broj[w', c] + 1)}$

vrati prior, izvjesnost, V

Naivni Bayes: Testiranje

TESTIRANJE ($testdoc, prior, izvjesnost, C, V$)

za svaku klasu $c \in C$

$posterior[c] \leftarrow prior[c]$

za svaku poziciju i iz $testdoc$

$w \leftarrow testdoc[i]$

ako $w \in V$

$posterior[c] = posterior[c] * izvjesnost[w, c]$

vrati $\operatorname{argmax}_{c \in C} posterior[c]$

Uvod u obradu prirodnog jezika

6.5. Odnos s modelom jezika

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Naivni Bayes i model jezika

- Naivni Bayesov klasifikator može koristiti bilo koje osobine
 - URL, email adresa, rječnici, svojstva mreže
- Ali ako
 - koristimo **samo** riječi kao osobine
 - koristimo **sve** riječi iz teksta (ne iz podskupa teksta)
- onda
 - Naivni Bayes ima velike sličnosti s modelom jezika

Svaka klasa je unigram

- Svakoj riječi w se pridružuje $P(w|c)$
- Svakoj rečenici s se pridružuje $P(s|c) = \prod P(w|c)$

Klasa = poz	
0.1	Ja
0.1	volim
0.01	ovaj
0.05	novi
0.1	film

$$P(s|\text{poz}) = 0.1 * 0.1 * 0.01 * 0.05 * 0.01 = 0.0000005$$

Svaka klasa je unigram

- Koja klasa pridružuje veću vjerojatnost rečenici s ?

Klasa = poz	
0.1	Ja
0.1	volim
0.01	ovaj
0.05	novi
0.1	film

Klasa = neg	
0.2	Ja
0.001	volim
0.01	ovaj
0.005	novi
0.1	film

$$P(s|poz) = 0.1 * 0.1 * 0.01 * 0.05 * 0.01 = 0.0000005$$

$$P(s|neg) = 0.2 * 0.001 * 0.01 * 0.005 * 0.1 = 0.0000001$$

$$P(s|poz) > P(s|neg)$$

Uvod u obradu prirodnog jezika

6.6. Multinominalni naivni Bayes: Radni primjer

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Naivni Bayes i model jezika

	Dokument	Riječi	Klasa
Treniranje	d ₁	Italija Rim Italija	IT
	d ₂	Italija Italija Firenca	IT
	d ₃	Italija Ankona	IT
	d ₄	Pariz Francuska Italija	FR
Test	d ₅	Italija Italija Italija Pariz Francuska	?

V = {Italija, Rim, Firenca, Ankona, Francuska, Pariz}

$$\hat{P}(c) = \frac{N_c}{N_{doc}}$$

$$\hat{P}(w_i|c) = \frac{\text{broj}(w_i, c) + 1}{\sum_{w \in V} \text{broj}(w, c) + |V|}$$

Prior

Uvjetna vjerojatnost

Izbor klase

Naivni Bayes i model jezika

	Dokument	Riječi	Klasa
Treniranje	d ₁	Italija Rim Italija	IT
	d ₂	Italija Italija Firenca	IT
	d ₃	Italija Ankona	IT
	d ₄	Pariz Francuska Italija	FR
Test	d ₅	Italija Italija Italija Pariz Francuska	?

$$V = \{\text{Italija, Rim, Firenca, Ankona, Francuska, Pariz}\}$$

$$\hat{P}(c) = \frac{N_c}{N_{doc}}$$

$$\hat{P}(w_i|c) = \frac{\text{broj}(w_i, c) + 1}{\sum_{w \in V} \text{broj}(w, c) + |V|}$$

Prior

$$P(\text{IT}) = \frac{3}{4}$$

$$P(\text{FR}) = \frac{1}{4}$$

Uvjetna vjerojatnost

$$P(\text{Italija|IT}) = \frac{5+1}{8+6} = \frac{6}{14} = \frac{3}{7}$$

$$P(\text{Pariz|IT}) = \frac{0+1}{8+6} = \frac{1}{14}$$

$$P(\text{Francuska|IT}) = \frac{0+1}{8+6} = \frac{1}{14}$$

$$P(\text{Italija|FR}) = \frac{1+1}{3+6} = \frac{2}{9}$$

$$P(\text{Pariz|FR}) = \frac{1+1}{3+6} = \frac{2}{9}$$

$$P(\text{Francuska|FR}) = \frac{1+1}{3+6} = \frac{2}{9}$$

Izbor klase

$$P(\text{IT}|d_5) \propto \frac{3}{4} \cdot \left(\frac{3}{7}\right)^3 \cdot \frac{1}{14} \cdot \frac{1}{14} \approx 0.0003$$

$$P(\text{FR}|d_5) \propto \frac{1}{4} \cdot \left(\frac{2}{9}\right)^3 \cdot \frac{2}{9} \cdot \frac{2}{9} \approx 0.0001$$

Naivni Bayes nije baš toliko naivan!

- Veoma brz, malo prostora zauzima
- robustan na nevažne osobine
 - nevažne osobine se međusobno poništavaju ne utječući na rezultat
- Dobar kod domena s mnogo jednakovražnih osobina
 - za razliku od stabla odluke koja pate od fragmentacije – pogotovo kod malo podataka
- Optimalan ako stoji pretpostavka o nezavisnosti: ako je pretpostavljena nezavisnost točna, onda se radi o optimalnom Bayesovom klasifikatoru
- dobra ovisna osnova za klasifikaciju teksta
- Postoje i drugi, precizniji klasifikatori

Uvod u obradu prirodnog jezika

6.7. Preciznost, odziv i F mjera (Precision, Recall and F measure)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

2 za 2 tablica slučaja

- 2 skupa podataka
 - točni entiteti
 - odabrani entiteti
- 4 moguća slučaja
 - TP – stvarno pozitivni (True positives)
 - FP – lažno pozitivni (False Positives)
 - FN – lažno negativni (False Negatives)
 - TN – stvarno negativni (True Negatives)

		točni entiteti	
		točno	nije točno
odabrani entiteti	odabрано	TP	FP
	nije odabрано	FN	TN

2 za 2 tablica slučaja: primjer

- Primjer

- TP – sustav je točno rekao za spam da je spam
- FP – sustav je pogrešno rekao za ne-spam da je spam
- FN – sustav je pogrešno rekao za spam da je ne-spam
- TN – sustav je točno rekao ne-spam da je ne-spam

	spam	ne-spam
spam	TP	FP
nije spam	FN	TN

Točnost

- Točnost (Acc = Accuracy) kao mjera

$$Acc = \frac{TP+TN}{TP+FP+TN+FN}$$

	točno	nije točno
odabrano	TP	FP
nije odabrano	FN	TN

Točnost

- Točnost kao mjera nije dobra za mali skup točnih podataka
- Recimo da promatramo 100000 Web stranica i samo 10 njih opisuje marku cipela.
- Ako napravimo najjednostavniji klasifikator koji za svaku stranicu kaže da ne opisuje marku cipela, onda ćemo dobiti veliku točnost

$$Acc = \frac{TP+TN}{TP+FP+FN+TN} =$$

$$\frac{0+99990}{0+0+10+99990} = \frac{99990}{100000} = 99.99\%$$

	marka cipela	ostalo
odabрано	0	0
nije odabрано	10	99990

Preciznost i odziv

- **Preciznost P :** % odabranih elemenata koji su točni
- **Odziv R :** % točnih elemenata koji su odabrani

	točno	nije točno
odabrano	TP	FP
nije odabrano	FN	TN

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

Preciznost i odziv

$$R = \frac{0}{10} = 0\%$$

	marka cipela	ostalo
odabrano	TP = 0	FP = 0
nije odabrano	FN = 10	TN = 99990

$$P = \frac{10}{40} = 25\%$$

$$R = \frac{10}{10} = 100\%$$

	marka cipela	ostalo
odabrano	TP = 10	FP = 30
nije odabrano	FN = 0	TN = 99960

Kombinirana mjera: F

- **F mjera:** Kombinirana mjera koja procjenjuje Preciznost/Odziv je (težinska harmonijska sredina)

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- Harmonijska sredina je konzervativni prosjek
- Obično se koristi balansirana F1 mjeru

$$\text{za } \beta = 1 \text{ (odnosno, } \alpha = \frac{1}{2} \text{)} F1 = \frac{2PR}{P+R}$$

Uvod u obradu prirodnog jezika

6.8. Evaluacija

Branko Žitko

prevedene od: Dan Jurafsky, Chris Manning

Više od dvije klase: skupovi binarnih klasifikatora

- "Bilo koja" viševrijednosna klasifikaciju
 - dokument može pripadati 0, 1, ili više klasa
- Za svaku klasu $c \in C$
 - napravi klasifikator γ_c kako bi razlikovali c od drugih klasa $c' \in C$
- Za dani testni dokument d ,
 - Evaluiraj pripadnost za svaku klasu koristeći svaku γ_c
 - d pripada svakoj klasi za koju γ_c vraća istinu

Više od dvije klase: skupovi binarnih klasifikatora

- "Jedna od" - viševrijednosna klasifikacija
 - Klase su međusobne isključive: svaki dokument pripada točno jednoj klasi
- Za svaku klasu $c \in C$
 - napravi klasifikator γ_c kako bi razlikovali c od drugih klasa $c' \in C$
- Za dani testni dokument d ,
 - Evaluiraj pripadnost za svaku klasu koristeći svaku γ_c
 - d pripada jednoj klasi za koju γ_c vraća najveću vjerojatnost

Evaluacija: jedna od - viševrijednosna klasifikacija

- Kategorizacija maila u 3 klase: Hitno, Normalno, Spam

		<i>Zlatni standard</i>		
		Hitno	Normalno	Spam
Hitno		8	10	1
Sustav	Normalno	5	60	50
	Spam	3	30	200
		$P_H = \frac{8}{8+10+1}$	$P_N = \frac{60}{5+60+50}$	$P_S = \frac{200}{3+30+200}$
		$R_H = \frac{8}{8+5+3}$	$R_N = \frac{60}{10+60+30}$	$R_S = \frac{200}{1+50+200}$

Evaluacija: jedna od - viševrijednosna klasifikacija

- Ako imamo više od jedne klase, kako se kombiniraju mjere u jednu mjeru?

	H	N	S
H	8	10	1
N	5	60	50
S	3	30	200

- Makro-prosjek:** izračunaj performanse za svaku klasu i onda prosjek

Hitno		Normalno		Spam	
	H	N	ne N	S	ne S
H	8	60	55	200	33
ne H	8	360	212	51	83
$P = \frac{8}{8+11} = 0.42$		$P = \frac{60}{60+55} = 0.52$		$P = \frac{200}{200+33} = 0.86$	

$$makroP = \frac{0.42 + 0.52 + 0.86}{3} = 0.60$$

Evaluacija: jedna od - viševrijednosna klasifikacija

- Ako imamo više od jedne klase, kako se kombiniraju mjere u jednu mjeru?

Hitno		Normalno		Spam	
	H	ne H	N	ne N	S
H	8	11	N	60	55
ne H	8	360	ne N	40	212
					S
					200
					33
				ne S	51
					83

- **Mikro-prosjek:** prikupi performanse svake klase, izračunaj tablicu slučaja, evaluiraj

	da	ne
da	268	99
ne	99	635

$$mikroP = \frac{268}{268+99} = 0.73$$

Razvojni testni skupovi i unakrsna validacija



- Mjera: $P/R/F1$ ili Acc
- Nevidjeni testni skup
 - izbjegći prekoračenja (ugađanje prema razvojnog skupu)
 - ponekad je testni skup (ili razvojni skup) malen
- Unakrsna validacija (cross validation) nad višestrukim podjelama
 - rukovanje greškama uzorkovanja nad više skupova podataka
 - skupljanje rezultata za svaku podjelu
 - izračunati prosjek rezultata

Razvojni testni skupovi i unakrsna validacija

Trening skup

Razvojni
skup

Testni skup

10-struka unakrsna validacija (10-fold cross validation)



Uvod u obradu prirodnog jezika

6.9. Testiranje statističke značajnosti

Branko Žitko

prevedene od: Dan Jurafsky, Chris Manning

Testitanje statističke značajnosti

Kako znamo koji klasifikator je bolji?

Za dane:

- Klasifikatore A i B
- Metriku M : $M(A, x)$ je performansa od A na testnom skupu x
- $\delta(x) = M(A, x) - M(B, x)$
- Želimo znati je li $\delta(x) > 0$ (A je bolji od B)
- $\delta(x)$ – **veličina učinka** (effect size)
- Ako se pokaže $\delta(x)$ pozitivnim, to može biti slučajnost samo za ovaj testni skup x .

Testitanje hipoteze

Dvije hipoteze

- nul-hipoteza: A nije bolji od B $H_0: \delta(x) \leq 0$
 - hipoteza: A je bolji od B $H_1 : \delta(x) > 0$
-
- Želimo isključiti H_0
 - Stvari se slučajna varijabla X za sve testne skupove
 - Pitamo, koliko je vjerojatno, ako je H_0 istina, da ćemo među testnim skupovima naići na vrijednost $\delta(x)$ koju promatramo.
 - Formaliziramo kao p-vrijednost:
$$P(\delta(X) \geq \delta(x) \mid H_0 \text{ je istina})$$

Testitanje hipoteze

$$P(\delta(X) \geq \delta(x) \mid H_0 \text{ je istina})$$

p-vrijednost je vjerojatnost da ćemo naići na $\delta(x)$ uz pretpostavku da A nije bolji od B .

Ako je $\delta(x)$ velik (A ima $F1 = 0.9$, B ima $F1 = 0.2$)

- to bi bilo iznenadnje
- uz činjenicu da je H_0 istinita.
(p-vrijednost niska)

Ako je $\delta(x)$ malen (A ima $F1 = 0.2$, B ima $F1 = 0.9$)

- to nebi bilo iznenadnje
- uz pretpostavku da je H_0 istinita i
- da A zaista nije bolji od B .
(p-vrijednost visoka)

Testitanje hipoteze

$$P(\delta(X) \geq \delta(x) \mid H_0 \text{ je istina})$$

Vrlo mala p-vrijednost znači da je razlika koju smo uočili malo vjerojatna pod nul-hipotezom.

Odbacujemo nul-hipotezu.

Veoma mala: 0.05 ili 0.01

Rezultat (A je bolji od B) je **statistički značajan** ako promatrani δ ima vjerojatnost koja je manja od praga.

Testitanje hipoteze

Koristi se neparametarsko testiranje temeljeno na uzorkovanju:

- umjetno stvaramo mnogo verzija postavki eksperimenta

Na primjer:

- kreiramo veliki broj testnih skupova x'
- za svaki izračunamo $\delta(x')$
- dobivamo distribuciju
- odaberemo prag (npr. 0.01)
- ako za 99% testnih skupova vrijedi $\delta(x) > \delta(x')$
- onda zaključujemo da je δ našeg testnog skupa prava δ , a ne umjetna

Upareno Bootstrap testiranje

Može se primijeniti na bilo koju metriku (Acc, P, R, F1)

Bootstrap znači iterativno uzimati veliki broj malih uzoraka sa zamjenom iz originalnog velikog uzorka.

Upareno Bootstrap testiranje

Jednostavan primjer:

Klasifikacija teksta s testnim skupom x od 10 dokumenata

Rezultati sustava A i B nad x s 4 moguća ishoda:

AB - oboje točno

AB - oboje pogrešno

AB - A točan, B pogrešan

AB - A pogrešan, B točan

	1	2	3	4	5	6	7	8	9	10	A%	B%	δ
x	AB	0.70	0.50	0.20									

Upareno Bootstrap testiranje

Sada stvorimo, recimo $b = 10000$ virtualnih testnih skupova
gdje svaki sadrži $n=10$ dokumenata

Virtualni testni skup $x^{(i)}$ dobijemo tako da iz originalnog testnog skupa x slučajnim odabirom s ponavljanjem uzmemos rezultat testiranja

Upareno Bootstrap testiranje

Imamo distribuciju!

Možemo provjeriti koliko često A ima **slučajnu** prednost

Uz pretpostavku H_0 očekujemo $\delta(x^{(i)}) = 0$

Prebrojimo koliko puta $\delta(x^{(i)})$ prelazi 0 u odnosu na $\delta(x)$

$$\text{p-value} = \frac{1}{b} \sum_{i=1}^b \mathbb{1}(\delta(x^{(i)}) - \delta(x) \geq 0)$$

Upareno Bootstrap testiranje

Međutim, uzorke nismo izvlačili iz distribucije čija je srednja vrijednost 0.

Koristili smo originalni testni skup x koji je pristran (0.20) u korist sustava A.

p-vrijednost stoga računamo koliko često $\delta(x^{(i)})$ premašuje očekivanu vrijednost $\delta(x)$ s $\delta(x)$ ili više:

$$\text{p-value} = \frac{1}{b} \sum_{i=1}^b 1(\delta(x^{(i)}) - \delta(x) \geq \delta(x))$$

$$\text{p-value} = \frac{1}{b} \sum_{i=1}^b 1(\delta(x^{(i)}) \geq 2\delta(x))$$

Uvod u obradu prirodnog jezika

7.1. Što je analiza sentimenta

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Pozitivna ili negativna kritika filma?

- nevjerljivo razočarenje 
- pun otkačenih likova i bogato primijenjena satira, s nekim velikim zapletima radnje 
- ovo je najveća ekscentrična komedija ikad snimljena 
- ovo je jadno, najgori dio je definitivno scena boksa 

Google Product Search



HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner

\$89 online, \$100 nearby 377 reviews

September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 sh

Reviews

Summary - Based on 377 reviews

1 star 2 3 4 stars 5 stars

What people are saying

ease of use		"This was very easy to setup to four computers."
value		"Appreciate good quality at a fair price."
setup		"Overall pretty easy setup."
customer service		"I DO like honest tech support people."
size		"Pretty Paper weight."
mode		"Photos were fair on the high quality mode."
colors		"Full color prints came out with great quality."

Bing Shopping

HP Officejet 6500A E710N Multifunction Printer

[Product summary](#) [Find best price](#) **Customer reviews** [Specifications](#) [Related items](#)



\$121.53 - \$242.39 (14 stores)

Compare

Average rating (144)



Most mentioned



Show reviews by source

Best Buy (140)
CNET (5)
Amazon.com (3)

Zašto analizirati sentiment

- Film: je li kritika pozitivna ili negativna?
- Proizvodi: što ljudi misle o novom štampaču?
- Javni sentiment: koliko je povjerenje potrošača?
- Politika: što ljudi misle o kandidatu?
- Predikcija: predviđanje rezultata izbora ili trendova na tržištu

Konotacija

- Riječi osim značenja imaju i konotaciju.
- Konotacija je sentiment koju riječ ili fraza posjeduje
- Možemo li izgraditi leksički resurs kojim se reprezentiraju konotacije?
- i iskoristiti ga u obradi prirodnog jezika?

Shrinerova tipologija afektivnih stanja

- **Emocija:** evaluacija glavnih događaja
 - ljut, tužan, vesel, strah, sram, ponosan, ushićen
- **Raspoloženje:** razlikovanje ne izazvanih, dugotrajnih promjena subjektivnog osjeta
 - vesel, tmuran, razdražljiv, bezvoljan, depresivan, poletan
- **Međuljudski odnosi:** afektivni stavovi prema drugoj osobi
 - prijateljski, koketirajući, hladan, topao, podržavajući, prijezirni
- **Stavovi:** postojeća, afektivno obojena vjerovanja, dispozicije prema predmetima ili osobama
 - naklonost, ljubav, mržnja, vrijedanje, želja
- **Osobne crte:** stabilna stanja osobe i tipično ponašanje
 - nervozan, tjeskoban, bezobziran, mrzovoljan, neprijateljski, ljubomoran

Analiza sentimenta

- Najjednostavniji zadatak
 - je li stav teksta pozitivan ili negativan
- Složeniji zadatak
 - Ocijeni stav u tekstu od 1 do 5
- Napredni zadatak
 - odredi cilj, izvor ili kompleksne vrste stavova

Uvod u obradu prirodnog jezika

7.2. Osnovni algoritam

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Klasifikacija sentimenta kritike filma

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. ACL, 271-278

- Detekcija polariteta
 - je li IMDB kritika filma pozitivna ili negativna
- Podaci: *Polarity Data 2.0*
 - <http://www.cs.cornell.edu/people/pabo/movie-review-data>

IMDB podaci u Pang i Lee bazi podataka



when _star wars_ came out some twenty years ago , the image of traveling throughout the stars has become a commonplace image . [...] when han solo goes light speed , the stars change to bright lines , going towards the viewer in lines that converge at an invisible point . cool .

october sky offers a much simpler image—that of a single white dot , traveling horizontally across the night sky . [. . .]



"snake eyes" is the most aggravating kind of movie : the kind that shows so much potential then becomes unbelievably disappointing . it's not just because this is a brian depalma film , and since he's a great director and one who's films are always greeted with at least some fanfare . and it's not even because this was a film starring nicolas cage and since he gives a brauvara performance , this film is hardly worth his talents .

Osnovni algoritam (adaptacija od Pang i Lee)

- Tokenizacija
- Ekstrakcija osobina
- Klasifikacija korištenjem različitih klasifikatora
 - Naivni Bayes
 - MaxEnt
 - SVN

Problemi tokenizacije sentimenta

- Problemi HTML i XML oznaka
- Twitter oznake (imena, hash oznake)
- Kapitalizacija (sačuvaj riječi koje su pisane u velikim znakovima)
- Brojevi telefona, datumi
- Emotikon

```
[<>] ?                                # kapa, obrve
[ : ; = 8 ]                               # oči
[ \-o\*\' ] ?                            # nos
[ \ ) \ ] \ ( \ [ dDpP / \ : \ } \ { @ \ | \ \ ]
|                                         # usta
[ \ ) \ ] \ ( \ [ dDpP / \ : \ } \ { @ \ | \ \ ]
[ \-o\*\' ] ?                            # nos
[ : ; = 8 ]                               # oči
[ <> ] ?                                # kapa, obrve
```

Christopher Potts tokenizator sentimenta:

<http://sentiment.christopherpotts.net/code-data/happyfuntokenizing.py>

Brendan O'Connor twitter tokenizator:

<https://github.com/brendano/tweetmotif>

Ekstrakcija osobina za klasifikaciju sentimenta

- Kako se odnositi s negacijom?
 - Ne sviđa mi se ovaj film
 - vs.
 - Zaista mi se sviđa ovaj film
- Koje riječi koristiti?
 - samo pridjeve
 - sve riječi
 - korištenje svih riječi se pokazuje boljim

Negacija

Das, Sanjiv and Mike Chen. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA).

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.

- Dodaj NE_ svakoj riječi između negacije i sljedeće interpunkcije:

Ne sviđa mi se ovaj film, ali ...



Ne NE_sviđa NE_mi NE_se NE_ovaj NE_film, ali ...

Naivni Bayes: podsjetnik

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i \in \text{pozicije}} P(w_i | c)$$

$$\hat{P}(w_i | c) = \frac{\text{broj}(w_i, c) + 1}{\sum_{w \in V} (\text{broj}(w, c) + 1)}$$

Binarni naivni Bayes

- Ideja:
 - za sentiment (i vjerojatno za druge domene klasifikacije teksta)
 - **pojavljivanje riječi** može značiti više od frekvencije riječi
 - pojavljivanje riječi "fantastično" govori nam mnogo
 - činjenica da se riječ "fantastično" pojavljuje 5 puta ne govori nam ništa više
 - Binarni (Booleov) naivni Bayes
 - sažima broj riječi u svakom dokumentu na 1

Binarni multinominalni naivni Bayes: učenje

UČENJE(D, C)

za svaku klasu $c \in C$

$N_{doc} \leftarrow$ broj dokumenata iz D

$N_c \leftarrow$ broj dokumenata iz D klase c

$prior[c] \leftarrow \frac{N_c}{N_{doc}}$

$V \leftarrow$ riječnik dokumenata iz D

$megadoc[c]$ proširi s jedinstvenim riječima iz $d \in D$
koji su klase c

za svaku riječ $w \in V$

$broj[w, c] \leftarrow$ broj pojavljivanja od w u $megadoc(c)$

$izvjesnost[w, c] \leftarrow \frac{broj[w, c] + 1}{\sum_{w' \in V} (broj[w', c] + 1)}$

vrati $prior, izvjesnost, V$

Binarni naivni Bayes na testnom dokumentu d

- Izbaci sve duplike riječi iz d
- Izračunaj NB koristeći istu formulu

$$c_{NB} = \operatorname{argmax}_{c \in C} P(c) \prod_{i \in pozicije} P(w_i | c)$$

Normalni vs. Binarni multinomialni naivni Bayes

- Normalni

	Dokument	Riječi	Klasa
Treniranje	d ₁	Italija Rim Italija	IT
	d ₂	Italija Italija Firenca	IT
	d ₃	Italija Ankona	IT
	d ₄	Pariz Francuska Italija	FR
Test	d ₅	Italija Italija Italija Pariz Francuska	?

- Binarni

	Dokument	Riječi	Klasa
Treniranje	d ₁	Italija Rim	IT
	d ₂	Italija Firenca	IT
	d ₃	Italija Ankona	IT
	d ₄	Pariz Francuska Italija	FR
Test	d ₅	Italija Pariz Francuska	?

Binarni naivni Bayes

- B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.
- V. Metsis, I. Androutsopoulos, G. Palioras. 2006. Spam Filtering with Naive Bayes – Which Naive Bayes? CEAS 2006 - Third Conference on Email and Anti-Spam.
- K.-M. Schneider. 2004. On word frequency information and negative evidence in Naive Bayes text classification. ICANLP, 474-485.
- JD Rennie, L Shih, J Teevan. 2003. Tackling the poor assumptions of naive bayes text classifiers. ICML 2003

- Binarni klasifikator se pokazuje boljim
 - On se razlikuje od multivarijabilnog Bernoullijevog naivnog Bayesa (MBNB)
 - MNBN ne radi dobro kod analize sentimenta i drugih zadataka obrade teksta
- Druga mogućnost
 - $\log(freq(w))$
- MaxEnt i SVN teže boljem učinku od naivnog Bayesa

Problemi: što čini kritikom teškom za obradu?

- Suptilnost
 - Kritika parfema

"Ako ovo čitate, jer se radi o vašem dragom mirisu, molimo vas da ga nosite isključivo kod kuće, a prozore zalijepite trakama"
 - Kritika glumice

"Ona iznosi sve emocije na skali A do B"

Oprečna očekivanja i efekt redoslijeda

- Ovaj film bi trebao biti **briljantan**. Izgleda da ima **sjajnu radnju**, glumci su **prvoklasni**, kao i sporedne uloge. Stallone pokušava ostvariti **dobru** izvedbu. Međutim, **ne može je održati**.
- Kao i obično Keanu Reeves nije ništa specijalan, ali iznenadujuće je što ni **vrlo talentirani** Laurence Fishbourne **nije toliko dobar**. Iznenaden sam.

Uvod u obradu prirodnog jezika

7.3. Leksikon sentimenta

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

General Inquirer

- Stranica:
 - <http://www.wjh.harvard.edu/~inquirer>
- Lista kategorija:
 - <http://www.wjh.harvard.edu/~inquirer/homecat.htm>
- Tablica:
 - <http://www.wjh.harvard.edu/~inquirer/inquirerbasic.xls>
- Kategorije:
 - Pozitivna (1915 riječi) i Negativna (2291 riječi)
 - Snažno - Slabo, Aktivno - Pasivno, Naglašeno - Nenametljivo
 - Ugoda, Bol, Vrlina, Porok, Motivacija, Kognitivna orijentacija, itd.
- Slobodan za istraživačke svrhe

Philip J. Stone, Dexter C Dunphy, Marshall S. Smith, Daniel M. Ogilvie. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press

General Inquirer

Positive	admire, amazing, assure, celebration, charm, eager, enthusiastic, excellent, fancy, fan-tastic, frolic, graceful, happy, joy, luck, majesty, mercy, nice, patience, perfect, proud, rejoice, relief, respect, satisfactorily, sensational, super, terrific, thank, vivid, wise, won- derful, zest
Negative	abominable, anger, anxious, bad, catastrophe, cheap, complaint, condescending, deceit, defective, disappointment, embarrass, fake, fear, filthy, fool, guilt, hate, idiot, inflict, lazy, miserable, mourn, nervous, objection, pest, plot, reject, scream, silly, terrible, unfriendly, vile, wicked

LIWC (Linguistic Inquiry and Word Count)

- Stranica:
 - <http://www.liwc.net/>
- 2300 riječi, preko 70 klasa
- **Afektivni procesi**
 - Negativne emocije (loš, čudan, mrzi, problem, naporan)
 - Pozitivne emocije (ljubav, lijepo, slatko)
- **Kognitivni procesi**
 - Probni (možda, valjda, prepostavljam)
 - Inhibicija (blokirati, ograničiti)
- **Zamjenice, Negacija (ne, nikad), Kvantifikatori (neki, mnogi)**
- \$30 do \$90

Pennebaker, J.W., Booth, R.J., & Francis, M.E. (2007). Linguistic Inquiry and Word Count: LIWC 2007. Austin, TX

LIWC (Linguistic Inquiry and Word Count)

Positive emotion	Negative emotion	Insight	Inhibition	Family	Negate
appreciat*	anger*	aware*	avoid*	brother*	aren't
comfort*	bore*	believe	careful*	cousin*	cannot
great	cry	decid*	hesitat*	daughter*	didn't
happy	despair*	feel	limit*	family	neither
interest	fail*	figur*	oppos*	father*	never
joy*	fear	know	prevent*	grandf*	no
perfect*	griev*	knew	reluctan*	grandm*	nobod*
please*	hate*	means	safe*	husband	none
safe*	panic*	notice*	stop	mom	nor
terrific	suffers	recogni*	stubborn*	mother	nothing
value	terrify	sense	wait	niece*	nowhere
wow*	violent*	think	wary	wife	without

Primjer 5 od 73 leksičke kategorije u LIWC

* prethodne riječi su prefiksi i sve riječi s ovim prefiksom su također uključene

MPQA Subjectivity Cues Lexicon

- Stranica:
 - http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
- 6885 riječi od 8221 lema
 - 2718 pozitivnih
 - 4912 negativnih
- Svaka riječ ima oznaku intenziteta (jak, slab)
- GNU GPL

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proc. of HLT-EMNLP-2005.

Riloff and Wiebe (2003). Learning extraction patterns for subjective expressions. EMNLP-2003.

Hu and Liu Opinion Lexicon

- Bing Liu stranica za Opinion Mining
 - <http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar>
- 6786 riječi
 - 2006 pozitivne
 - 4783 negativne

SentiWordNet

- Stranica:
 - <http://sentiwordnet.isti.cnr.it/>
- Svi WordNet-ovi skupovi sinonima imaju oznake stupnja pozitivnosti, negativnosti i neutralnosti (objektivnosti)
 - [procijenjen(J,3)] "može biti izračunat ili procijenjen"
 - Poz 0 Neg 0 Obj 1
 - [smatrati(J,1)] "smatra se dobrom učenikom"
 - Poz 0.75 Neg 0 Obj 0.25

Nesuglasice među polaritetima u leksikonima

Christopher Potts, Sentiment Tutorial, 2011

<http://sentiment.christopherpotts.net/lexicons.html>

	Opinion Lexicon	General Inquirer	SentiWordNet	LIWC
MPQA	33/5402 (0.6%)	49/2867 (2%)	1127/4214 (27%)	12/363 (3%)
Opinion Lexicon		32/2411 (1%)	1004/3994 (25%)	9/403 (2%)
General Inquirer			520/2306 (23%)	1/204 (0.5%)
SentiWordNet				174/694 (25%)
LIWC				

Analiza polariteta svake riječi u IMDB kritici

- Koliko vjerojatno će se svaka riječ pojaviti u svakoj klasi sentimenta?
 - Prebroji "loš" ("bad") u kritici od 1-zvijezde, 2-zvijezde, 3-zvijezde, itd.

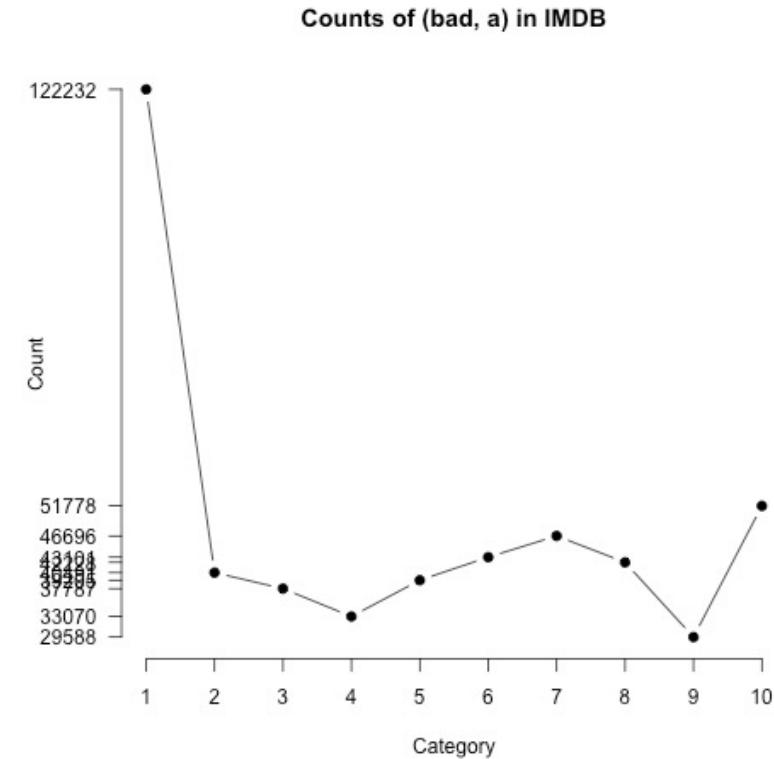
- Ali ne može se koristiti sirovo prebrojavanje:

- Umjesto **izglednosti**

$$P(w|c) = \frac{f(w, c)}{\sum_{w' \in c} f(w', c)}$$

- Potrebno je riječi učiniti međusobno usporedivim
 - **Skalirana izglednost**

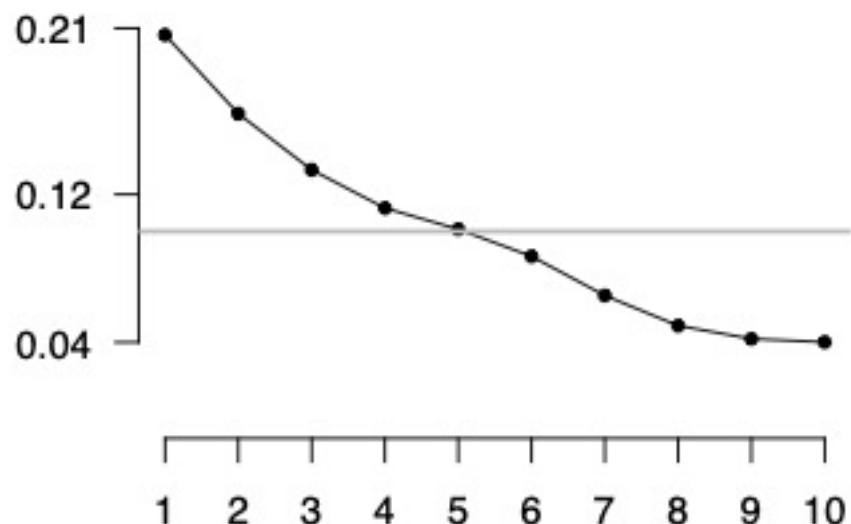
$$Pott(w|c) = \frac{P(w|c)}{\sum_c P(w|c)}$$



Potts, Christopher. 2011. On the negativity of negation. SALT 20, 636-659.

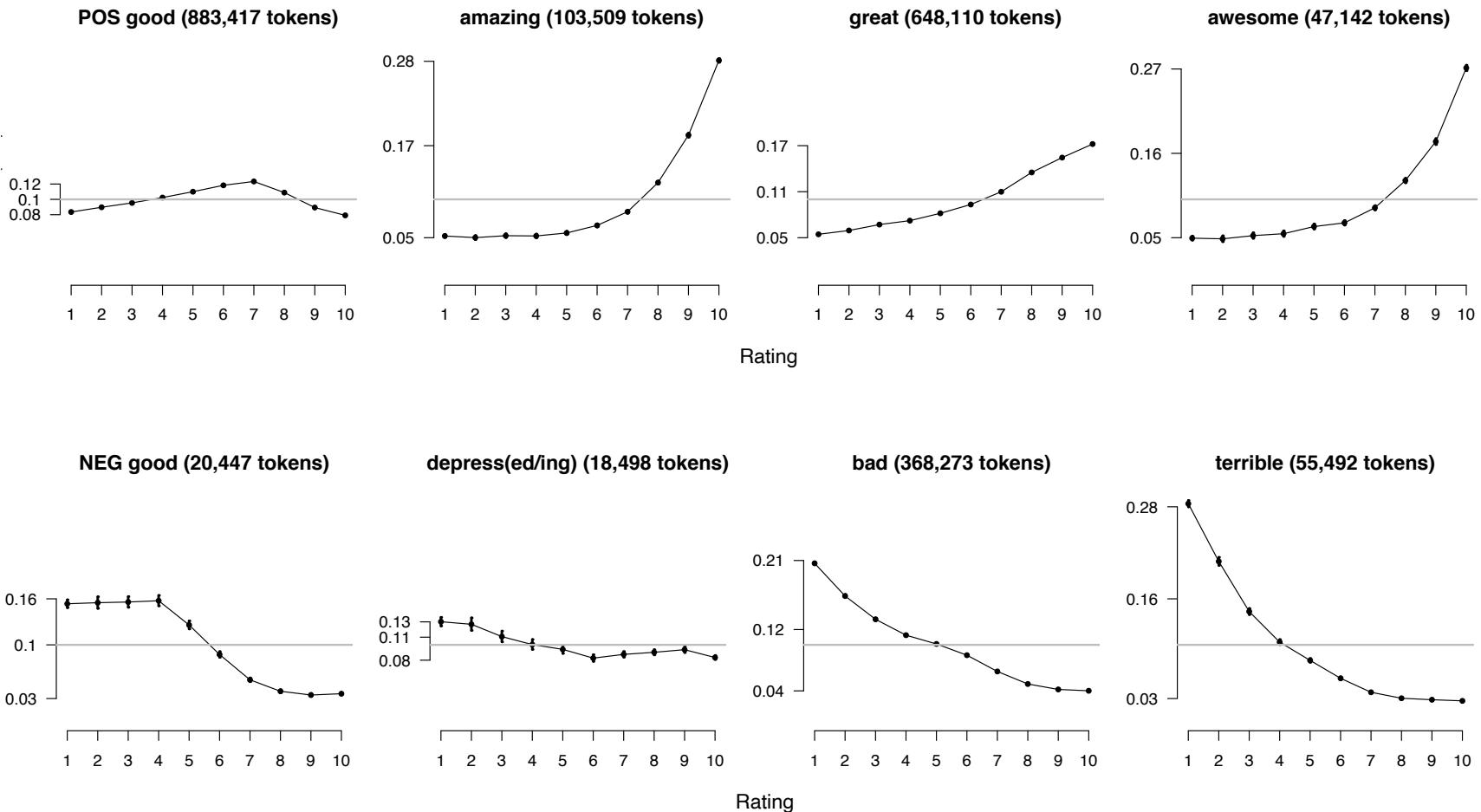
Analiza polariteta svake riječi u IMDB kritici

- IMDB kritika ima 10 klasa, stoga svakoj riječi se dodjeljuje vektor.
- Vektor od "bad" je
[0.21 0.14 0.13 0.11 0.10 0.09 0.07 0.06 0.05 0.04]



Analiza polariteta svake riječi u IMDB kritici

Skalirana vjerodostojnost $\frac{P(W|C)}{P(w)}$



Potts, Christopher. 2011. On the negativity of negation. SALT 20, 636-659.

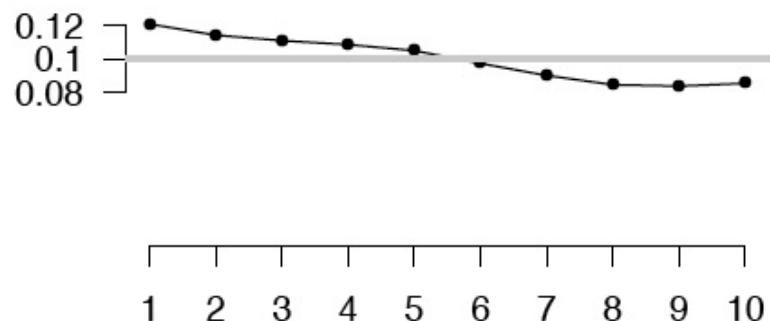
Ostale osobine sentimenta

- Je li logička negacija (ne, nije) povezana s negativnim sentimentom?
- Pottsov eksperiment:
 - Izbroji negacije (ne, nije, nikad) u kritikama
 - Regresivno usporedi s ocjenom kritike

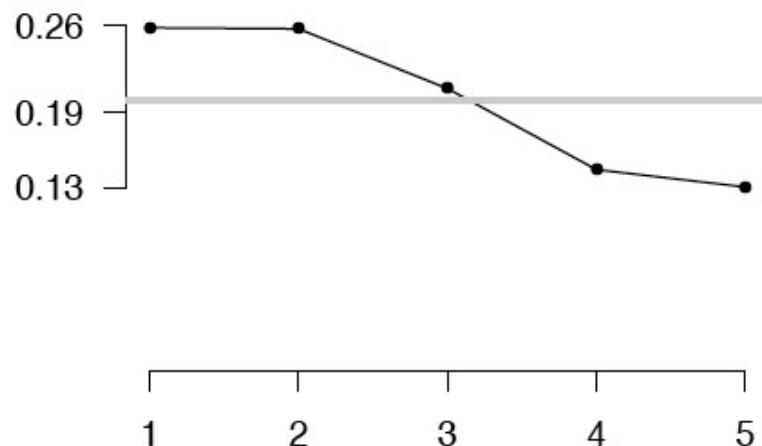
Potts 2011 rezultati

više negacije u negativnom sentimentu

IMDB (4,073,228 tokens)



Five-star reviews (846,444 tokens)



Uvod u obradu prirodnog jezika

7.4. Učenje leksikona sentimenta

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Polunadzirano učenje leksikona

- Koristiti malu količinu informacija
 - Nekoliko označenih primjera
 - Nekoliko ručno izrađenih uzoraka
- Podizanje (bootstrap) leksikona

Hatzivassiloglou i McKeown ideja za identifikaciju polariteta

- Pridjevi spojeni s "i" imaju isti polaritet
 - Pošten i odan
 - Pokvaren i svirep
- Pridjevi spojeni s "ali" nemaju isti polaritet
 - Pošten ali strog

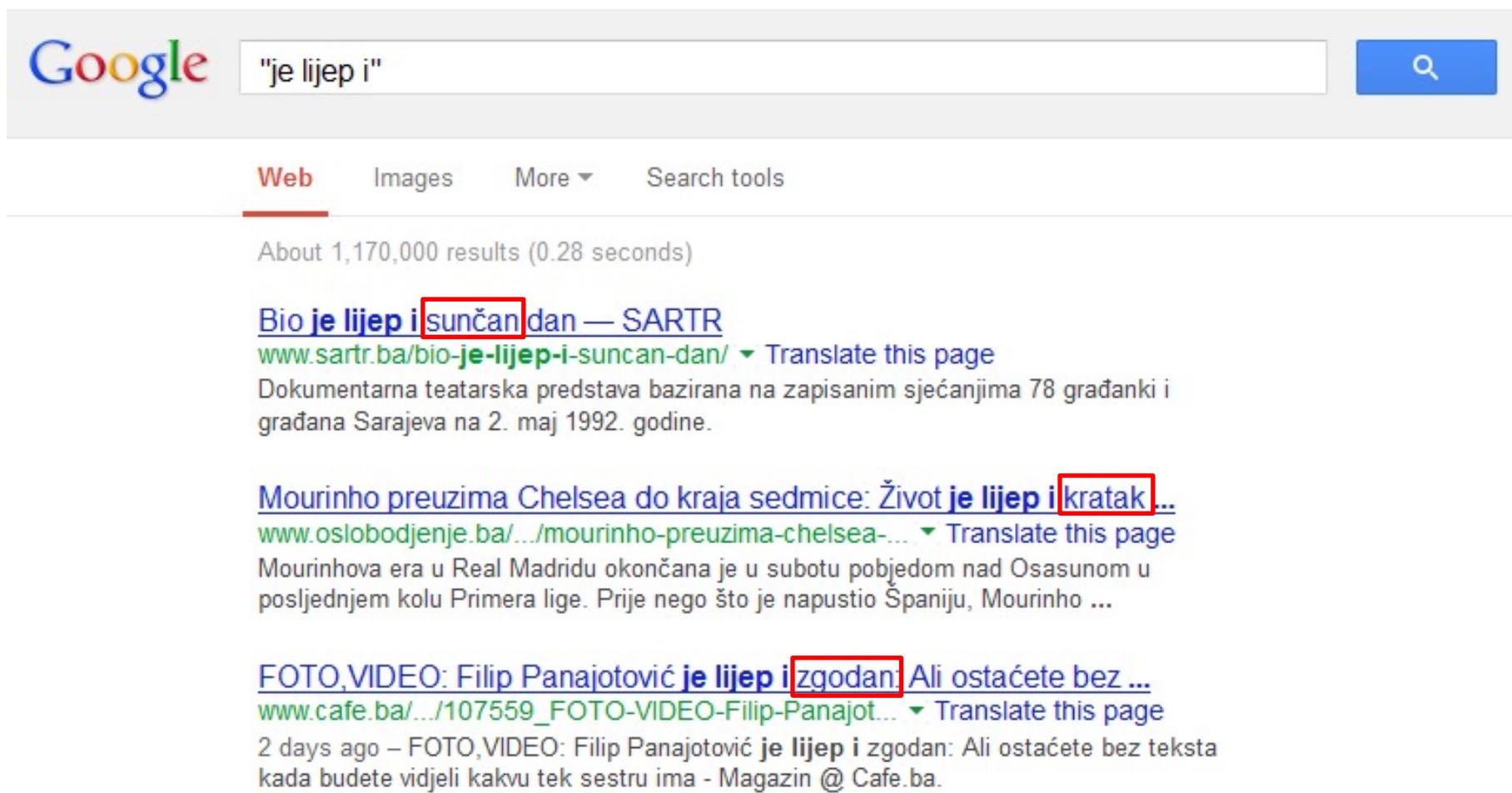
Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. ACL, 174–181

Hatzivassiloglou i McKeown 1997: 1. Korak

- Označi početni skup od 1336 pridjeva (iz WSJ korpusa od 21 miliona riječi)
 - 657 pozitivnih
 - odgovarajuće, centralno, pametno, poznato, inteligentno, izvanredno, znano, osjetljivo, vitko, uspješno...
 - 679 negativnih
 - zarazno, pijano, neznano, koščato, bezvoljno, primitivno, uznemirujuće, neriješeno, zlobno...

Hatzivassiloglou i McKeown 1997: 2. Korak

- Proširi početni skup na spojene pridjeve



Google search results for the query "je lijep i". The results page shows three search results, each with a red box highlighting the word "i" in the phrase "je lijep i".

"je lijep i"

Web Images More Search tools

About 1,170,000 results (0.28 seconds)

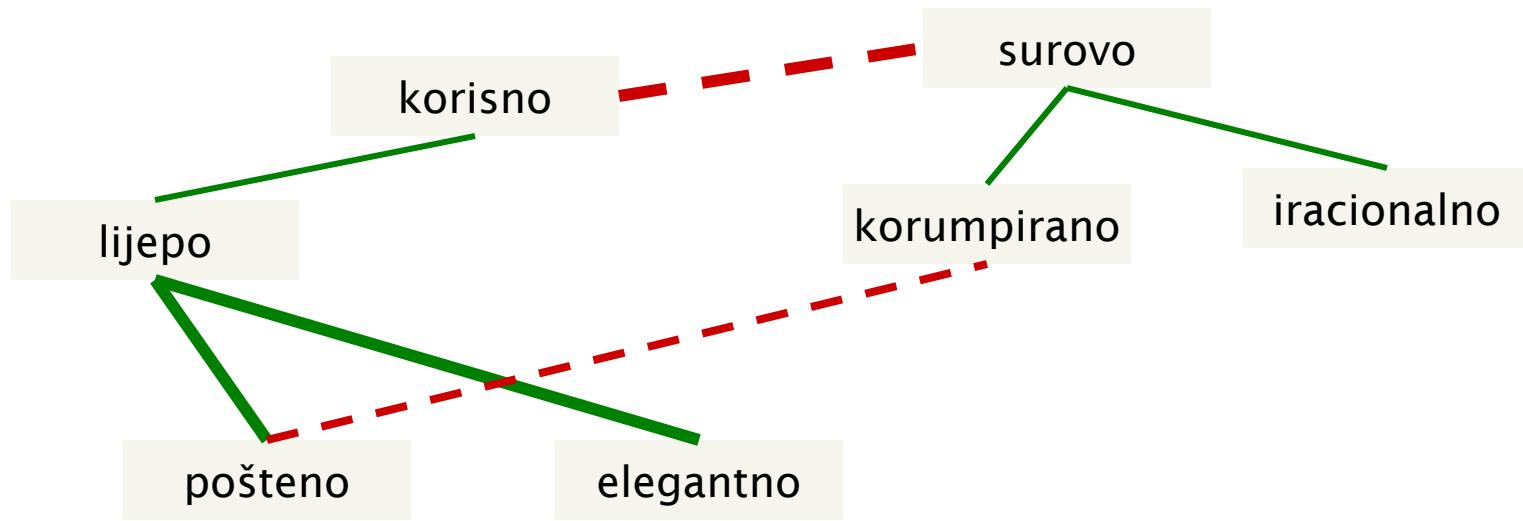
Bio je lijep i sunčan dan — SARTR
www.sartr.ba/bio-je-lijep-i-suncan-dan/ ▾ Translate this page
Dokumentarna teatarska predstava bazirana na zapisanim sjećanjima 78 građanki i građana Sarajeva na 2. maj 1992. godine.

Mourinho preuzima Chelsea do kraja sedmice: Život je lijep i kratak ...
www.oslobodjenje.ba/.../mourinho-preuzima-chelsea-... ▾ Translate this page
Mourinhova era u Real Madridu okončana je u subotu pobjedom nad Osasunom u posljednjem kolu Primera lige. Prije nego što je napustio Španiju, Mourinho ...

FOTO,VIDEO: Filip Panajotović je lijep i zgodan: Ali ostaćete bez ...
www.cafe.ba/.../107559_FOTO-VIDEO-Filip-Panajot... ▾ Translate this page
2 days ago – FOTO,VIDEO: Filip Panajotović je lijep i zgodan: Ali ostaćete bez teksta kada budete vidjeli kakvu tek sestru ima - Magazin @ Cafe.ba.

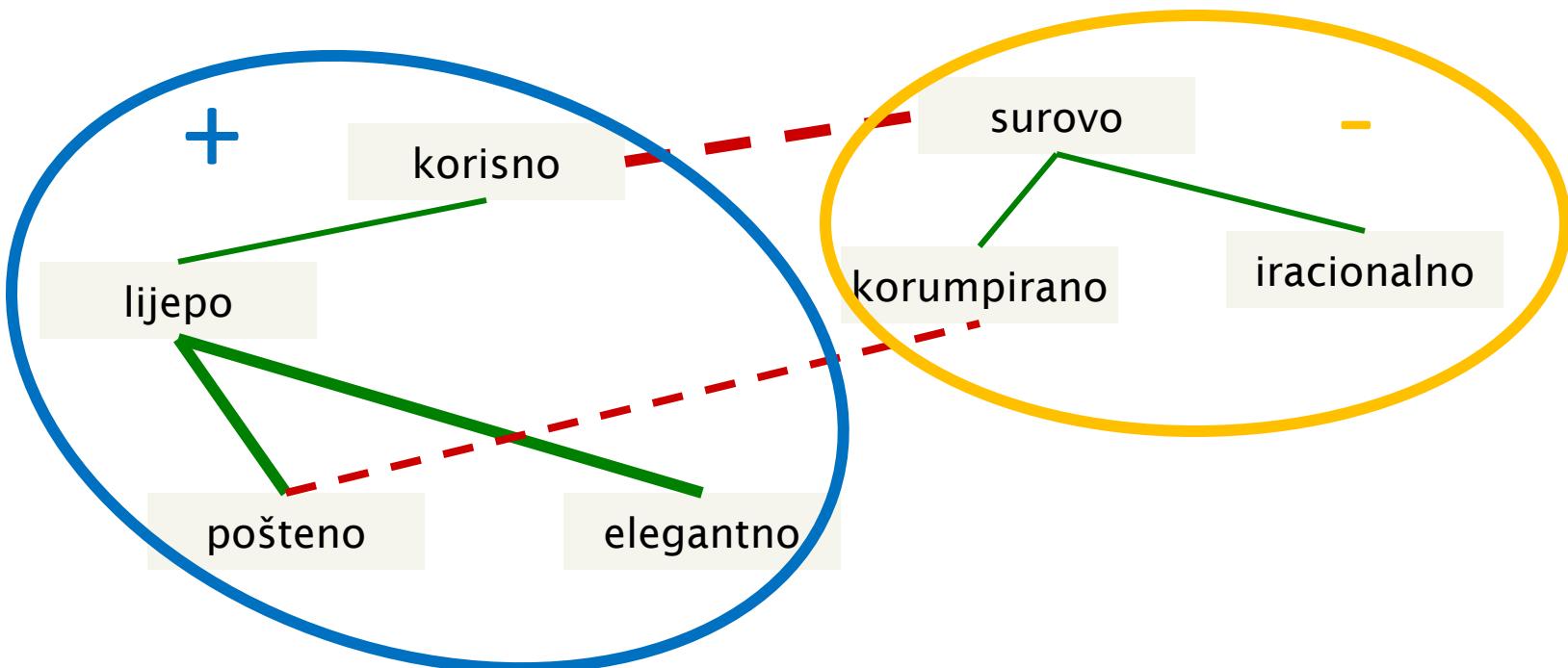
Hatzivassiloglou i McKeown 1997: 3. Korak

- Nadzirani klasifikator pridružuje "sličnost polariteta" svakom paru riječi



Hatzivassiloglou i McKeown 1997: 4. Korak

- Grupiranje grafa



Izlaz iz leksikona polarnosti

- Pozitivno
 - hrabar odlučujuća uznemirujuće velikodušni dobri pošteni važna velika zrela strpljivi mirni pozitivno ponosni zvuk poticanje jednostavan neobično snažno talentirana duhovita ...
- Negativno
 - dvosmislena oprezni cinični izbjegavajući štetna licemjerno neučinkovit nesigurno iracionalno neodgovorno maloljetnika gorljivi ugodno osvrće rizično sebični zamorno nepotkrijepljene ranjiva rastrošna ...

Izlaz iz leksikona polarnosti

- Pozitivno
 - hrabar odlučujuća **uznemirujuće** velikodušni dobri poštenu važna velika zrela strpljivi mirni pozitivno ponosni zvuk poticanje jednostavan neobično **stran** talentirana duhovita ...
- Negativno
 - dvosmislena **oprezni** cinični izbjegavajući štetna licemjerno neučinkovit nesigurno iracionalno neodgovorno maloljetnika gorljivi **ugodno** osvrće rizično sebični zamorno nepotkrijepljene ranjiva rastrošna ...

Turney-ov algoritam

1. Izvuci leksikon fraza iz kritika
2. Nauči polaritet za svaku frazu
3. Ocjeni kritiku temeljem srednje vrijednosti polarnosti njenih fraza

Turney (2002): Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews

Izvlačenje fraze od dvije riječi koje imaju pridjev

Prva riječ	Druga riječ	Treća riječ (nije izvučena)
JJ	NN ili NNS	bilo što
RB, RBR, RBS	JJ	nije NN ni NNS
JJ	JJ	nije NN ni NNS
NN ili NNS	JJ	nije NN ni NNS
RB, RBR, or RBS	VB, VBD, VBN, VBG	bilo što

J – pridjev

N – imenica

R – prilog

V - glagol

Kako izmjeriti polaritet fraza

- Pozitivne fraze
 - pozitivne fraze se češće ko-javljuju s "dobro"
 - negativne fraze se češće ko-javljuju s "loše"
- ali kako izmjeriti ko-javljanje?

Srednji uzajamni sadržaj informacije

- Uzajamni sadržaj informacije dviju slučajnih varijabli X i Y
- mjeri količinu informacija koju imamo o pojavljivanju jedne riječi, ako imamo informacije o pojavljivanju druge riječi

$$I(X, Y) = \sum_x \sum_y P(x, y) \log_2 \frac{P(x, y)}{P(x)P(y)}$$

- Srednji uzajamni sadržaj informacije
(Pointwise mutual information)
- koliko često se događaji X i Y ko-javljujaju ako su nezavisni?

$$PMI(X, Y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

Srednji uzajamni sadržaj informacije

- Srednji uzajamni sadržaj informacije između dvije riječi
 - koliko često se riječi w_1 i w_2 ko-javljuju ako su nezavisni?

$$PMI(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

Kako procijeniti srednji uzajamni sadržaj informacije?

- Corpus Query System (<http://filip.ffzg.hr/bonito2/>)
 - $P(w)$ procjenjuje se kao $\frac{\text{broj}(w)}{|V|}$
 - $P(w_1, w_2)$ procjenjuje se kao broj $\frac{\text{broj}(w_1 \text{ POKRAJ } w_2)}{|V|^2}$

$$PMI(w_1, w_2) = \log_2 \frac{\text{broj}(w_1 \text{ POKRAJ } w_2)}{\text{broj}(w_1)\text{broj}(w_2)}$$

Da li se fraza pojavljuje više uz "izvrstan" ili "jadan"?

$$Polarnost(fraza) = PMI(fraza, \text{dobro}) - PMI(fraza, \text{loše})$$

$$= \log_2 \left(\frac{\text{broj}(fraza \text{ POKRAJ } \text{ dobro})}{\text{broj}(fraza) \text{ broj(dobro)}} \right) - \log_2 \left(\frac{\text{broj}(fraza \text{ POKRAJ } \text{ loše})}{\text{broj}(fraza) \text{ broj(loše)}} \right)$$

$$= \log_2 \left(\frac{\text{broj}(fraza \text{ POKRAJ } \text{ dobro})}{\text{broj}(fraza) \text{ broj(dobro)}} \right) - \log_2 \left(\frac{\text{broj}(fraza \text{ POKRAJ } \text{ loše})}{\text{broj}(fraza) \text{ broj(loše)}} \right)$$

$$= \log_2 \left(\frac{\text{broj}(fraza \text{ POKRAJ } \text{ dobro}) \text{broj}(fraza) \text{ broj(loše)}}{\text{broj}(fraza) \text{ broj(dobro)} \text{broj}(fraza \text{ POKRAJ } \text{ loše})} \right)$$

$$= \log_2 \left(\frac{\text{broj}(fraza \text{ POKRAJ } \text{ dobro}) \text{ broj(loše)}}{\text{broj}(fraza \text{ POKRAJ } \text{ loše}) \text{broj(dobro)}} \right)$$

Fraze iz pozitivne kritike

Fraza	POS	Polarnost
online usluga	JJ NN	2.8
online iskustvo	JJ NN	2.3
izravna uplata	JJ NN	1.3
lokalna grana	JJ NN	0.42
...		
niske pristojbe	JJ NNS	0.33
prava usluga	JJ NN	-0.73
druga banka	JJ NN	-0.85
nezgodno nalazi	JJ NN	-1.5
Prosjek		0.32

Fraze iz negativne kritike

Fraza	POS	Polarnost
izravna uplata	JJ NNS	5.8
online web	JJ NN	1.9
veoma praktično	RB JJ	1.4
...		
virtualni monopol	JJ NN	-2.0
manje zlo	RBR JJ	-2.3
drugi problemi	JJ NNS	-2.8
niska sredstva	JJ NNS	-6.8
neetične prakse	JJ NNS	-8.5
Prosjek		-1.2

Rezultati Turneyovog algoritma

- 410 kritika iz www.epinions.com
 - 170 (41% negativnih)
 - 210 (59% pozitivnih)
- Većinska klasa: 59%
- Turney algoritam: 74%

- Koristiti fraze rađe nego same riječi
- Učiti nad područnim informacijama

Korištenje WordNet-a za određivanje polarnosti

- WordNet: online leksikon sinonima (thesaurus)
- Ideja
 - Stvoriti pozitivni skup ("dobar") i negativni skup ("loš") riječi
 - Pronađi sinonime i antonime
 - Pozitivni skup: Dodaj sinonime pozitivne riječi ("sjajan") i antonime negativne riječi ("užasan")
 - Negativni skup: Dodaj sinonime negativne riječi ("grozан") i antonime pozitivne riječi ("odličan")
 - Ponovi, koristeći lanac sinonima
 - Filtriraj

Zaključak

- Prednosti:
 - namijenjen za specifično područje
 - može biti robusniji (više riječi)
- Ideja
 - Započeti s inicijalnim skupom riječi ("dobar", "loš")
 - Pronaći ostale riječi koje imaju sličan polaritet:
 - koristeći "i" i "ili"
 - koristeći riječi koje se ko-javljuju u istom dokumentu
 - koristeći sinonime i antonime

Uvod u obradu prirodnog jezika

8.1. Generativni protiv diskriminativnih modela

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Uvod

- Do sada smo razmatrali **generativne modele**
 - modeli jezika, Naivni Bayes
- Danas se često koriste **uvjetni** odnosno **diskriminativni probabilistički modeli** kod obrade prirodnog jezika, prepoznavanja govora, vraćanju informacija (strojnom učenju općenito) jer:
 - su veoma precizni
 - omogućavaju uključivanje mnogih lingvistički značajnih osobina
 - omogućavaju automatsku izgradnju jezično nezavisnih modula za obradu prirodnog jezika

Združeni protiv uvjetnog modela

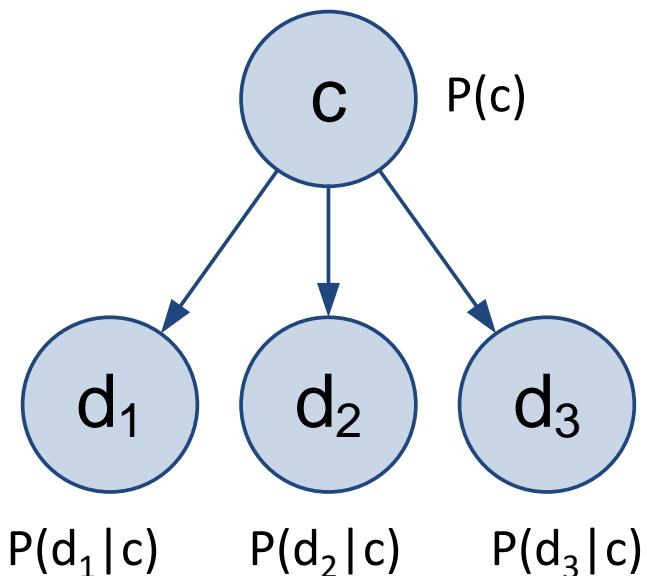
- Neka imamo skup $\{(d, c)\}$ uparenih promatranih podataka d i skrivenih klasa c
- **Združeni (generativni) model** postavljaju vjerojatnosti nad promatranim podacima d i skrivenim stvarima $P(c,d)$ (generiraju promatrane podatke temeljem skrivenih stvari)
 - Svi klasični statistički modeli obrade prirodnog jezika:
 - n-gram modeli, Naivni Bayesovi klasifikatori, skriveni Markovljevi modeli, probabilističke kontekstno neovisne gramatike, ...

Združeni protiv uvjetnog modela

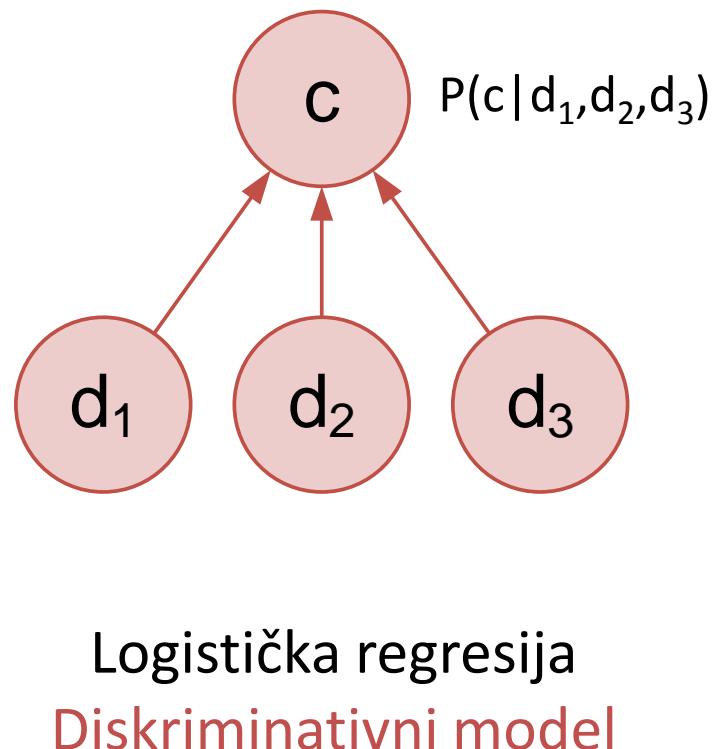
- **Uvjetni (diskriminativni) model** ostavlja podatke kako jesu i postavljaju vjerojatnost nad skrivenim stvarima za dane podatke $P(c|d)$
 - logistička regresija, modeli maksimalne entropije, uvjetna slučajna polja
 - stroj s potpornim vektorima, perceptroni ... su diskriminativni modeli koji nisu probabilistički

Bayesova mreža / grafički model

- Bayesova mreža za vrhove ima slučajne variable i lukove za direktne zavisnosti
- Neke varijable se promatraju, dok su neke skrivene
- Svaki čvor je mali klasifikator (tablica uvjetne vjerojatnosti) temeljem dolaznih lukova



Naivni Bayes
Generativni model



Logistička regresija
Diskriminativni model

Uvjetna protiv združene vjerodostojnosti

- Združeni model daje vjerojatnosti $P(d,c)$ i pokušava maksimizirati združenu vjerodostojnost
 - trivijalnim izborom težina: samo relativne frekvencije.
- Uvjetni model daje vjerojatnost $P(c|d)$. Uzima podatke kakve jesu i modelira samo uvjetnu vjerojatnost klase
 - teži se maksimizaciji uvjetne vjerodostojnosti
 - teže za napraviti

Uvjetni modeli rade dobro

- Primjer određivanja smisla riječi
- Čak i s istim osobimama, promjenom metode na uvjetnu se povećavaju performanse

Skup za treniranje	
Metoda	Točnost
Združena	86.8
Uvjetna	98.5

Skup za testiranje	
Metoda	Točnost
Združena	73.6
Uvjetna	76.1

(Klein and Manning 2002, using Senseval-1 Data)

Uvod u obradu prirodnog jezika

8.2. Osobine diskriminativnog modela

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Osobine (features)

- Osobina f je elementarni dio dokaza koji povezuje aspekte onoga što promatramo d (podatak) i onoga što predviđamo c (klasa)
- Osobina je realna funkcija

$$f : C \times D \rightarrow R$$

Primjer osobina

$f_1(c,d) \equiv [c = \text{LOKACIJA} \wedge w_{-1} = "u" \wedge \text{imaPrvoSlovoVeliko}(w)]$

$f_2(c,d) \equiv [c = \text{LOKACIJA} \wedge \text{imaHrvatskeZnakove}(w)]$

$f_3(c,d) \equiv [c = \text{LIJEK} \wedge \text{završava}(w, "c")]$

LOKACIJA

u Split

LOKACIJA

u Čakovec

LIJEK

uzima Prozac

OSOBA

vidi Ivana

- Model će svakoj osobini pridružiti **težinu**:
 - Pozitivna težina govori da je vjerojatno točno
 - Negativna težina govori da vjerojatno nije točno

Očekivanja osobina

- Iskorištavaju se dva **očekivanja**:
 - točan ili predviđeni broj korištenja osobina

- Empirijski broj (očekivanje) osobine

$$\text{empijski } E(f_i) = \sum_{(c,d) \in \text{promatran} \& C,D} f_i(c,d)$$

- Model očekivanja osobine

$$E(f_i) = \sum_{(c,d) \in (C,D)} P(c,d) f_i(c,d)$$

Osobine

- Kod obrade prirodnog jezika, osobina najčešće specificira
 1. binarnu funkciju svojstava ulaznih podataka
 2. određenu klasu

$$f_i(c, d) \equiv [\Phi(d) \wedge c = c_j] \quad [vrijednost\ je\ 0\ ili\ 1]$$

- Svaka osobina bira podskup podataka i sugerira kako će ih označiti

Modeli zasnovani na osobinama

- Odlučivanje o mjestu u podacima se temelji na aktivnim **osobinama** na tom mjestu

Podaci:

POSAO: Dionice su dosegle vrh ...

Oznaka: POSAO

Osobine: {..., dionice, su, dosegle, vrh, ...}

Kategorizacija teksta

Podaci:

... za rekonstrukciju bankovnog: NOVAC duga

Oznaka: NOVAC

Osobine: {..., w_{-1} = rekonstrukciju, w_1 = duga, $L=9$, ...}

Smisao riječi

Podaci:

N V A ...
ulica je duga ...

Oznaka: A

Osobine: { $w =$ duga, $t_{-1} = V$, $w_{-1} =$ je, ...}

POS označavanje

Primjer: Kategorizacija teksta

- (Zhang and Oles 2001)
- Osobine su prisutnosti svake riječi u dokumentu i klasi dokumenta
- Testovi na klasičnim Reuters skupovima podataka
 - Naivni Bayes: 77.0% F1
 - Linearna regresija: 86.0%
 - Logistička regresija: 86.4%
 - SVM: 86.5%
- Rad ističe važnost regularizacije (izglađivanja) za uspješnu upotrebu diskriminativnih metoda.

Drugi primjeri klasifikacije s maksimalnom entropijom

- MAXENT klasifikator koristiti kad god se želi pridružiti mjesto podataka jednoj od brojnih klasa:
 - Detekcija granice rečenice (Mikheev 2000)
 - Je li točka kraj rečenice ili kratica?
 - Sentimentalna analiza (Pang and Lee 2002)
 - unigrami, bigrami riječi, brojenje POS, ...
 - prijedlozi u POS (Ratnaparkhi 1998)
 - Pridružiti glagolu ili imenici? Osobine glavne imenice, prijedloga itd.
 - Opće odlučivanje prilikom parsiranja (Ratnaparkhi 1997; Johnson et al. 1999, etc.)

Uvod u obradu prirodnog jezika

8.3. Linearni klasifikator temeljen na osobinama (Feature-based linear classifier)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Linearni klasifikatori temeljeni na osobinama

- Linearni klasifikatori tijekom klasifikacije:
 - Linearna funkcija iz skupa osobina $\{f_i\}$ u skup klase $\{c\}$
 - Pridruživanje težine λ_i svakoj osobini f_i
 - Promatra se svaka klasa c za dane podatke d
 - Za par (c, d) , osobine biraju njihovim težinama
 - $\text{biraj}(c) = \sum \lambda_i f_i(c, d)$

$$\begin{array}{ccc} \lambda_1=1.8 & \lambda_2=-0.6 & \lambda_3=0.3 \\ \text{OSOBA} & \text{LOKACIJA} & \text{LIJEK} \\ u \check{\text{C}}akovec & u \check{\text{C}}akovec & u \check{\text{C}}akovec \end{array}$$

- Izabrana klasa c koja maksimizira vjerojatnost
 $\sum \lambda_i f_i(c, d) = \text{LOKACIJA}$

Linearni klasifikatori temeljeni na osobinama

- Postoje razni načini odabira težina za osobine:
 - Perceptron: pronađi trenutni pogrešno klasificiran primjer i guraj težine u smjeru njegove korektne klasifikacije
 - Granične metode (SVM)

Linearni klasifikatori temeljeni na osobinama

- Eksponencijalni (log-linearni, maxent, logički, Gibbs) modeli:
 - Napravi probabilistički model iz linearne kombinacije $\sum \lambda_i f_i(c, d)$

$$P(c | d, \lambda) = \frac{\exp \sum \lambda_i f_i(c, d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c', d)}$$

Čini biranje pozitivnim

Normalizira biranje

- $P(\text{LOKACIJA} | \text{u Čakovec}) = e^{1.8}e^{-0.6} / (e^{1.8}e^{-0.6} + e^{0.3} + e^0) = 0.586$
- $P(\text{LIJEK} | \text{u Čakovec}) = e^{0.3} / (e^{1.8}e^{-0.6} + e^{0.3} + e^0) = 0.238$
- $P(\text{OSOBA} | \text{u Čakovec}) = e^0 / (e^{1.8}e^{-0.6} + e^{0.3} + e^0) = 0.176$

- Težine su parametri probabilističkog modela

Linearni klasifikatori temeljeni na osobinama

- Eksponencijalni (log-linearni, maxent, logistički, Gibbs) modeli:
 - Za ovaj oblik modela, izabrat ćemo parametre $\{\lambda_i\}$ koji maksimiziraju uvjetnu vjerojatnost podataka po ovom modelu.
 - Ne gradi se samo klasifikacija, već i distribucija vjerojatnosti nad klasifikacijom
 - postoje drugi (dobri) načini diskriminiranja klase – SVM, perceptroni – ali nisu toliko trivijalne za interpretaciju distribucija nad klasama.

Veza sa logističkom regresijom

- Maxent modeli u OPJ su esencijalno isti kao i modeli višeklasne logističke regresije u statistici (ili strojnom učenju)

Uvod u obradu prirodnog jezika

8.4. Izgradnja MaxEnt modela

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Izgradnja MaxEnt modela

- Definiranje osobina nad mjestima u podacima
 - osobine predstavljaju skupove mjesta u podacima koji su dovoljno karakteristični da zasluže parametre modela:
 - riječi, riječi s brojem, riječi koje završavaju na "iti", itd.
- Kodiranje Φ osobine kao jedinstveni string
 - Podatak će dovesti do niza stringova; aktivnih Φ osobina
 - Svaka osobina $f_i(c, d) \equiv [\Phi(d) \wedge c = c_j]$ dobiva težinu kao realni broj

Izgradnja MaxEnt modela

- Osobine se često dodavaju tijekom razvoja modela kako bi označili pogreške
 - Često, najjednostavnije je dodavati osobine koje označavaju loše kombinacije
- Tada, za svaku težinu osobine, želimo biti u stanju izračunati:
 - uvjetnu vjerojatnost podataka
 - derivaciju vjerojatnosti sa težinom modela
 - koristi očekivanja za svaku osobinu po modelu
- Sada se mogu odrediti optimalne težine osobina

Uvod u obradu prirodnog jezika

8.5. Problem nadbrojavanja dokaza

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Klasifikacija teksta: Azija ili Evropa?

Evropa

Monaco
Monaco

Monaco

Monaco
Monaco

Podaci za treniranje

Monaco
Hong
Kong



Azija

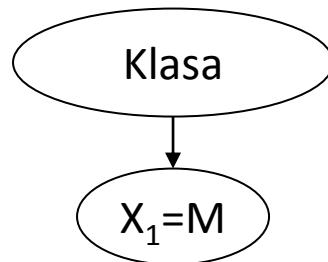
Hong
Kong
Monaco

Monaco

Hong
Kong

Hong
Kong

Naivni Bayes (NB)
model



NB faktori

- $P(A) = P(E) = 4/8 = 1/2$
- $P(M|A) = 2/8 = 1/4$
- $P(M|E) = 6/8 = 3/4$

Predikcije

- $P(A,M) = 1/2 * 1/4 = 1/8$
- $P(E,M) = 1/2 * 3/4 = 3/8$
- $P(A|M) = (1/8) / (8/16) = 1/4$
- $P(E|M) = (3/8) / (8/16) = 3/4$

Klasifikacija teksta: Azija ili Evropa?

Evropa

Monaco
Monaco

Monaco

Monaco
Monaco

Podaci za treniranje

Monaco
Hong
Kong



Azija

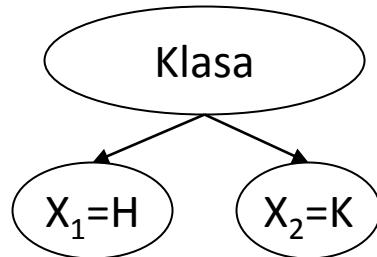
Hong
Kong
Monaco

Monaco

Hong
Kong

Hong
Kong

Naivni Bayes (NB)
model



NB faktori

- $P(A) = 4/8 = 1/2$
- $P(E) = 4/8 = 1/2$
- $P(H|A) = P(K|A) = 3/8$
- $P(H|E) = P(K|E) = 1/8$

Predikcije

- $P(A,H,K) = 1/2 * 3/8 * 3/8 \propto 9$
- $P(E,H,K) = 1/2 * 1/8 = 1/8 \propto 1$
- $P(A|H,K) = 9/(9+1) = 9/10$
- $P(E|H,K) = 1/(9+1) = 1/10$

Klasifikacija teksta: Azija ili Evropa?

Evropa

Monaco
Monaco

Monaco

Monaco
Monaco

Podaci za treniranje

Monaco
Hong
Kong



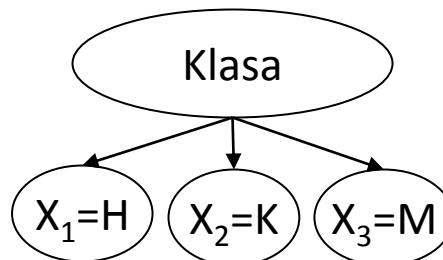
Hong
Kong
Monaco

Monaco

Hong
Kong

Azija

Naivni Bayes (NB)
model



NB faktori

- $P(A) = 4/8 = 1/2$
- $P(E) = 4/8 = 1/2$
- $P(M|A) = 2/8 = 1/4$
- $P(M|E) = 6/8 = 3/4$
- $P(H|A) = P(K|A) = 3/8$
- $P(H|E) = P(K|E) = 1/8$

Predikcije

- $P(A,H,K,M) = 1/2 * 3/8 * 3/8 * 1/4 \approx 9$
- $P(E,H,K,M) = 1/2 * 1/8 * 1/8 * 3/4 \approx 3$
- $P(A|H,K,M) = 9/(9+3) = 9/12 = 3/4$
- $P(E|H,K,M) = 3/(9+3) = 3/12 = 1/4$

Naivni Bayes protiv MaxEnt modela

- Naivni Bayesovi modeli višestruko broje dokaze koji su povezani
 - Svaka osobina je pomnožena, iako postoji više osobina koje govore o istoj stvari
- MaxEnt modeli (u većini) rješavaju ovaj problem
 - To se postiže određivanjem težina osobinama rezultirajući da očekivanja modela odgovaraju promatranim očekivanjima

Uvod u obradu prirodnog jezika

8.6. Maksimizacija vjerodostojnosti

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Vjerodostojnost eksponencijalnog modela

- Maksimalni (uvjetni) modeli vjerodostojnosti
 - za dani model, odaberi parametre tako da se maksimizira (uvjetne) vjerodostojnosti podataka

$$\log P(C \mid D, \lambda) = \sum_{(c,d) \in (C,D)} \log P(c \mid d, \lambda) = \sum_{(c,d) \in (C,D)} \log \frac{\exp \sum_i \lambda_i f_i(c, d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c', d)}$$

Vrijednost vjerodostojnosti

- (Logaritamska) uvjetna vjerodostojnost promatranih podataka (C, D) po maxent modelu je funkcija podataka i parametara λ :

$$\log P(C | D, \lambda) = \log \prod_{(c,d) \in (C,D)} P(c | d, \lambda) = \sum_{(c,d) \in (C,D)} \log P(c | d, \lambda)$$

- Ako nema mnogo vrijednosti od c , jednostavno se izračuna:

$$\log P(C | D, \lambda) = \sum_{(c,d) \in (C,D)} \log \frac{\exp \sum_i \lambda_i f_i(c, d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c', d)}$$

Vrijednost vjerodostojnosti

- Možemo izdvojiti dvije komponente:

$$\log P(C | D, \lambda) = \sum_{(c,d) \in (C,D)} \log \exp \sum_i \lambda_i f_i(c,d) - \sum_{(c,d) \in (C,D)} \log \sum_{c'} \exp \sum_i \lambda_i f_i(c',d)$$

$$\log P(C | D, \lambda) = N(\lambda) - M(\lambda)$$

- Derivacija je razlika derivacija svake komponente

Derivacija brojnika

$$\begin{aligned}\frac{\partial N(\lambda)}{\partial \lambda_i} &= \frac{\partial \sum_{(c,d) \in (C,D)} \log \exp \sum_i \lambda_i f_i(c,d)}{\partial \lambda_i} = \frac{\partial \sum_{(c,d) \in (C,D)} \sum_i \lambda_i f_i(c,d)}{\partial \lambda_i} \\ &= \sum_{(c,d) \in (C,D)} \frac{\partial \sum_i \lambda_i f_i(c,d)}{\partial \lambda_i} = \sum_{(c,d) \in (C,D)} f_i(c,d)\end{aligned}$$

- Derivacija brojnika je stvarni broj(f_i, c)

Derivacija nazivnika

$$\begin{aligned}\frac{\partial M(\lambda)}{\partial \lambda_i} &= \frac{\partial \sum_{(c,d) \in (C,D)} \log \sum_{c'} \exp \sum_i \lambda_i f_i(c',d)}{\partial \lambda_i} \\&= \sum_{(c,d) \in (C,D)} \frac{1}{\sum_{c''} \exp \sum_i \lambda_i f_i(c'',d)} \frac{\partial \sum_{c'} \exp \sum_i \lambda_i f_i(c',d)}{\partial \lambda_i} \\&= \sum_{(c,d) \in (C,D)} \frac{1}{\sum_{c''} \exp \sum_i \lambda_i f_i(c'',d)} \sum_{c'} \frac{\exp \sum_i \lambda_i f_i(c',d)}{1} \frac{\partial \sum_i \lambda_i f_i(c',d)}{\partial \lambda_i} \\&= \sum_{(c,d) \in (C,D)} \sum_{c'} \frac{\exp \sum_i \lambda_i f_i(c',d)}{\sum_{c''} \exp \sum_i \lambda_i f_i(c'',d)} \frac{\partial \sum_i \lambda_i f_i(c',d)}{\partial \lambda_i} \\&= \sum_{(c,d) \in (C,D)} \sum_{c'} P(c' | d, \lambda) f_i(c',d)\end{aligned}$$

- Derivacija nazivnika je predviđen broj(f_i, c)

Derivacija

$$\frac{\partial \log P(C | D, \lambda)}{\partial \lambda_i} = \text{stvaran broj}(f_i, c) - \text{predviđen broj}(f_i, c)$$

- Optimalni parametri su oni kod kojih je predviđeno očekivanje jednako epmirijskom očekivanju.
- Optimalna distribucija:
 - uvijek je jedinstvena (ali parametri ne moraju biti jedinstveni)
 - uvijek postoji (ako su odabране osobine nad aktualnim podacima)
- Ovi modeli se zovu modeli maksimalne entropije

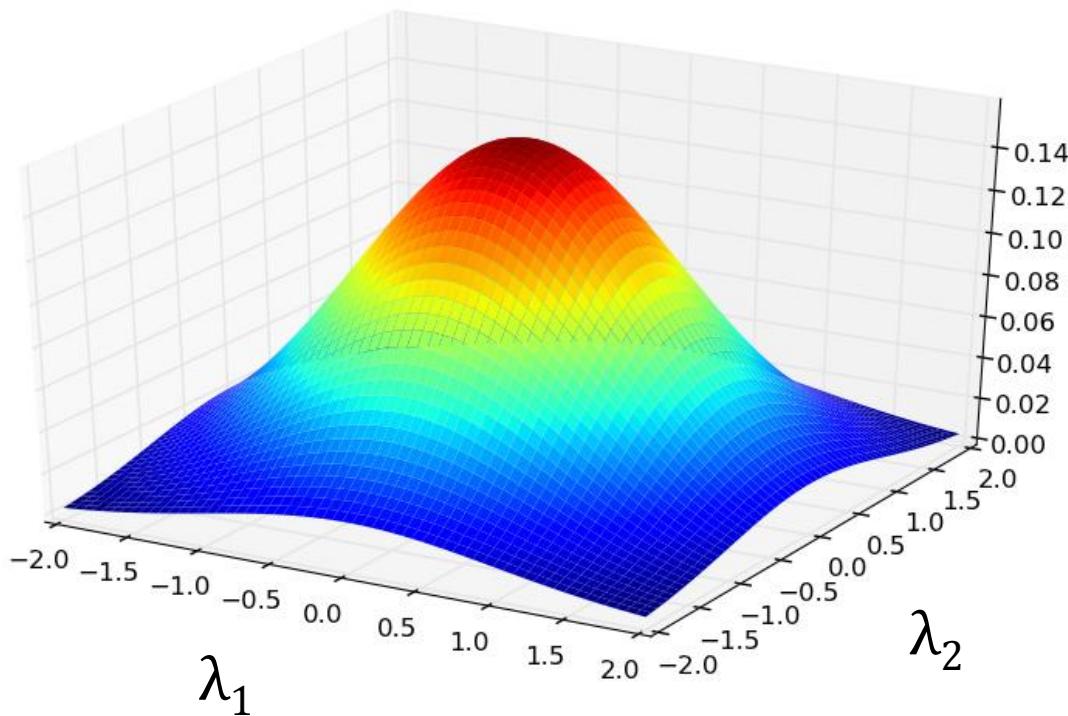
Pronalaženje optimalnih parametara

- Želimo odabratи parametre $\lambda_1, \lambda_2, \lambda_3 \dots$ koji maksimiziraju uvjetni logaritam vjerodostojnosti (conditional log likelihood) podataka za treniranje

$$CLogLik(D) = \sum_{i=1}^n \log P(c_i | d_i)$$

- Kako bi to napravili, pokazali smo kako se računaju parcijalne derivacije (gradijent)

Površina vjerodostojnosti



Pronalaženje optimalnih parametara

1. GD - Gradijentno spuštanje (Gradient Descent)
SGD - Stohastičko gradijentno spuštanje
2. GIS – Generalized Iterative Scaling
IIS – Improved iterative Scaling
3. CG – Conjugate Gradient
4. Kvazi-Newtonove metode

Uvod u obradu prirodnog jezika

8.1. Logistička regresija

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Cilj logističke regresije

- učenje klasifikatora koji može donijeti binarnu odluku za klasu nekog novog ulaznog promatranja

Komponente logističke regresije

- Skup za treniranje od m promatranja
 - promatranje je par ulaza i izlaza $(x^{(i)}, y^{(i)})$
- Strojno učenje za klasifikaciju ima sljedeće komponente:
 1. **Reprezentacija osobina** za ulaze
 - svaki ulaz $x^{(i)}$ je predstavljen vektorom osobina $[x_1, x_2, \dots, x_n]$
 - osobina i za ulaz $x^{(j)}$ je $x_i^{(j)}$ (ili f_i ili $f_i(x)$)
 2. **Funkcija klasifikacije** koja računa \hat{y} - procjenu klase pomoću $p(y|x)$
 - sigmoid, softmax, ...
 3. **Aktivacijska funkcija** za učenje koja obično uključuje minimizaciju greške
 - Unakrsna entropija gubitka
 4. **Algoritam za optimizaciju** aktivacijske funkcije:
 - stohastičko opadanje gradijenta

Faze logističke regresije

1. Učenje (treniranje)

- pomoću stohastičkog opadanja gradijenta i gubitka unakrsne entropije

2. Testiranje

- Za dani testni primjer x računa se $p(y|x)$ i vraća klasa s većom vjerojatnošću $y = 1$ ili $y = 0$

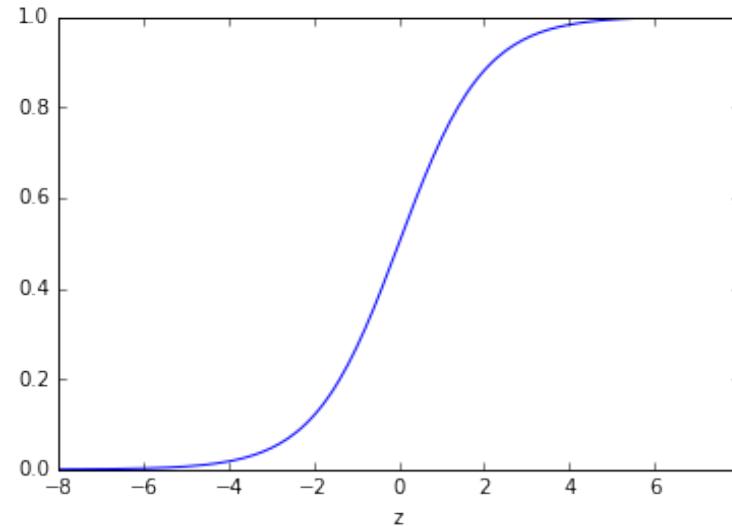
Klasifikacija

Uči se **vektor težina** $w = [w_1 \ w_2 \ \dots \ w_n]$ i **pristranost** b

- Težina w_i govori koliko je osobina x_i bitna za odluku
- $z = \sum_{i=1}^n w_i x_i + b$ - težinska suma
- $z = w \cdot x + b$ gdje je \cdot skalarni produkt

Težinska suma z je realni broj iz $\langle -\infty, +\infty \rangle$
kojeg treba prebaciti u vjerojatnostni prostor $[0, 1]$

- Sigmoid – logistička funkcija $y = \sigma(z) = \frac{1}{1+e^{-z}}$



Klasifikacija

Sigmoid klasifikator

- x promatranje (ulaz)
- $[x_1 \ x_2 \ \dots \ x_n]$ vektor osobina za x
- $y = 1$ ili $y = 0$ klasa (izlaz)

Želimo izračunati $p(y = 1|x)$

Primjer: Odluka "pozitivan sentiment" ili "negativan sentiment" za osobinu koja prebrojava riječi u dokumentu:

- $p(y = 1|x)$ je vjerojatnost da je dokument "pozitivan"
- $p(y = 0|x)$ je vjerojatnost da je dokument "negativan"

Klasifikacija

Izračun vjerojatnosti:

$$p(y = 1|x) = \vartheta(w \cdot x + b) = \frac{1}{1 + e^{-(w \cdot x + b)}}$$

$$p(y = 0|x) = 1 - p(y = 1|x) = 1 - \frac{1}{1 + e^{-(w \cdot x + b)}} = \frac{e^{-(w \cdot x + b)}}{1 + e^{-(w \cdot x + b)}}$$

Kako odlučiti?

Neka je 0.5 **granična vrijednost**

$$\hat{y} = \begin{cases} 1 & \text{ako } p(y = 1|x) > 0.5 \\ 0 & \text{u suprotnom} \end{cases}$$

Klasifikacija

Primjer: Binarna klasifikacija sentimenta kritike filma

Je li kritika pozitivna (+) ili negativna (-)

Osobine promatranja dokumenta d

Osobina	Definicija	Vrijednost
x_1	$\text{broj}(\text{pozitivni leksikon}) \in d$	3
x_2	$\text{broj}(\text{negativni leksikon}) \in d$	2
x_3	$\begin{cases} 1, & \text{ako "ne" } \in d \\ 0, & \text{inače} \end{cases}$	1
x_4	$\text{broj}(\text{zamjenice prvog i drugog lica } \in d)$	3
x_5	$\begin{cases} 1, & \text{ako "!" } \in d \\ 0, & \text{inače} \end{cases}$	0
x_6	$\log(\text{broj riječi od } d)$	$\ln(64) = 4.15$

Klasifikacija

Osobina	Definicija	Vrijednost
x_1	$\text{broj}(\text{pozitivni leksikon}) \in d$	3
x_2	$\text{broj}(\text{negativni leksikon}) \in d$	2
x_3	$\begin{cases} 1, & \text{ako "ne" } \in d \\ 0, & \text{inače} \end{cases}$	1
x_4	$\text{broj}(\text{zamjenice prvog i drugog lica } \in d)$	3
x_5	$\begin{cases} 1, & \text{ako "!" } \in d \\ 0, & \text{inače} \end{cases}$	0
x_6	$\log(\text{broj riječi od } d)$	$\ln(64) = 4.15$

$x_2=2$

$x_3=1$

$x_1=3$

Sve je **isfolirano**. Gotovo **nema** iznenadjenja, a scenarij je

drugorazredan. Pa zašto je onda bio **užitak** gledati?

Kao prvo, glumci su **sjajni**. Još jedna **dobra** stvar je glazba.

Prevladao **me** nagon da se maknem s kauča i počnem plesati. Uvuklo **me** potpunosti, a i **vas** će.

$x_4=3$

$x_5=0$

$x_6=4.15$

Klasifikacija

Osobina	Definicija	Vrijednost
x_1	$\text{broj}(\text{pozitivni leksikon}) \in d$	3
x_2	$\text{broj}(\text{negativni leksikon}) \in d$	2
x_3	$\begin{cases} 1, & \text{ako "ne" } \in d \\ 0, & \text{inače} \end{cases}$	1
x_4	$\text{broj}(\text{zamjenice prvog i drugog lica } \in d)$	3
x_5	$\begin{cases} 1, & \text{ako "!" } \in d \\ 0, & \text{inače} \end{cases}$	0
x_6	$\log(\text{broj riječi od } d)$	$\ln(64) = 4.15$

- Težinski vektor $w = [2.5, -5.0, -1.2, 0.5, 2.0, 0.7]$
- Pristranost $b = 0.1$
- $p(+|x) = \partial(w \cdot x + b) =$
 $= \partial([2.5, -5.0, -1.2, 0.5, 2.0, 0.7] \cdot [3, 2, 1, 3, 0, 4.15] + 0.1)$
 $= \partial(1.805)$
 $= 0.86$
- $p(-|x) = 1 - p(+|x)$
 $= 0.14$

Učenje

Kako se uče parametri modela w i b ?

- Želimo da \hat{y} bude što bliži stvarnom y
- Odnosno da udaljenost između \hat{y} i y bude što manja

Funkcija gubitka $L(\hat{y}, y) =$ koliko mnogo se \hat{y} razlikuje od y

- Primjer funkcije gubitka je srednja vrijednost kvadrata (mean square error)
- $L_{\text{MSE}}(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$
- teško za optimizirati jer nije konveksna

Procjena uvjetne maksimalne izglednosti:

- biramo w i b koji **maksimiziraju** log **vjerodjnost stvarnih vrijednosti** od y podataka za učenje
- dobivena funkcija gubitka je **unakrsna entropija gubitka (cross-entropy loss)**

Želimo naučiti težine koje maksimiziraju vjerodjnost točne klase za $p(y|x)$

- Imamo dvije klase (1 ili 0)
- Bernoullijeva distribucija $p(y|x) = \hat{y}^y(1 - \hat{y})^{1-y}$
 - za $y = 1, p(y = 1|x) = \hat{y}$
 - za $y = 0, p(y = 0|x) = 1 - \hat{y}$

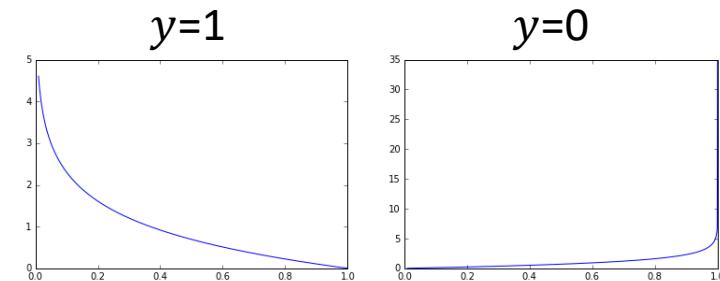
Učenje

Maksimizacija od $p(y|x)$ je isto što i maksimizacija od $\log(p(y|x))$

$$\begin{aligned}\log(p(y|x)) &= \log(\hat{y}^y(1 - \hat{y})^{1-y}) \\ &= y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\end{aligned}$$

Funkcija gubitka se minimizira, stoga

$$\begin{aligned}L_{CE}(\hat{y}, y) &= -\log(p(y|x)) \\ &= -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]\end{aligned}$$



Proširujemo na cijeli skup za učenje $\{(x^{(i)}, y^{(i)}) \mid i \in \{1, \dots, m\}\}$

$$\begin{aligned}\log\left(p(\{(x^{(i)}, y^{(i)}) \mid i \in \{1, \dots, m\}\})\right) &= \log\left(\prod_{i=1}^m p(y^{(i)}|x^{(i)})\right) = \\ &= \sum_{i=1}^m \log(p(y^{(i)}|x^{(i)})) \\ &= -\sum_{i=1}^m L_{CE}(\hat{y}^{(i)}, y^{(i)})\end{aligned}$$

Učenje

Funkcija gubitka na cijelom skupu za učenje

$$\begin{aligned} cost(w, b) &= \frac{1}{m} \sum_{i=1}^m L_{CE}(\hat{y}^{(i)}, y^{(i)}) \\ &= -\frac{1}{m} \sum_{i=1}^m y^{(i)} \log(\sigma(w \cdot x^{(i)} + b)) + (1 - y^{(i)}) \log(1 - \sigma(w \cdot x^{(i)} + b)) \end{aligned}$$

Za ovu funkciju je potrebno pronaći minimum

Opadanje gradijenta

Neka su θ parametri po kojima se minimizira

$$\theta = (w, b) \text{ kod logističke regresije}$$

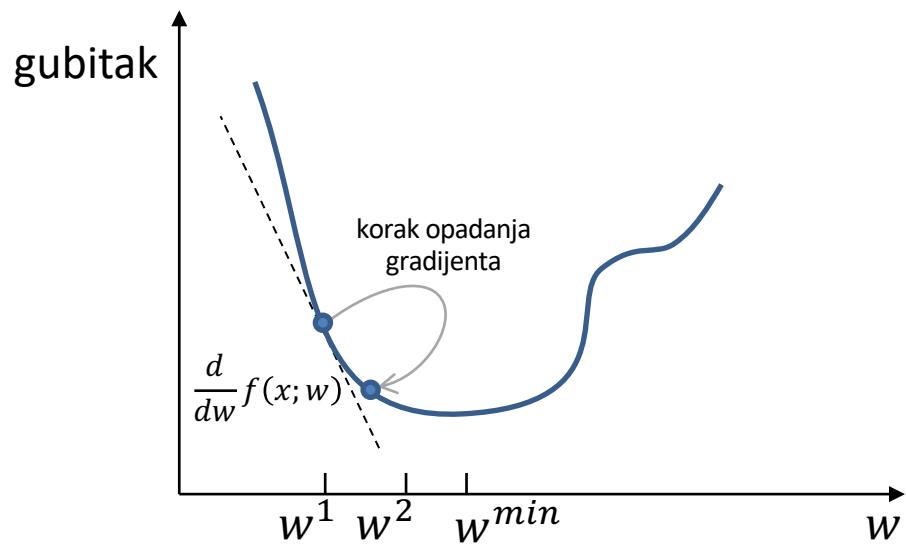
$$\hat{\theta} = \operatorname{argmin}_{\theta} \frac{1}{m} \sum_{i=1}^m L_{\text{CE}}(y^{(i)}, \hat{y}^{(i)}; \theta)$$

Po dogovoru, funkcija gubitka je konveksna funkcija (jedan minimum)

Metoda opadanja gradijenta garantira da će se minimum pronaći

Opadanje gradijenta

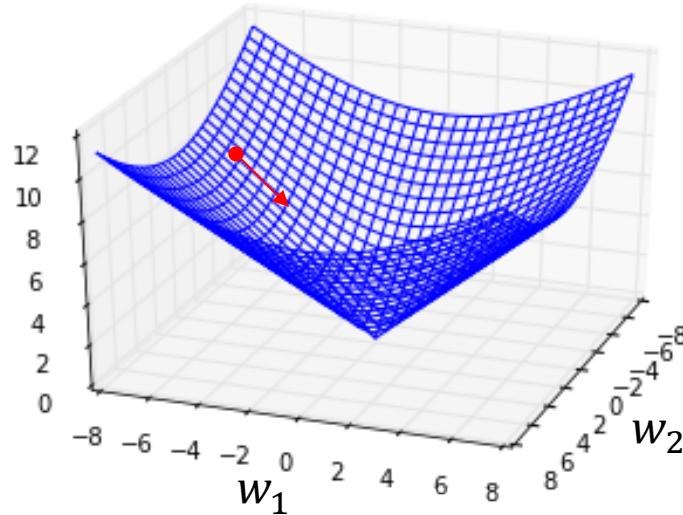
Prepostavimo da je funkcija gubitka $f(x; w)$ od jednog parametra w



$$w^{t+1} = w^t - \eta \frac{d}{dw} f(x; w) \text{ gdje je } \eta \text{ stopa učenja}$$

Opadanje gradijenta

Poopćimo funkciju gubitka $f(x; \theta)$ na više parametra w_i



$$\theta^{t+1} = \theta^t - \eta \nabla L(f(x; \theta), y) \text{ gdje je}$$

$$\nabla L(f(x; \theta), y) = \begin{bmatrix} \frac{\partial}{\partial w_1} L(f(x; \theta), y) \\ \frac{\partial}{\partial w_2} L(f(x; \theta), y) \\ \vdots \\ \frac{\partial}{\partial w_m} L(f(x; \theta), y) \end{bmatrix}$$

Opadanje gradijenta

Opadanje gradijenta kod logističke regresije

Kako izračunati $\nabla L(f(x; \theta), y)$

$$L_{CE}(w, b) = -[y \log(\sigma(w \cdot x + b)) + (1 - y) \log(1 - \sigma(w \cdot x + b))]$$

$$\begin{aligned}\frac{\partial L_{CE}(w, b)}{\partial w_j} &= \frac{\partial}{\partial w_j} - [y \log(\sigma(w \cdot x + b)) + (1 - y) \log(1 - \sigma(w \cdot x + b))] \\ &= - \left[\frac{\partial}{\partial w_j} y \log(\sigma(w \cdot x + b)) + \frac{\partial}{\partial w_j} (1 - y) \log(1 - \sigma(w \cdot x + b)) \right] \\ &= - \frac{y}{\partial(w \cdot x + b)} \frac{\partial}{\partial w_j} \sigma(w \cdot x + b) - \frac{1 - y}{1 - \sigma(w \cdot x + b)} \frac{\partial}{\partial w_j} (1 - \sigma(w \cdot x + b)) \\ &= - \left[\frac{y}{\partial(w \cdot x + b)} - \frac{1 - y}{1 - \sigma(w \cdot x + b)} \right] \frac{\partial}{\partial w_j} \sigma(w \cdot x + b) \\ &= - \left[\frac{y - \sigma(w \cdot x + b)}{\partial(w \cdot x + b)[1 - \sigma(w \cdot x + b)]} \right] \sigma(w \cdot x + b)[1 - \sigma(w \cdot x + b)] \frac{\partial \sigma(w \cdot x + b)}{\partial w_j} \\ &= -[y - \sigma(w \cdot x + b)]x_j \\ &= [\sigma(w \cdot x + b) - y]x_j\end{aligned}$$

Opadanje gradijenta kod grupnog treniranja

Grupno treniranje (batch training)

- određivanje gradijenta za cijeli skup podataka

Treniranje u mini grupama (mini-batch tranining)

- određivanje gradijenta za m podatak iz skupa podataka
($m = 512, 1024, \dots$)

$$cost(w, b) = \frac{1}{m} \sum_{i=1}^m L_{CE}(\hat{y}^{(i)}, y^{(i)})$$

$$cost(w, b) = -\frac{1}{m} \sum_{i=1}^m y^{(i)} \log \left(\sigma(w \cdot x^{(i)} + b) \right) + (1 - y^{(i)}) \log \left(1 - \sigma(w \cdot x^{(i)} + b) \right)$$

$$\frac{\partial cost(w, b)}{\partial w_j} = \frac{1}{m} \sum_{i=1}^m [\sigma(w \cdot x^{(i)} + b) - y^{(i)}] x_j^{(i)}$$

Stohastičko opadanje gradijenta

Algoritam stohastičkog opadanja gradijenta

$$\theta \leftarrow 0$$

ponovi T puta

za svaki $(x^{(i)}, y^{(i)})$ po slučajnom redoslijedu

izračunaj $\hat{y}^{(i)} = f(x^{(i)}; \theta)$

izračunaj gubitak $L(\hat{y}^{(i)}, y^{(i)})$

$\theta \leftarrow \theta - \eta \nabla L(f(x^{(i)}; \theta), y^{(i)})$

vrati θ

Stohastičko opadanje gradijenta

Primjer: neka je

- $x = [x_1, x_2] = [3, 2]$
- za θ^0 imamo $w = [w_1, w_2] = [0, 0], b = 0$
- $\eta = 0.1$

Znamo $\theta^{t+1} = \theta^t - \eta \nabla_{\theta} L(f(x^{(i)}; \theta), y^{(i)})$ stoga

$$\begin{aligned}\nabla_{w,b} &= \begin{bmatrix} \frac{\partial}{\partial w_1} L_{\text{CE}}(w, b) \\ \frac{\partial}{\partial w_2} L_{\text{CE}}(w, b) \\ \frac{\partial}{\partial b} L_{\text{CE}}(w, b) \end{bmatrix} = \begin{bmatrix} (\sigma(w \cdot x + b) - y)x_1 \\ (\sigma(w \cdot x + b) - y)x_2 \\ \sigma(w \cdot x + b) - y \end{bmatrix} \\ &= \begin{bmatrix} (\sigma(0) - 1)x_1 \\ (\sigma(0) - 1)x_2 \\ \sigma(0) - 1 \end{bmatrix} = \begin{bmatrix} -0.5x_1 \\ -0.5x_2 \\ -0.5 \end{bmatrix} = \begin{bmatrix} -0.5 \cdot 3 \\ -0.5 \cdot 2 \\ -0.5 \end{bmatrix} = \begin{bmatrix} -1.5 \\ -1.0 \\ -0.5 \end{bmatrix}\end{aligned}$$

Stohastičko opadanje gradijenta

Primjer:

- $x = [x_1, x_2] = [3, 2]$
- $w = [w_1, w_2] = [0, 0], b = 0$
- $\eta = 0.1$

$$\theta^1 = \theta^0 - \eta \nabla_{w,b}$$

$$\theta^1 = \begin{bmatrix} w_1 \\ w_2 \\ b \end{bmatrix} - \eta \nabla_{w,b} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} - 0.1 \begin{bmatrix} -1.5 \\ -1.0 \\ -0.5 \end{bmatrix} = \begin{bmatrix} -0.15 \\ -0.10 \\ -0.05 \end{bmatrix}$$

Regularizacija

- Ako je osobina savršeno prediktivna (pojavljuje se samo u jednoj klasi), dobit će veliku težinu
- Težine osobina će nastojati savršeno odgovarati detaljima u podacima za učanje (**prenaučenost modela- overfitting**)
- Dobro naučen model mora moći generalizirati na dosad neviđenim podacima za testiranje.
- Regularizacija $R(w)$ se dodaje funkciji gubitka

$$\widehat{w} = \operatorname{argmax}_w \sum_{i=1}^m \log P(y^{(i)}|x^{(i)}) - \alpha R(w)$$

- $R(w)$ služi za "kažnjavanje" velikih težina

Regularizacija

- Dvije često korištene regularizacije
 - L2 regularizacija $R(W) = \|W\|_2^2 = \sum_{j=1}^n w_j^2$ - Euklidska udaljenost
 - L1 regularizacija $R(W) = \|W\|_1 = \sum_{j=1}^n |w_j|$ - Manhattan udaljenost
- L2 regularizacija se lakše optimizira (jednostavnija derivacija)

Primjer: Podaci

$$Train = \left\{ \left(\begin{bmatrix} x_1^{(1)} \\ \vdots \\ x_n^{(1)} \end{bmatrix}, y^{(1)} \right), \dots, \left(\begin{bmatrix} x_1^{(m)} \\ \vdots \\ x_n^{(m)} \end{bmatrix}, y^{(m)} \right) \right\} = \left\{ \left(\begin{bmatrix} 23 \\ 9 \\ 1 \\ 7 \\ 1 \\ 5 \end{bmatrix}, 0 \right), \left(\begin{bmatrix} 12 \\ 5 \\ 1 \\ 10 \\ 1 \\ 5 \end{bmatrix}, 0 \right), \left(\begin{bmatrix} 14 \\ 12 \\ 1 \\ 1 \\ 0 \\ 5 \end{bmatrix}, 1 \right) \right\}$$

$$X = [x^{(1)} \quad \dots \quad x^{(m)}] = \begin{bmatrix} x_1^{(1)} & \dots & x_1^{(m)} \\ \vdots & & \vdots \\ x_n^{(1)} & \dots & x_n^{(m)} \end{bmatrix} = \begin{bmatrix} 23 & 12 & 14 \\ 9 & 5 & 12 \\ 1 & 1 & 1 \\ 7 & 10 & 1 \\ 1 & 1 & 0 \\ 5 & 5 & 5 \end{bmatrix}$$

$$Y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Primjer: Inicijalizacija

$$W = [w_1 \quad \dots \quad w_n] = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$$

$$b = 0$$

$$\theta = [W \quad b] = [w_1 \quad \dots \quad w_n \quad b] = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$$

$$\eta = 0.01$$

Primjer: Učenje - predikcija

$$\hat{y}^{(i)} = \sigma(W \cdot x^{(i)} + b) = \sigma\left(\begin{bmatrix} w_1 & \dots & w_n \end{bmatrix} \cdot \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix} + b\right) = \sigma\left(\sum_{k=1}^n w_k x_k^{(i)} + b\right)$$

$$\hat{Y} = \sigma(W \cdot X + b)$$

$$\begin{bmatrix} \hat{y}^{(1)} \\ \vdots \\ \hat{y}^{(m)} \end{bmatrix} = \sigma\left(\begin{bmatrix} w_1 & \dots & w_n \end{bmatrix} \cdot \begin{bmatrix} x_1^{(1)} & \dots & x_1^{(m)} \\ \vdots & & \vdots \\ x_n^{(1)} & \dots & x_n^{(m)} \end{bmatrix} + \begin{bmatrix} b \\ \vdots \\ b \end{bmatrix}\right) = \begin{bmatrix} \sigma\left(\sum_{k=1}^n w_k x_k^{(1)} + b\right) \\ \vdots \\ \sigma\left(\sum_{k=1}^n w_k x_k^{(m)} + b\right) \end{bmatrix}$$

$$\begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \hat{y}^{(3)} \end{bmatrix} = \sigma\left(\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 23 & 12 & 14 \\ 9 & 5 & 12 \\ 1 & 1 & 1 \\ 7 & 10 & 1 \\ 1 & 1 & 0 \\ 5 & 5 & 5 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}\right) =$$

$$\begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \hat{y}^{(3)} \end{bmatrix} = \begin{bmatrix} \sigma(0 \cdot 23 + 0 \cdot 9 + 0 \cdot 1 + 0 \cdot 7 + 0 \cdot 1 + 0 \cdot 5 + 0) \\ \sigma(0 \cdot 12 + 0 \cdot 5 + 0 \cdot 1 + 0 \cdot 10 + 0 \cdot 1 + 0 \cdot 5 + 0) \\ \sigma(0 \cdot 14 + 0 \cdot 12 + 0 \cdot 1 + 0 \cdot 1 + 0 \cdot 0 + 0 \cdot 5 + 0) \end{bmatrix} = \begin{bmatrix} \sigma(0) \\ \sigma(0) \\ \sigma(0) \end{bmatrix} = \begin{bmatrix} \frac{1}{1+e^{-0}} \\ \frac{1}{1+e^{-0}} \\ \frac{1}{1+e^{-0}} \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}$$

Primjer: Učenje - predikcija

$$\hat{y}^{(i)} = \sigma\left(\theta \cdot \begin{bmatrix} x^{(i)} \\ 1 \end{bmatrix}\right) = \sigma\left([W \quad b] \cdot \begin{bmatrix} x^{(i)} \\ 1 \end{bmatrix}\right) = \sigma\left([w_1 \quad \dots \quad w_n \quad b] \cdot \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_n^{(i)} \\ 1 \end{bmatrix}\right) = \sigma\left(\sum_{k=1}^n w_k x_k^{(i)} + b \cdot 1\right)$$

$$\hat{Y} = \sigma\left(\theta \cdot \begin{bmatrix} X \\ 1 \end{bmatrix}\right)$$

$$\begin{bmatrix} \hat{y}^{(1)} \\ \vdots \\ \hat{y}^{(m)} \end{bmatrix} = \sigma\left([w_1 \quad \dots \quad w_n \quad b] \cdot \begin{bmatrix} x_1^{(1)} & \dots & x_1^{(m)} \\ \vdots & & \vdots \\ x_n^{(1)} & \dots & x_n^{(m)} \\ 1 & \dots & 1 \end{bmatrix}\right) = \begin{bmatrix} \sigma\left(\sum_{k=1}^n w_k x_k^{(1)} + b \cdot 1\right) \\ \vdots \\ \sigma\left(\sum_{k=1}^n w_k x_k^{(m)} + b \cdot 1\right) \end{bmatrix}$$

$$\begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \hat{y}^{(3)} \end{bmatrix} = \sigma\left([0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0] \cdot \begin{bmatrix} 23 & 12 & 14 \\ 9 & 5 & 12 \\ 1 & 1 & 1 \\ 7 & 10 & 1 \\ 1 & 1 & 0 \\ 5 & 5 & 5 \\ 1 & 1 & 1 \end{bmatrix}\right) =$$

$$\begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \hat{y}^{(3)} \end{bmatrix} = \sigma\left(\begin{bmatrix} 0 \cdot 23 + 0 \cdot 9 + 0 \cdot 1 + 0 \cdot 7 + 0 \cdot 1 + 0 \cdot 5 + 0 \cdot 1 \\ 0 \cdot 12 + 0 \cdot 5 + 0 \cdot 1 + 0 \cdot 10 + 0 \cdot 1 + 0 \cdot 5 + 0 \cdot 1 \\ 0 \cdot 14 + 0 \cdot 12 + 0 \cdot 1 + 0 \cdot 1 + 0 \cdot 0 + 0 \cdot 5 + 0 \cdot 1 \end{bmatrix}\right) = \begin{bmatrix} \sigma(0) \\ \sigma(0) \\ \sigma(0) \end{bmatrix} = \begin{bmatrix} \frac{1}{1+e^{-0}} \\ \frac{1}{1+e^{-0}} \\ \frac{1}{1+e^{-0}} \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}$$

Primjer: Učenje - trošak

$$L_{\text{CE}}(\hat{y}^{(i)}, y^{(i)}) = -(y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}))$$

$$L_{\text{CE}}(\hat{Y}, Y) = -\frac{1}{m} \sum_{i=1}^m (y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}))$$

$$\begin{aligned} L_{\text{CE}}(\hat{Y}, Y) &= L_{\text{CE}} \left(\begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right) = \\ &= -\frac{1}{3} \left[(0 \log(0.5) + (1 - 0) \log(1 - 0.5)) + \right. \\ &\quad \left. (0 \log(0.5) + (1 - 0) \log(1 - 0.5)) + \right. \\ &\quad \left. (1 \log(0.5) + (1 - 1) \log(1 - 0.5)) \right] \\ &= -\frac{1}{3} [-0.69 - 0.69 - 0.69] = 0.69 \end{aligned}$$

Primjer: Učenje – opadanje gradijenta

$$\theta^{t+1} = \theta^t - \eta \nabla_{\theta}$$

$$\nabla_{\theta} = \begin{bmatrix} \frac{\partial}{\partial w_1} L_{\text{CE}}(W, b) \\ \vdots \\ \frac{\partial}{\partial w_n} L_{\text{CE}}(W, b) \\ \frac{\partial}{\partial b} L_{\text{CE}}(W, b) \end{bmatrix} = \frac{1}{m} \begin{bmatrix} X \\ 1 \end{bmatrix} (\hat{Y} - Y) = \frac{1}{m} \begin{bmatrix} x_1^{(1)} & \dots & x_1^{(m)} \\ \vdots & & \vdots \\ x_n^{(1)} & \dots & x_n^{(m)} \\ 1 & \dots & 1 \end{bmatrix} \left(\begin{bmatrix} \hat{y}^{(1)} \\ \vdots \\ \hat{y}^{(m)} \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix} \right) = \frac{1}{m} \begin{bmatrix} \sum_{j=1}^m (\hat{y}^{(j)} - y^{(j)}) x_1^{(j)} \\ \vdots \\ \sum_{j=1}^m (\hat{y}^{(j)} - y^{(j)}) x_n^{(j)} \\ \sum_{j=1}^m (\hat{y}^{(j)} - y^{(j)}) \end{bmatrix}$$

$$\nabla_{\theta} = \frac{1}{3} \begin{bmatrix} 23 & 12 & 14 \\ 9 & 5 & 12 \\ 1 & 1 & 1 \\ 7 & 10 & 1 \\ 1 & 1 & 0 \\ 5 & 5 & 5 \\ 1 & 1 & 1 \end{bmatrix} \left(\begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right) = \frac{1}{3} \begin{bmatrix} (0.5 - 0) \cdot 23 + (0.5 - 0) \cdot 12 + (0.5 - 1) \cdot 14 \\ (0.5 - 0) \cdot 9 + (0.5 - 0) \cdot 5 + (0.5 - 1) \cdot 12 \\ (0.5 - 0) \cdot 1 + (0.5 - 0) \cdot 1 + (0.5 - 1) \cdot 1 \\ (0.5 - 0) \cdot 7 + (0.5 - 0) \cdot 10 + (0.5 - 1) \cdot 1 \\ (0.5 - 0) \cdot 1 + (0.5 - 0) \cdot 1 + (0.5 - 1) \cdot 0 \\ (0.5 - 0) \cdot 5 + (0.5 - 0) \cdot 5 + (0.5 - 1) \cdot 5 \\ (0.5 - 0) \cdot 1 + (0.5 - 0) \cdot 1 + (0.5 - 1) \cdot 1 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 10.5 \\ 1 \\ 0.5 \\ 8 \\ 1 \\ 2.5 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 3.5 \\ 0.33 \\ 0.17 \\ 2.67 \\ 0.33 \\ 0.83 \\ 0.17 \end{bmatrix}$$

Primjer: Učenje – opadanje gradijenta

$$\theta^{t+1} = \theta^t - \eta \nabla_{\theta^t}$$

$$\begin{bmatrix} w_1 \\ \vdots \\ w_n \\ b \end{bmatrix}^{t+1} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \\ b \end{bmatrix}^t - \eta \nabla_{\theta^t}$$

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ b \end{bmatrix}^2 = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ b \end{bmatrix}^1 - \eta \begin{bmatrix} \frac{\partial}{\partial w_1} L_{\text{CE}}(W, b) \\ \frac{\partial}{\partial w_2} L_{\text{CE}}(W, b) \\ \frac{\partial}{\partial w_3} L_{\text{CE}}(W, b) \\ \frac{\partial}{\partial w_4} L_{\text{CE}}(W, b) \\ \frac{\partial}{\partial w_5} L_{\text{CE}}(W, b) \\ \frac{\partial}{\partial w_6} L_{\text{CE}}(W, b) \\ \frac{\partial}{\partial b} L_{\text{CE}}(W, b) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - 0.01 \begin{bmatrix} 3.5 \\ 0.33 \\ 0.17 \\ 2.67 \\ 0.33 \\ 0.83 \\ 0.17 \end{bmatrix} = \begin{bmatrix} -0.035 \\ -0.0033 \\ -0.0017 \\ -0.0267 \\ -0.0033 \\ -0.0083 \\ -0.0017 \end{bmatrix}$$

Primjer: Učenje

Nakon 2000 iteracija

$$\theta^1 = [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]$$

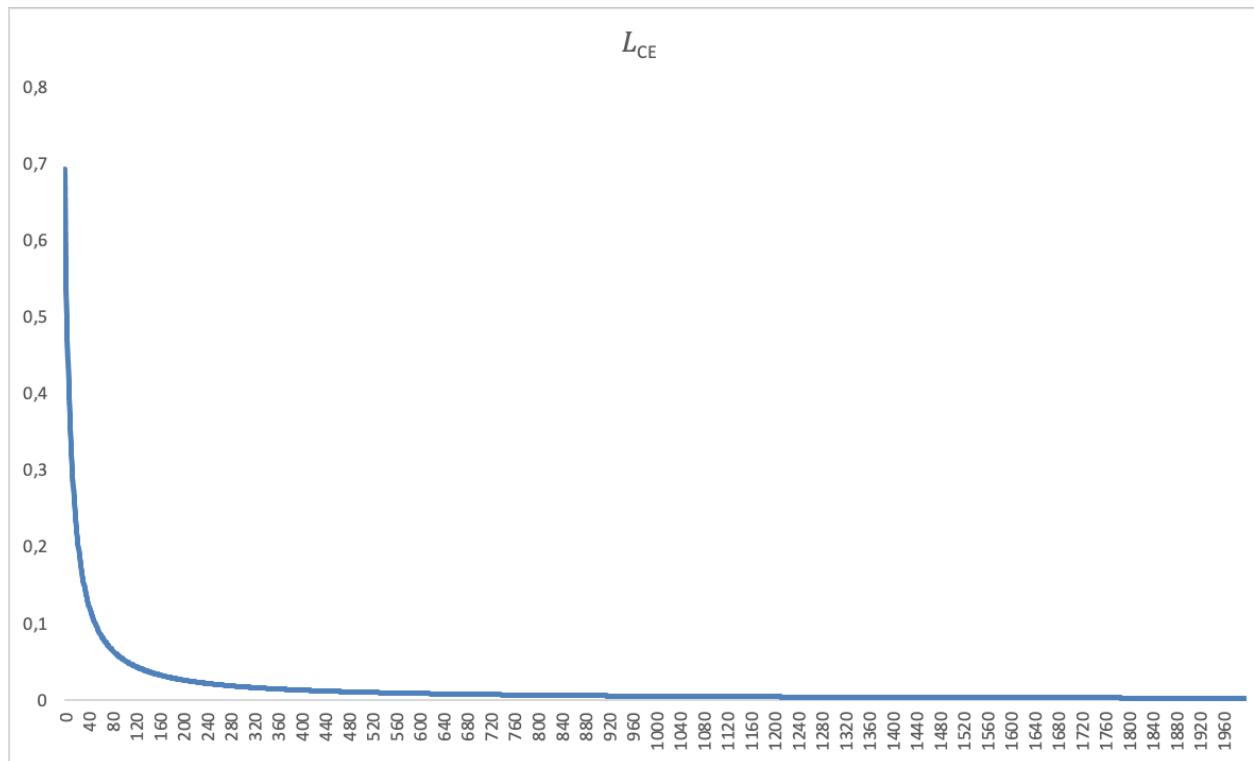
$$\theta^2 = [-0.035 \ -0.0033 \ -0.0017 \ -0.0267 \ -0.0033 \ -0.0083 \ -0.0017]$$

$$\theta^3 = [-0.037 \ -0.0095 \ -0.0014 \ -0.0412 \ -0.0053 \ -0.0072 \ -0.0014]$$

...

$$\theta^{1999} = [-0.3665 \ -0.8842 \ -0.0293 \ -0.8574 \ -0.1299 \ -0.1465 \ -0.0293]$$

$$\theta^{2000} = [-0.3665 \ -0.8843 \ -0.0293 \ -0.8574 \ -0.1299 \ -0.1465 \ -0.0293]$$



Primjer: Testiranje

$$Test = \left\{ \left(\begin{pmatrix} 18 \\ 15 \\ 1 \\ 4 \\ 1 \\ 5 \end{pmatrix}, 1 \right), \left(\begin{pmatrix} 8 \\ 4 \\ 0 \\ 0 \\ 0 \\ 5 \end{pmatrix}, 0 \right) \right\}$$

$$X = \begin{bmatrix} 18 & 8 \\ 15 & 4 \\ 1 & 0 \\ 4 & 0 \\ 1 & 0 \\ 5 & 5 \end{bmatrix}$$

$$Y = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\theta = [-0.3665 \quad -0.8843 \quad -0.0293 \quad -0.8574 \quad -0.1299 \quad -0.1465 \quad -0.0293]$$

$$\hat{Y} = \sigma(\theta \cdot \begin{bmatrix} X \\ 1 \end{bmatrix})$$

$$\hat{Y} = \sigma \left(\theta \cdot \begin{bmatrix} X \\ 1 \end{bmatrix} \right) = \sigma \left(\theta \cdot \begin{bmatrix} -0.3665 & -0.8843 & -0.0293 & -0.8574 & -0.1299 & -0.1465 & -0.0293 \end{bmatrix} \begin{bmatrix} 18 & 8 \\ 15 & 4 \\ 1 & 0 \\ 4 & 0 \\ 1 & 0 \\ 5 & 5 \\ 1 & 1 \end{bmatrix} \right) = \begin{bmatrix} 0.98 \\ 0.797 \end{bmatrix}$$

Uvod u obradu prirodnog jezika

9.2. Višeklasna logistička regresija (MaxEnt)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Višeklasna logistička regresija

- **Višeklasna logistička regresija** se još zove
 - Softmax regresija
 - Maxent klasifikator

- Klase $C = \{c_1, c_2, \dots, c_K\}$

- Funkcija klasifikacije softmax za vektor

$$z = [z_1, z_2, \dots, z_K]$$

$$\text{softmax}(z_j) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad 1 \leq i \leq K$$

$$\text{softmax}(z) = \left[\frac{e^{z_1}}{\sum_{j=1}^K e^{z_j}}, \frac{e^{z_2}}{\sum_{j=1}^K e^{z_j}}, \dots, \frac{e^{z_m}}{\sum_{j=1}^K e^{z_j}} \right]$$

- Nazivnik $\sum_{j=1}^K e^{z_j}$ služi za normalizaciju vrijednosti u vjerojatnosti

Višeklasna logistička regresija

- **Osobine**

$$f_i(x) = f_i(c, x) \text{ osobina } i \text{ za klasu } c$$

- Primjer, klasifikacija teksta u 3 klase: $\{+, -, 0\}$

osobina	definicija	W
$f_1(+, x)$	$\begin{cases} 1, & \text{ako "!" } \in d \\ 0, & \text{inače} \end{cases}$	-4.5
$f_1(-, x)$	$\begin{cases} 1, & \text{ako "!" } \in d \\ 0, & \text{inače} \end{cases}$	2.6
$f_1(0, x)$	$\begin{cases} 1, & \text{ako "!" } \in d \\ 0, & \text{inače} \end{cases}$	1.3

Višeklasna logistička regresija

- **Vjerojatnost za klasu c_i**

$$p(y = c_i | x) = \frac{e^{w_{c_i}x + b_{c_i}}}{\sum_{j=1}^K e^{w_{c_j}x + b_{c_j}}}$$

- Funkcija gubitka

$$\begin{aligned} L_{CE}(\hat{y}, y) &= - \sum_{k=1}^K 1\{y = c_k\} \log p(y = c_k | x) \\ &= - \sum_{k=1}^K 1\{y = c_k\} \log \frac{e^{w_{c_k}x + b_{c_k}}}{\sum_{j=1}^K e^{w_{c_j}x + b_{c_j}}} \end{aligned}$$

Višeklasna logistička regresija

- Gradijent

$$\begin{aligned}\frac{\partial L_{CE}(\hat{y}, y)}{\partial w_{c_m}} &= \left(1\{y = c_k\} - p(y = c_m | x) \right) x_m \\ &= \left(1\{y = c_k\} - \log \frac{e^{w_{c_m}x + b_{c_m}}}{\sum_{j=1}^K e^{w_{c_j}x + b_{c_j}}} \right) x_m\end{aligned}$$

Primjer: Podaci

$$TrainSet = \{(d_1, c_1), (d_2, c_1), (d_3, c_2), (d_4, c_3)\}$$

$$C = \{c_1, c_2, c_3\} \quad hot(C) = \{[1 \ 0 \ 0], [0 \ 1 \ 0], [0 \ 0 \ 1]\}$$

$$FeatureSet = \left\{ \begin{array}{l} \left[f_1^{c_1}(d_1) \quad f_2^{c_1}(d_1) \right], [1 \ 0 \ 0] \\ \left[f_1^{c_1}(d_2) \quad f_2^{c_1}(d_2) \right], [1 \ 0 \ 0] \\ \left[f_1^{c_2}(d_3) \quad f_2^{c_2}(d_3) \right], [0 \ 1 \ 0] \\ \left[f_1^{c_3}(d_4) \quad f_2^{c_3}(d_4) \right], [0 \ 0 \ 1] \end{array} \right\}$$

$$= \left\{ \begin{array}{l} \left[x_1^{d_1} \quad x_2^{d_1} \right], [1 \ 0 \ 0] \\ \left[x_1^{d_2} \quad x_2^{d_2} \right], [1 \ 0 \ 0] \\ \left[x_1^{d_3} \quad x_2^{d_3} \right], [0 \ 1 \ 0] \\ \left[x_1^{d_4} \quad x_2^{d_4} \right], [0 \ 0 \ 1] \end{array} \right\} = \{X, Y\}$$

Primjer: Predikcija

$$\Phi = (W, B)$$

$$\hat{Y} = \text{softmax}(XW + B)$$

$$\hat{Y} = \text{softmax} \left(\begin{bmatrix} x_1^{d_1} & x_2^{d_1} \\ x_1^{d_2} & x_2^{d_2} \\ x_1^{d_3} & x_2^{d_3} \\ x_1^{d_4} & x_2^{d_4} \end{bmatrix} \begin{bmatrix} w_1^{c_1} & w_1^{c_2} & w_1^{c_3} \\ w_2^{c_1} & w_2^{c_2} & w_2^{c_3} \end{bmatrix} + \begin{bmatrix} b^{c_1} \\ b^{c_2} \\ b^{c_3} \end{bmatrix} \right)$$

$$\hat{Y} = \text{softmax} \left(\begin{bmatrix} x_1^{d_1}w_1^{c_1} + x_2^{d_1}w_2^{c_1} + b^{c_1} & x_1^{d_1}w_1^{c_2} + x_2^{d_1}w_2^{c_2} + b^{c_2} & x_1^{d_1}w_1^{c_3} + x_2^{d_1}w_2^{c_3} + b^{c_3} \\ x_1^{d_2}w_1^{c_1} + x_2^{d_2}w_2^{c_1} + b^{c_1} & x_1^{d_2}w_1^{c_2} + x_2^{d_2}w_2^{c_2} + b^{c_2} & x_1^{d_2}w_1^{c_3} + x_2^{d_2}w_2^{c_3} + b^{c_3} \\ x_1^{d_3}w_1^{c_1} + x_2^{d_3}w_2^{c_1} + b^{c_1} & x_1^{d_3}w_1^{c_2} + x_2^{d_3}w_2^{c_2} + b^{c_2} & x_1^{d_3}w_1^{c_3} + x_2^{d_3}w_2^{c_3} + b^{c_3} \\ x_1^{d_4}w_1^{c_1} + x_2^{d_4}w_2^{c_1} + b^{c_1} & x_1^{d_4}w_1^{c_2} + x_2^{d_4}w_2^{c_2} + b^{c_2} & x_1^{d_4}w_1^{c_3} + x_2^{d_4}w_2^{c_3} + b^{c_3} \end{bmatrix} \right)$$

$$\hat{Y} = \text{softmax} \left(\begin{bmatrix} x^{d_1} \cdot w^{c_1} + b^{c_1} & x^{d_1} \cdot w^{c_2} + b^{c_2} & x^{d_1} \cdot w^{c_3} + b^{c_3} \\ x^{d_2} \cdot w^{c_1} + b^{c_1} & x^{d_2} \cdot w^{c_2} + b^{c_2} & x^{d_2} \cdot w^{c_3} + b^{c_3} \\ x^{d_3} \cdot w^{c_1} + b^{c_1} & x^{d_3} \cdot w^{c_2} + b^{c_2} & x^{d_3} \cdot w^{c_3} + b^{c_3} \\ x^{d_4} \cdot w^{c_1} + b^{c_1} & x^{d_4} \cdot w^{c_2} + b^{c_2} & x^{d_4} \cdot w^{c_3} + b^{c_3} \end{bmatrix} \right)$$

Primjer: Predikcija

$$\hat{Y} = \text{softmax} \left(\begin{bmatrix} x^{d_1} \cdot w^{c_1} + b^{c_1} & x^{d_1} \cdot w^{c_2} + b^{c_2} & x^{d_1} \cdot w^{c_3} + b^{c_3} \\ x^{d_2} \cdot w^{c_1} + b^{c_1} & x^{d_2} \cdot w^{c_2} + b^{c_2} & x^{d_2} \cdot w^{c_3} + b^{c_3} \\ x^{d_3} \cdot w^{c_1} + b^{c_1} & x^{d_3} \cdot w^{c_2} + b^{c_2} & x^{d_3} \cdot w^{c_3} + b^{c_3} \\ x^{d_4} \cdot w^{c_1} + b^{c_1} & x^{d_4} \cdot w^{c_2} + b^{c_2} & x^{d_4} \cdot w^{c_3} + b^{c_3} \end{bmatrix} \right)$$

$$\hat{Y} = \text{softmax} \left(\begin{bmatrix} d_1 c_1 & d_1 c_2 & d_1 c_3 \\ d_2 c_1 & d_2 c_2 & d_2 c_3 \\ d_3 c_1 & d_3 c_2 & d_3 c_3 \\ d_4 c_1 & d_4 c_2 & d_4 c_3 \end{bmatrix} \right)$$

$$\hat{Y} = \begin{bmatrix} e^{d_1 c_1} / \sum_i e^{d_1 c_i} & e^{d_1 c_2} / \sum_i e^{d_1 c_i} & e^{d_1 c_3} / \sum_i e^{d_1 c_i} \\ e^{d_2 c_1} / \sum_i e^{d_2 c_i} & e^{d_2 c_2} / \sum_i e^{d_2 c_i} & e^{d_2 c_3} / \sum_i e^{d_2 c_i} \\ e^{d_3 c_1} / \sum_i e^{d_3 c_i} & e^{d_3 c_2} / \sum_i e^{d_3 c_i} & e^{d_3 c_3} / \sum_i e^{d_3 c_i} \\ e^{d_4 c_1} / \sum_i e^{d_4 c_i} & e^{d_4 c_2} / \sum_i e^{d_4 c_i} & e^{d_4 c_3} / \sum_i e^{d_4 c_i} \end{bmatrix}$$

Neka je $z^{d^i} = \sum_j e^{d_i c_j}$

$$\hat{Y} = \begin{bmatrix} \frac{e^{d_1 c_1}}{z^{d^1}} & \frac{e^{d_1 c_2}}{z^{d^1}} & \frac{e^{d_1 c_3}}{z^{d^1}} \\ \frac{e^{d_2 c_1}}{z^{d^2}} & \frac{e^{d_2 c_2}}{z^{d^2}} & \frac{e^{d_2 c_3}}{z^{d^2}} \\ \frac{e^{d_3 c_1}}{z^{d^3}} & \frac{e^{d_3 c_2}}{z^{d^3}} & \frac{e^{d_3 c_3}}{z^{d^3}} \\ \frac{e^{d_4 c_1}}{z^{d^4}} & \frac{e^{d_4 c_2}}{z^{d^4}} & \frac{e^{d_4 c_3}}{z^{d^4}} \end{bmatrix}$$

Primjer: Učenje

$$\Theta^{t+1} = \Theta^t - \eta \nabla_{\Theta} L$$

za d_1 imamo klasu $c_1 = [1 \quad 0 \quad 0]$

$$\begin{bmatrix} w_1^{c_1} & w_1^{c_2} & w_1^{c_3} \\ w_2^{c_1} & w_2^{c_2} & w_2^{c_3} \\ b^{c_1} & b^{c_2} & b^{c_3} \end{bmatrix}^{t+1} = \begin{bmatrix} w_1^{c_1} & w_1^{c_2} & w_1^{c_3} \\ w_2^{c_1} & w_2^{c_2} & w_2^{c_3} \\ b^{c_1} & b^{c_2} & b^{c_3} \end{bmatrix}^t - \eta \begin{bmatrix} \frac{\partial L}{\partial w_1^{c_1}} & \frac{\partial L}{\partial w_1^{c_2}} & \frac{\partial L}{\partial w_1^{c_3}} \\ \frac{\partial L}{\partial w_2^{c_1}} & \frac{\partial L}{\partial w_2^{c_2}} & \frac{\partial L}{\partial w_2^{c_3}} \\ \frac{\partial L}{\partial b^{c_1}} & \frac{\partial L}{\partial b^{c_2}} & \frac{\partial L}{\partial b^{c_3}} \end{bmatrix}$$

$$\Theta^{t+1} = \Theta^t - \eta \begin{bmatrix} -\left(1 - \log\left(\frac{e^{d_1 c_1}}{z^{d_1}}\right)\right) x_1^{d_1} & -\left(0 - \log\left(\frac{e^{d_1 c_2}}{z^{d_1}}\right)\right) x_1^{d_1} & -\left(0 - \log\left(\frac{e^{d_1 c_3}}{z^{d_1}}\right)\right) x_1^{d_1} \\ -\left(1 - \log\left(\frac{e^{d_1 c_1}}{z^{d_1}}\right)\right) x_2^{d_1} & -\left(0 - \log\left(\frac{e^{d_1 c_2}}{z^{d_1}}\right)\right) x_2^{d_1} & -\left(0 - \log\left(\frac{e^{d_1 c_3}}{z^{d_1}}\right)\right) x_2^{d_1} \\ -\left(1 - \log\left(\frac{e^{d_1 c_1}}{z^{d_1}}\right)\right) 1 & -\left(0 - \log\left(\frac{e^{d_1 c_2}}{z^{d_1}}\right)\right) 1 & -\left(0 - \log\left(\frac{e^{d_1 c_3}}{z^{d_1}}\right)\right) 1 \end{bmatrix}$$

Uvod u obradu prirodnog jezika

9.1. Ekstrakcija informacija i prepoznavanje imenovanih entiteta (Information Extraction and Named Entity Recognition)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Ekstrakcija informacija (IE)

- Sustavi za ekstrakciju informacija (IE)
 - pronalaženje i razumijevanje relevantnih dijelova teksta
 - skupljanje informacija iz mnogih izvora teksta
 - produkcija strukturne reprezentacije relevantnih informacija
 - relacije
 - baza znanja
 - Ciljevi:
 1. organizacija informacija tako da budu korisne ljudima
 2. postavljanje informacija u semantički preciznom obliku čime se omogućava daljnje zaključivanje uz pomoć računalnih algoritama

Ekstrakcija informacija (IE)

- IE sustavi ekstraktiraju čiste, činjenične informacije:
 - Ugrubo: Tko je učinio nešto nekome kada?
- Npr:
 - Prikupljanje zarade, profita, članova odbora, sjedišta, itd. iz izvještaja kompanije
 - Sjedišta ABC Trade d.o.o. i globalna sjedišta kombinirane ABC Trade Grupe, su locirane u Splitu, Hrvatska
 - *sjedišta("ABC Trade d.o.o.", "Split Hrvatska")*
 - Učenje lijek-gen interakcije iz znanstvene medicinske literature

IE na niskom nivou

- Dostupno – relativno popularno – u aplikacijama kao Apple ili Google mail i kod web indeksiranja
- Izgleda da su temeljena na regularnim izrazima i listama naziva

Prepoznavanje imenovanih entiteta (NER)

- Važan podzadatak: pronalaženje i klasifikacija naziva u tekstu, npr:
 - odluka nezavisnog kandidata Ivana Mimača da obustavi njegovu podršku za manjinsku stranku Rada je zvučala dramatično, ali u buduće neće prijetiti stabilnosti. Kada su, nakon izbora 2010 godine, Ivan, Ante Jukić, Marija Anitovska i Milanić odlučili podržati Rad, dali su samo dvije garancije: povjerenje i opskrbu.

Prepoznavanje imenovanih entiteta (NER)

- Važan podzadatak: **pronalaženje** i klasifikacija naziva u tekstu, npr:
 - odluka nezavisnog kandidata **Ivana Mimača** da obustavi njegovu podršku za manjinsku stranku **Rada** je zvučala dramatično, ali u buduće neće prijetiti stabilnosti. Kada su, nakon izbora **2010** godine, **Ivan, Ante Jukić, Marija Anitovska** i **Milanić** odlučili podržati **Rad**, dali su samo dvije garancije: povjerenje i opskrbu.

Prepoznavanje imenovanih entiteta (NER)

- Važan podzadatak: pronalaženje i **klasifikacija** naziva u tekstu, npr:
 - odluka nezavisnog kandidata **Ivana Mimača** da obustavi njegovu podršku za manjinsku stranku **Rada** je zvučala dramatično, ali u buduće neće prijetiti stabilnosti. Kada su, nakon izbora **2010** godine, **Ivan**, **Ante Jukić**, **Marija Anitovska** i **Milanić** odlučili podržati **Rad**, dali su samo dvije garancije: povjerenje i opskrbu.

Osoba
Datum
Lokacija
Organizacija

Prepoznavanje imenovanih entiteta (NER)

- Korištenje:
 - imenovani entiteti se mogu indeksirati, povezati, itd.
 - Sentiment se može pridružiti kompanijama ili produktima
 - Mnoge IE relacije su veze između imenovanih entiteta
 - Za odgovaranje na pitanja, odgovori su često imenovani entiteti
- Konkretno:
 - Mnoge Web stranice označavaju razne entitete, s vezama na biografiju, tematske stranice i slično
 - Reuter's OpenCalais, Evri, AlchemyAPI, Yahoo's Term Extraction
 - Apple/Google/Microsoft/ ... pametni prepoznavatelji za sadržaj dokumenta

Uvod u obradu prirodnog jezika

9.2. Evaluacija prepoznavanja imenovanih entiteta (Evaluation of Named Entity Recognition)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Zadatak NER prepoznavanja

- Zadatak: Predvidjeti entitete u tekstu

– govornik	O	
– Ministarstva	ORG	
– Vanjskih	ORG	
– Poslova	ORG	
– Ivan	PER	
– Ivanić	PER	
– rekao	O	
– je	O	
– Vjesniku	ORG	
:	:	

Standardna evaluacija je po entitetu,
ne po pojavnici (tokenu)

Zadatak NER prepoznavanja

- Zadatak: Predvidjeti entitete u tekstu

– govornik	O	
– Ministarstva	ORG	
– Vanjskih	ORG	
– Poslova	ORG	
– Ivan	PER	
– Ivanić	PER	
– rekao	O	
– je	O	
– Vjesniku	ORG	
:	:	

Standardna evaluacija je po entitetu,
ne po pojavnici (tokenu)

Zadatak NER prepoznavanja

- Zadatak: Predvidjeti entitete u tekstu

– govornik	O
– Ministarstva	ORG
– Vanjskih	ORG
– Poslova	ORG
– Ivan	PER
– Ivanić	PER
– rekao	O
– je	O
– Vjesniku	ORG
:	:



sustav
detektirao
sustav nije
detektirao

točno	pogrešno
2	0
1	0

Preciznost: 100%

Odziv: 66%

Preciznost/Odziv/F1 za IE/NER

- Odziv i preciznost su odlične mjere za dohvati informacija (IR) i kategorizaciju teksta
- Mjera se ponaša čudno kod IE/NER kada ima **graničnih grešaka** (koje su **česte**):
 - Prva Banka za Splićane je proglašila...
- Ovim se obuhvaća i lažno pozitivne i lažno negativne vrijednosti
- Izbor ničega bi bilo bolje
- Neke druge metrike (npr. MUC bodovanje) daju djelomičan utjecaj (prema složenim pravilima)

IE na niskom nivou

- Dostupno – relativno popularno – u aplikacijama kao Apple ili Google mail i kod web indeksiranja
- Izgleda da su temeljena na regularnim izrazima i listama naziva

Prepoznavanje imenovanih entiteta (NER)

- Važan podzadatak: pronalaženje i klasifikacija naziva u tekstu, npr:
 - odluka nezavisnog kandidata Ivana Mimača da obustavi njegovu podršku za manjinsku stranku Rada je zvučala dramatično, ali u buduće neće prijetiti stabilnosti. Kada su, nakon izbora 2010 godine, Ivan, Ante Jukić, Marija Anitovska i Milanić odlučili podržati Rad, dali su samo dvije garancije: povjerenje i opskrbu.

Prepoznavanje imenovanih entiteta (NER)

- Važan podzadatak: **pronalaženje** i klasifikacija naziva u tekstu, npr:
 - odluka nezavisnog kandidata **Ivana Mimača** da obustavi njegovu podršku za manjinsku stranku **Rada** je zvučala dramatično, ali u buduće neće prijetiti stabilnosti. Kada su, nakon izbora **2010** godine, **Ivan, Ante Jukić, Marija Anitovska** i **Milanić** odlučili podržati **Rad**, dali su samo dvije garancije: povjerenje i opskrbu.

Prepoznavanje imenovanih entiteta (NER)

- Važan podzadatak: pronalaženje i **klasifikacija** naziva u tekstu, npr:
 - odluka nezavisnog kandidata **Ivana Mimača** da obustavi njegovu podršku za manjinsku stranku **Rada** je zvučala dramatično, ali u buduće neće prijetiti stabilnosti. Kada su, nakon izbora **2010** godine, **Ivan**, **Ante Jukić**, **Marija Anitovska** i **Milanić** odlučili podržati **Rad**, dali su samo dvije garancije: povjerenje i opskrbu.

Osoba
Datum
Lokacija
Organizacija

Prepoznavanje imenovanih entiteta (NER)

- Korištenje:
 - imenovani entiteti se mogu indeksirati, povezati, itd.
 - Sentiment se može pridružiti kompanijama ili produktima
 - Mnoge IE relacije su veze između imenovanih entiteta
 - Za odgovaranje na pitanja, odgovori su često imenovani entiteti
- Konkretno:
 - Mnoge Web stranice označavaju razne entitete, s vezama na biografiju, tematske stranice i slično
 - Reuter's OpenCalais, Evri, AlchemyAPI, Yahoo's Term Extraction
 - Apple/Google/Microsoft/ ... pametni prepoznavatelji za sadržaj dokumenta

Uvod u obradu prirodnog jezika

9.3. Modeli sekvenci za prepoznavanje imenovanih entiteta (Sequence Models for Named Entity Recognition)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

NER i model sekvence iz strojnog učenja

Treniranje

1. Prikupi skup reprezentativnih dokumenata za treniranje
2. Označi svaku pojavniciu entitetskom klasom ili ostalo (O)
3. Oblikuj ekstraktore osobina prikladne za tekst i klase
4. Treniraj sekvenički klasifikator za predviđanje oznaka iz podataka

Testiranje

1. Primi skup dokumenata za testiranje
2. Pokreni zaključivanje pomoću modela sekvence radi označavanja svake pojavnice
3. Prikladno vrati prepoznate entitete

Kodne klase za označavanje sekvence

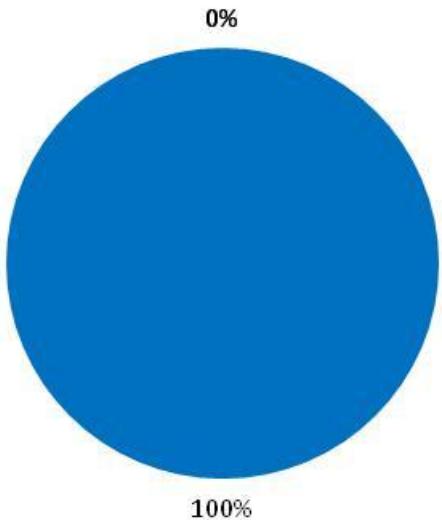
	IO kodiranje	IOB kodiranje
Luka	PER	B-PER
pokazuje	O	O
Sanji	PER	B-PER
Ivo	PER	B-PER
Ivičevu	PER	I-PER
novu	O	O
sliku	O	O

Osobine za označavanje kod sekvenci

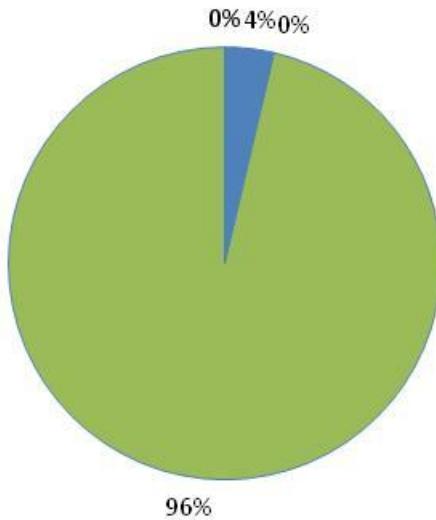
- Riječi
 - Trenutna riječ (kao naučeni rječnik)
 - Prethodna/sljedeća riječ (sadržaj)
- Druge vrste naslijedjenih lingvističkih klasifikacija
 - POS
- Sadržaj oznake
 - prethodna (i možda sljedeća) oznaka

Osobine: Podnizovi riječi

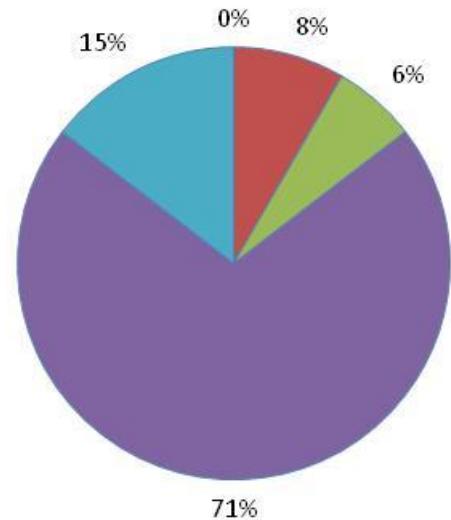
oxa



:



field



lijek
tvrtka
film
mjesto
osoba

Cotrimoxazole

Wethersfield

Rambo: First Blood

Osobine: Oblik riječi

- Oblik riječi
 - pridruživanje pojednostavljenog prikaza riječi koji kodira attribute kao što su duljina, velika/mala slova, brojevi, grčka slova, unutrašnje interpunkcije, itd.

Varicella-zoster	Xx-xxx
mRNA	xXXX
CPA1	XXXd

Uvod u obradu prirodnog jezika

9.4. Maksimalna entropija Markovljevog modela (Maximum Entropy Markov Models)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Problemi sekvenci

- Mnogi problemi OPJ imaju podatke kao sekvence znakova, riječi, fraza, linija, rečenica ...
- Naš zadatak je označavanje svakog elementa sekvence

POS označavanje

N	V	C	V	A	N
Stručnjaci	navode	kako	će	metalurški	sektor

NER

PERS	O	O	O	ORG	ORG
Matić	diskutira	o	budućnosti	Fakulteta	strojarstva

Segmentacija riječi

B	B	I	I	B	I	B	I	B	B
而	相	对	于	这	些	品	牌	的	价

Segmentacija teksta



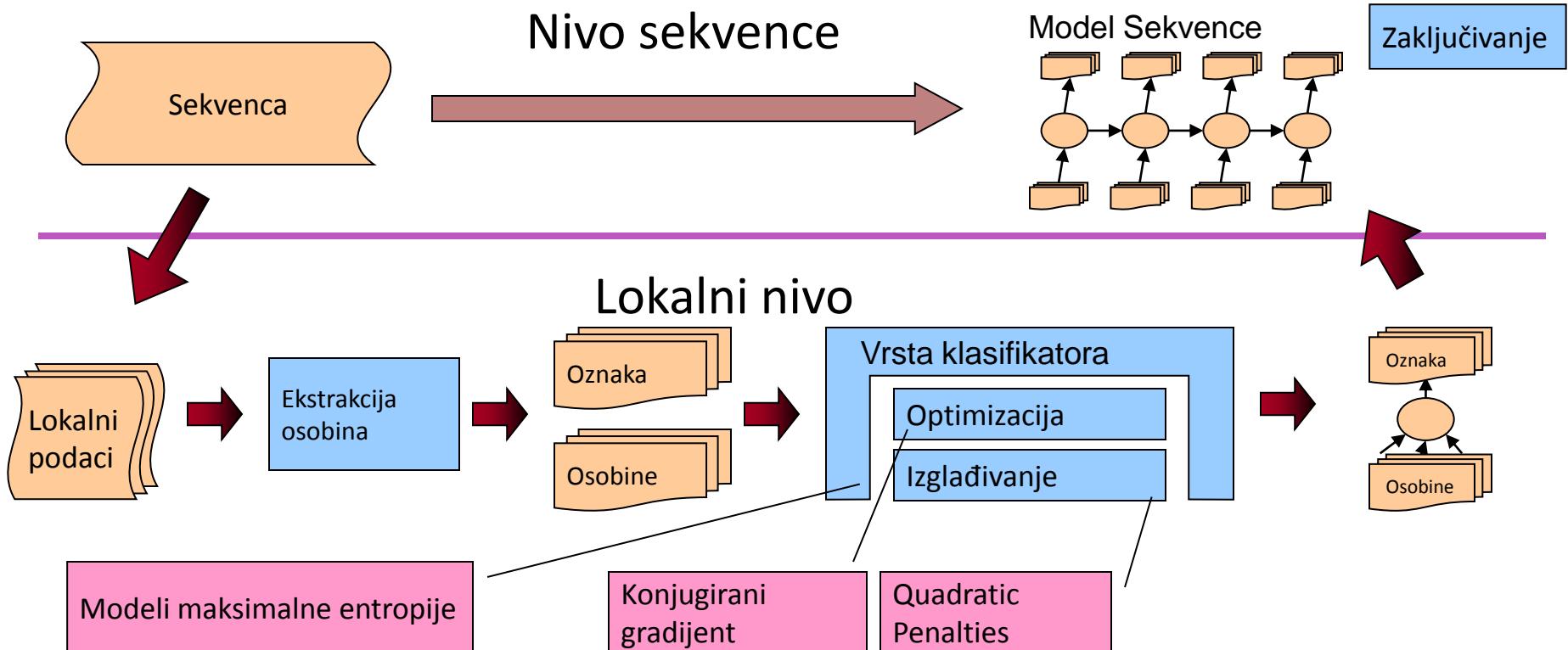
MEMM zaključivanje

- **Uvjetni Markovljev model** (Conditional Markov Model) tj. **Markovljev model maksimalne entropije (MEMM)** je klasifikator koji donosi odluku ovisno o opservacijama i **prethodnim odlukama**.
- Naš zadatak je označiti svaki element sekvene.

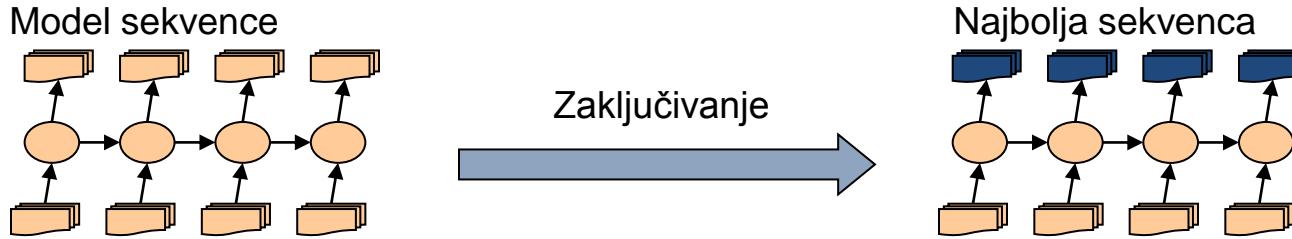
Lokalni sadržaj					Točka odluke
-3	-2	-1	0	1	
N	V	V	???	???	
Dionice	su	pale	22.6	%	

Osobine	
W_0	22.6
W_{+1}	%
W_{-1}	pale
T_{-1}	V
$T_{-1} T_{-2}$	V V
imaBroj?	da
...	...

Sustav za zaključivanje

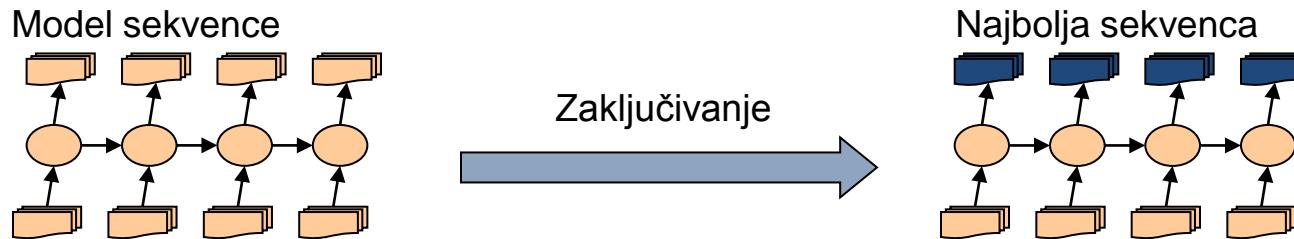


Pohlepno (greedy) zaključivanje



- Pohlepno zaključivanje
 - Počinjemo s lijeva i koristimo klasifikator na svakoj poziciji kako bi pridružili oznaku
 - Klasifikator može ovisiti o prethodnoj odluci kao i o promatranom podatku
- Prednosti
 - Brz, ne zahtjeva dodatnu memoriju
 - Jednostavan za implementaciju
 - Obogaćivanjem osobina tako da uključuju opservacije s desna mogu se postići dobri rezultati
- Mane
 - Pohlepan. Rade se greške od kojih se ne može oporaviti.

Zaključivanje zrakama (Beam)



- Zaključivanje zrakama
 - Na svakoj poziciji zadrži najboljih K kompletiranih sekvenci
 - Proširivanje sekvence se vrši lokalno
 - Proširivanjem s oznakom se dobiva novi skup K kompletiranih sekvenci
- Prednosti
 - Brz, zrake veličine 3-5 su u većini slučajeva dobre kao i egzaktno zaključivanje
 - Jednostavno za implementirati (ne zahtjeva dinamičko programiranje)
- Mane
 - Nije egzaktno: globalno najbolje sekvence mogu ispasti sa zrake

Viterbi zaključivanje



- Viterbi zaključivanje
 - Dinamičko programiranje ili memoizacija
 - Zahtjeva mali prozor utjecaja stanja (npr. prethodna dva stanja su relevantna)
- Prednosti
 - Egzaktan: Globalno najbolja sekvenca se dobiva
- Mane
 - Teže za implementirati duže interakcije stanja (ali zaključivanje zrakama ne dopušta duže interakcije)

Uvjetna slučajna polja

- Još jedan model sekvenci: Conditional Random Fields (CRF)
- Uvjetni model cijele sekvence u odnosu na ulančavanje lokalnih modela

$$P(\vec{c} \mid \vec{d}, \lambda) = \frac{\exp \sum_i \lambda_i f_i(c, d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c', d)}$$

- Prostor od C je sada prostor sekvenci
 - Ako osobine f_i ostaju lokalne, onda se uvjetna vjerodostojnost može izračunati dinamičkim programiranjem
- Treniranje je sporije, ali CRF izbjegava natjecanje pristranosti
- U praksi obično rade dobro kao i MEMM

Uvod u obradu prirodnog jezika

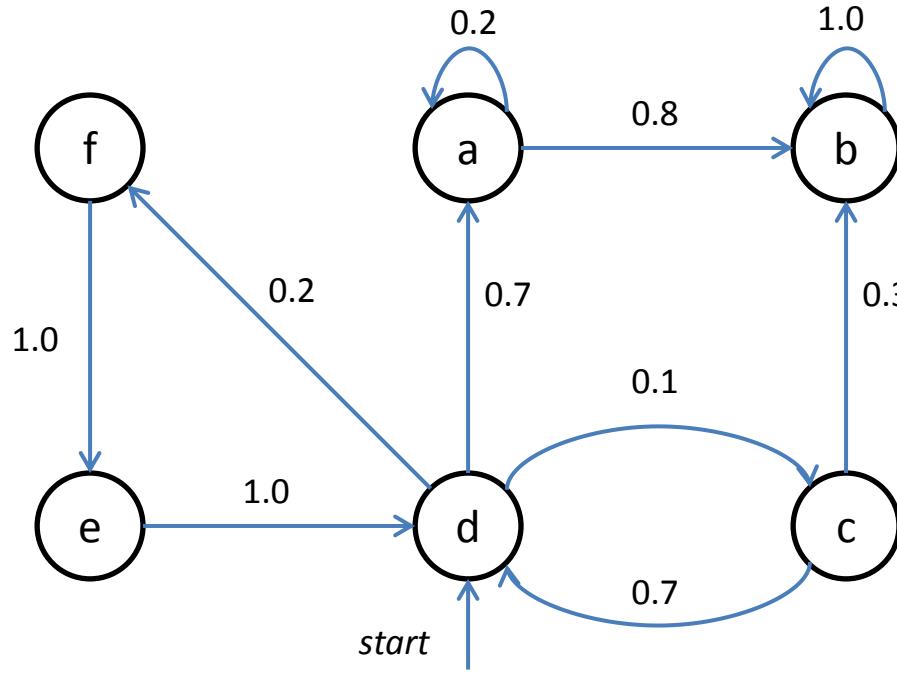
9.5. Markovljev model (Markov model)

Branko Žitko

Markovljev model

- Sekvenca slučajnih varijabli koja nije nezavisna
- Primjer
 - prognoza vremena
 - tekst
- Svojstva:
 - $P(X_{t+1} = s_k | X_1, \dots, X_t) = P(X_{t+1} = s_k | X_t)$
 - Vremenska invarijanta
 - $P(X_2 = s_k | X_1)$
- Definicija:
 - u terminima tranzicijske matrice A i vjerojatnosti početnog stanja Π

(Vidljivi) Markovljev model (VMM)



$$\begin{aligned} P(X_1, \dots, X_T) &= P(X_1) P(X_2 | X_1) P(X_3 | X_1, X_2) \dots P(X_T | X_1, X_2, \dots, X_{T-1}) \\ &= P(X_1) P(X_2 | X_1) P(X_3 | X_2) \dots P(X_T, X_{T-1}) \\ &= \end{aligned}$$

$$\begin{aligned} P(d, a, b) &= P(X_1=d) P(X_2=a | X_1=d) P(X_3=b | X_2=a) \\ &= 1.0 * 0.7 * 0.8 \\ &= 0.56 \end{aligned}$$

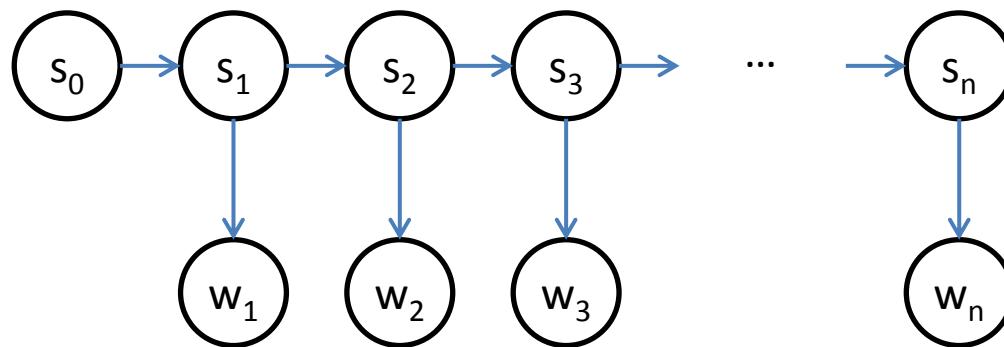
Skriveni Markovljev model

- Hidden Markov Model (HMM)
 - Promatra sekvencu simbola
 - Sekvenca stanja koja vodi do generiranja simbola je skrivena
- Definicija
 - Q = skup stanja
 - O = skup opservacija, napravljena iz rječnika
 - q_0, q_f = specijalna stanja (početno i završno stanje)
 - A = matrica vjerojatnosti tranzicija stanja
 - B = matrica vjerojatnosti emisije simbola
 - Π = vjerojatnosti početnog stanja
 - $\mu = (A, B, \Pi)$ = potpuni probabilistički model

Skriveni Markovljev model

- Koristi se za modeliranja sekvenca stanja i sekvenca opservacija
- Primjer:

$$P(S|W) = \prod_i P(s_i|s_{i-1}) P(w_i|s_i)$$

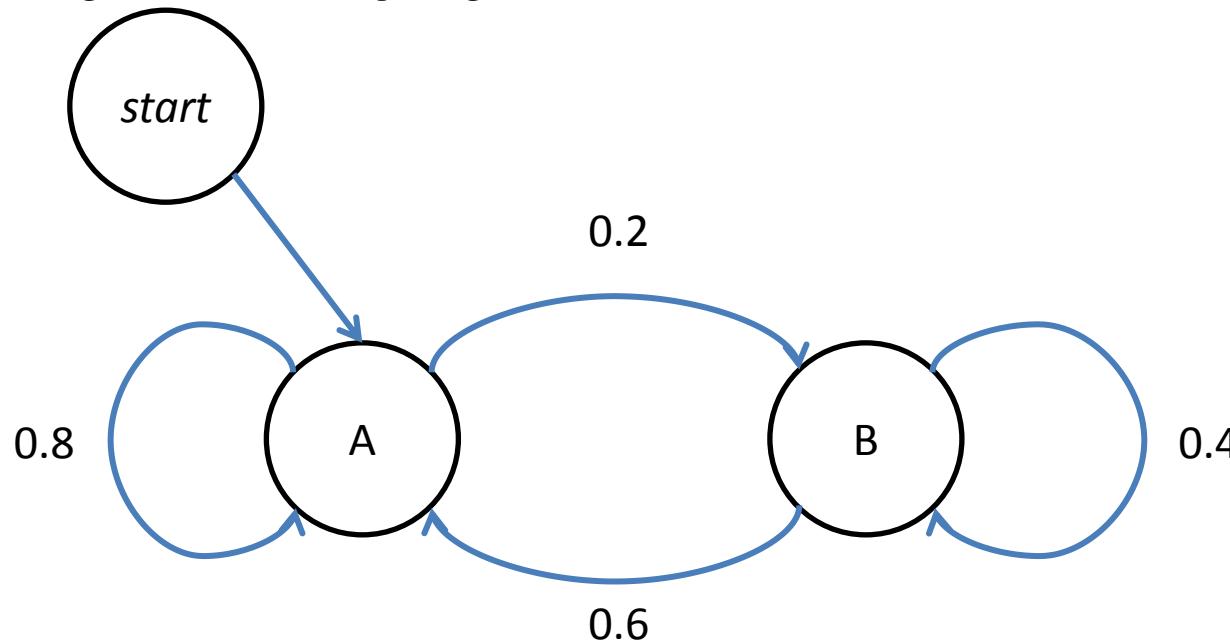


Generativni algoritam

1. Izaberi početak iz Π
2. za $t = 1..T$
 1. Prijedji u sljedeće stanje temeljem A
 2. Emitiraj opservaciju temeljem B

Vjerojatnosti skrivenog Markovljevog modela

Vjerojatnosti prijelaza



Emisijeske vjerojatnosti

	x	y	z
A	0.7	0.2	0.1
B	0.3	0.5	0.2

Svi parametri skrivenog Markovljevog modela

- **Početak**

- $P(A|start) = 1.0 \quad P(B|start) = 0.0$

- **Tranzicije**

- $P(A|A) = 0.8 \quad P(A|B) = 0.6$

- $P(B|A) = 0.2 \quad P(B|B) = 0.4$

- **Emisije**

- $P(x|A) = 0.7 \quad P(y|A) = 0.2 \quad P(z|A) = 0.1$

- $P(x|B) = 0.3 \quad P(y|B) = 0.5 \quad P(z|B) = 0.2$

Opservacijska sekvenca "yz"

- Počevši u stanju A, koliki je $P(yz)$?
- Moguće sekvence stanja
 - AA
 - AB
 - BA
 - BB
- $$\begin{aligned} P(yz) &= P(yz|AA) + P(yz|AB) + P(yz|BA) + P(yz|BB) \\ &= 0.8 \times 0.2 \times 0.8 \times 0.1 \\ &\quad + 0.8 \times 0.2 \times 0.2 \times 0.2 \\ &\quad + 0.2 \times 0.5 \times 0.4 \times 0.2 \\ &\quad + 0.2 \times 0.5 \times 0.6 \times 0.1 \\ &= 0.0128 + 0.0064 + 0.0080 + 0.0060 = 0.0332 \end{aligned}$$

HMM zadaci

- Zadaci
 - Za dani model $\mu=(A,B,\Pi)$ pronađi vjerojatnost opservacija $P(O|\mu)$
 - Za dane opservacije O , koji je slijed stanja (X_1, \dots, X_{T+1})
 - Za dane opservacije O i sve moguće modele μ , odredi model koji najbolje opisuje O
- Dekodiranje
 - označiti svaku pojavniciu oznakom
- Vjerodostojnost opservacije
 - klasificiraj sekvencu
- Učenje
 - treniraj model da odgovara empiričkim podacima

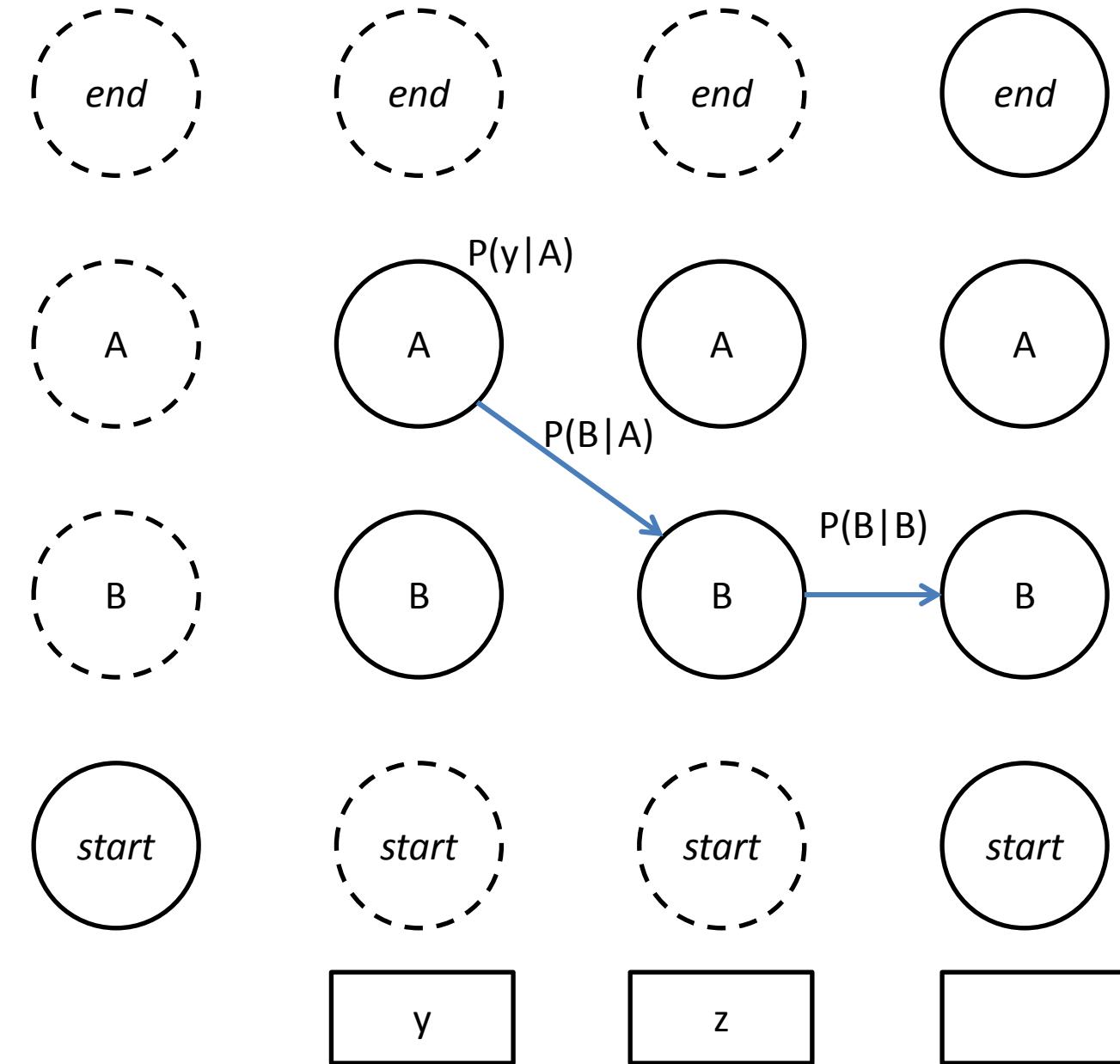
Zaključivanje

- Pronađi najvjerojatniji slijed oznaka, za dani slijed riječi
 - $t^* = \operatorname{argmax}_t P(t|w)$
- Za dani model μ jeli moguće izračunati $P(t|w)$ za sve vrijednosti od t
- U praksi, postoji previše kombinacija
- Moguća rješenja:
 - koristiti pretraživanje po zrakama (beam search) – djelomična hipoteza
 - U svakom stanju, čuvati k najboljih hipoteza do sada
 - Ne mora dobro raditi

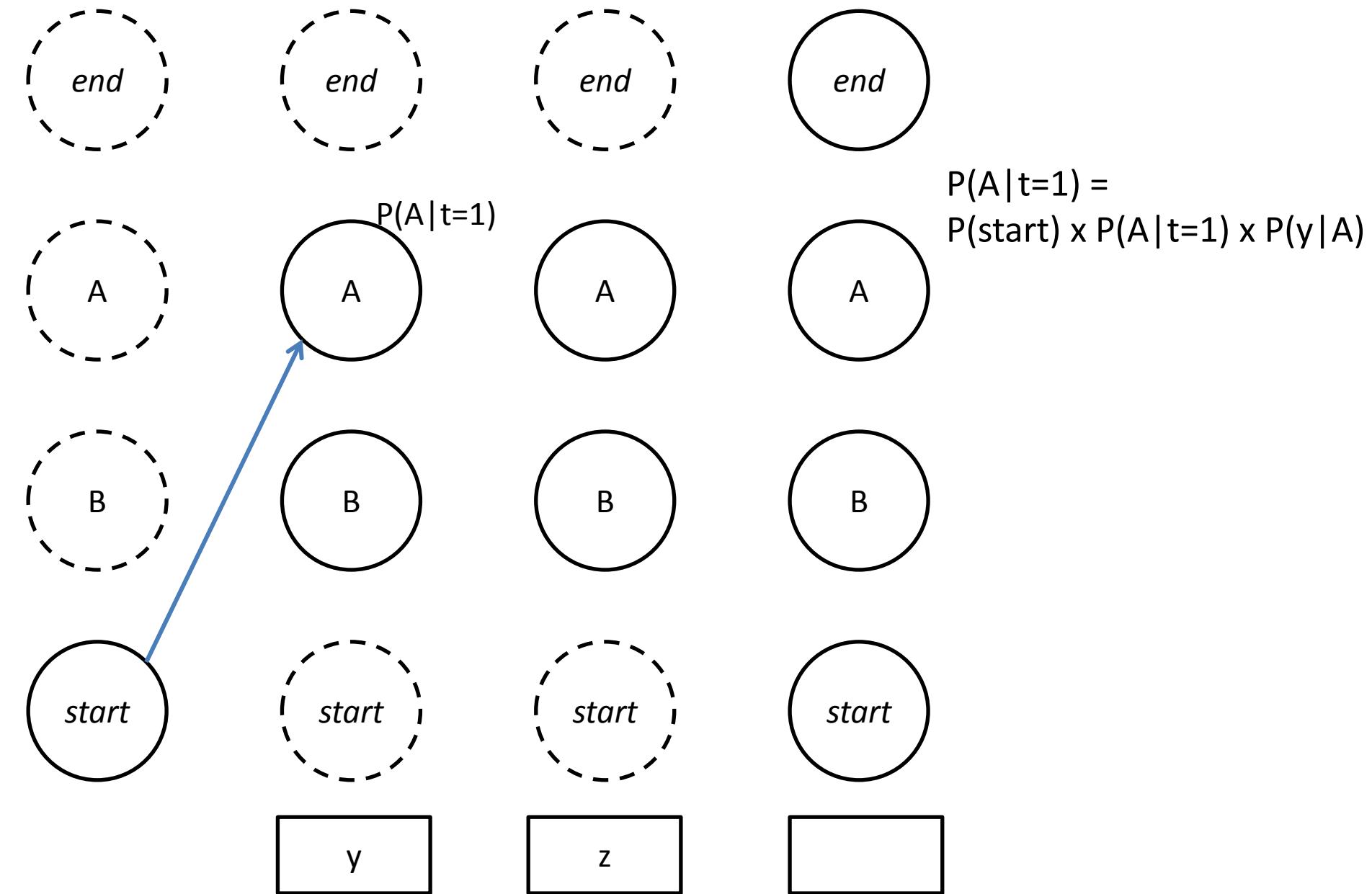
Viterbi algoritam

- Pronađi najbolju putanju do opservacije O i stanja S
- Karakteristike
 - koristi dinamičko programiranje
 - memoizacija
 - praćenje unatrag

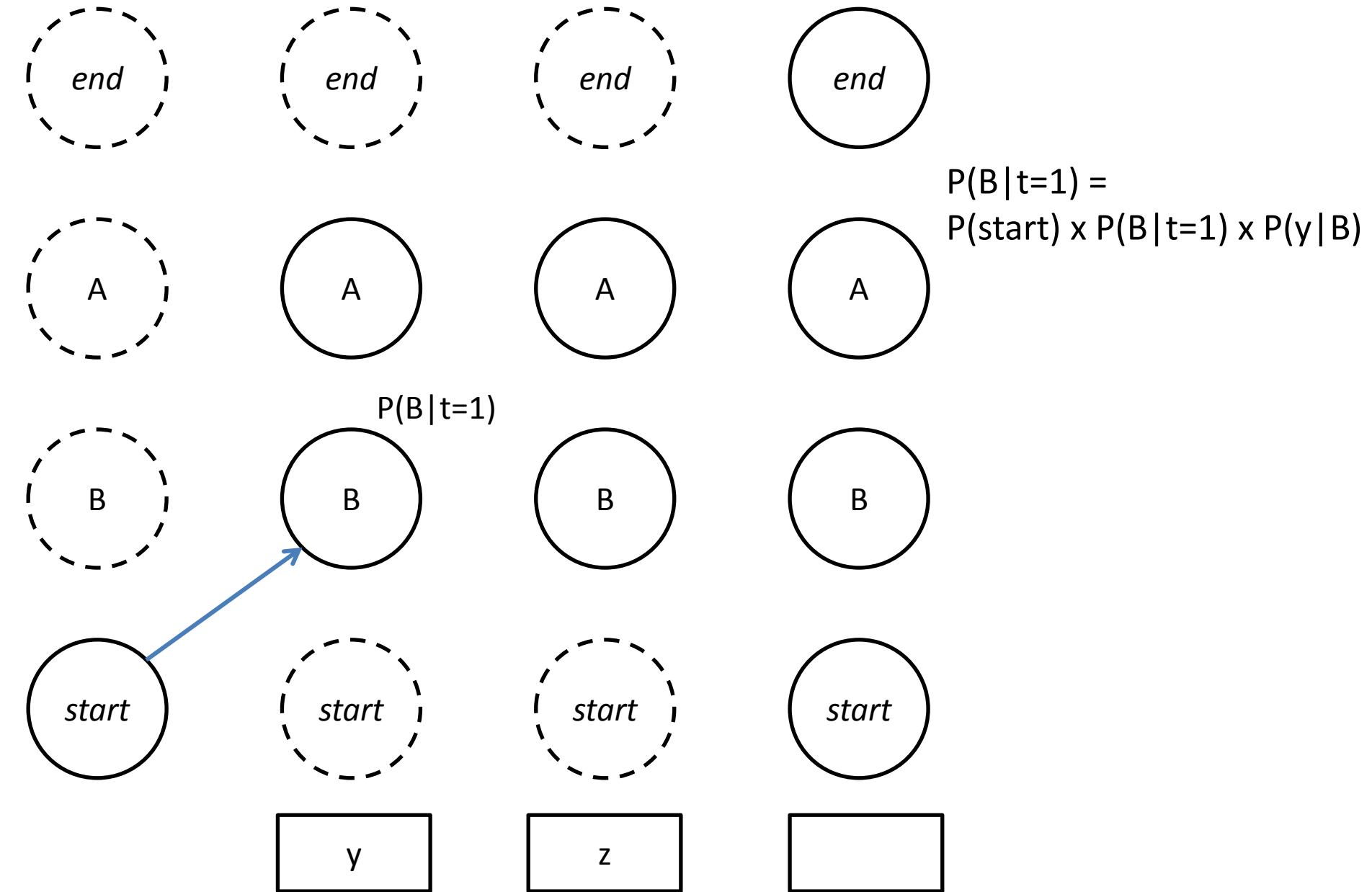
HMM rešetka (trellis)



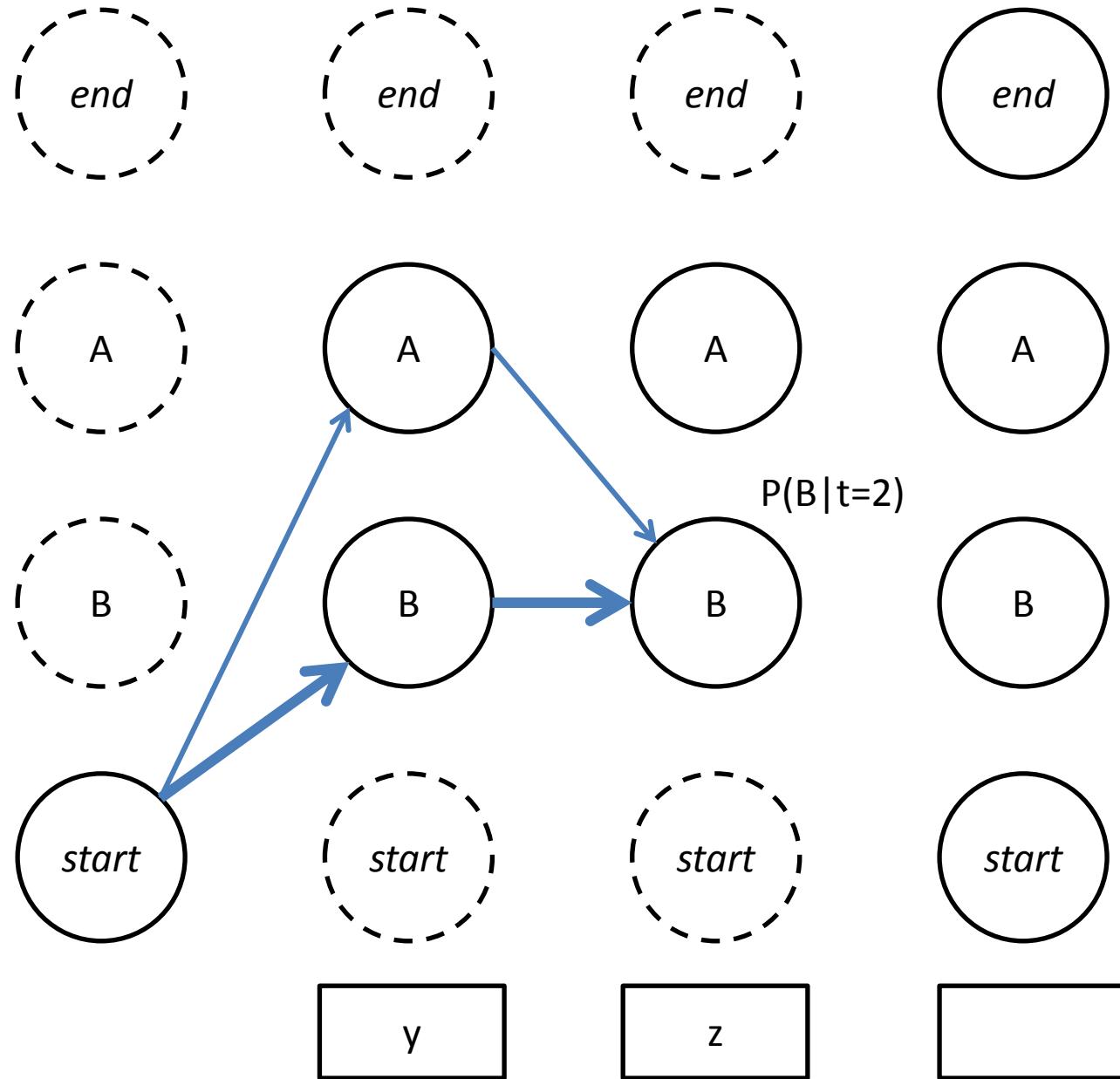
HMM rešetka (trellis)



HMM rešetka (trellis)

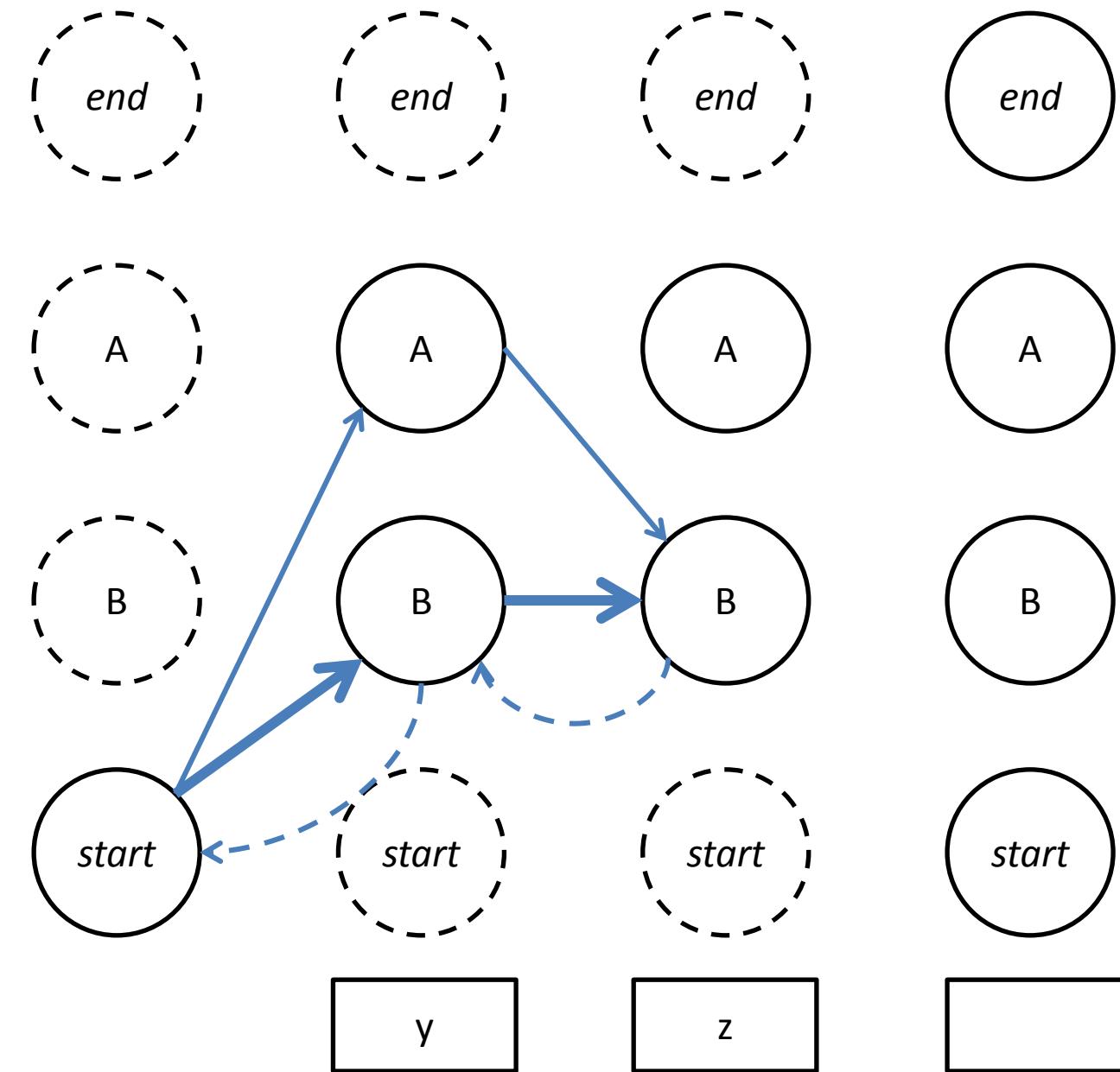


HMM rešetka (trellis)

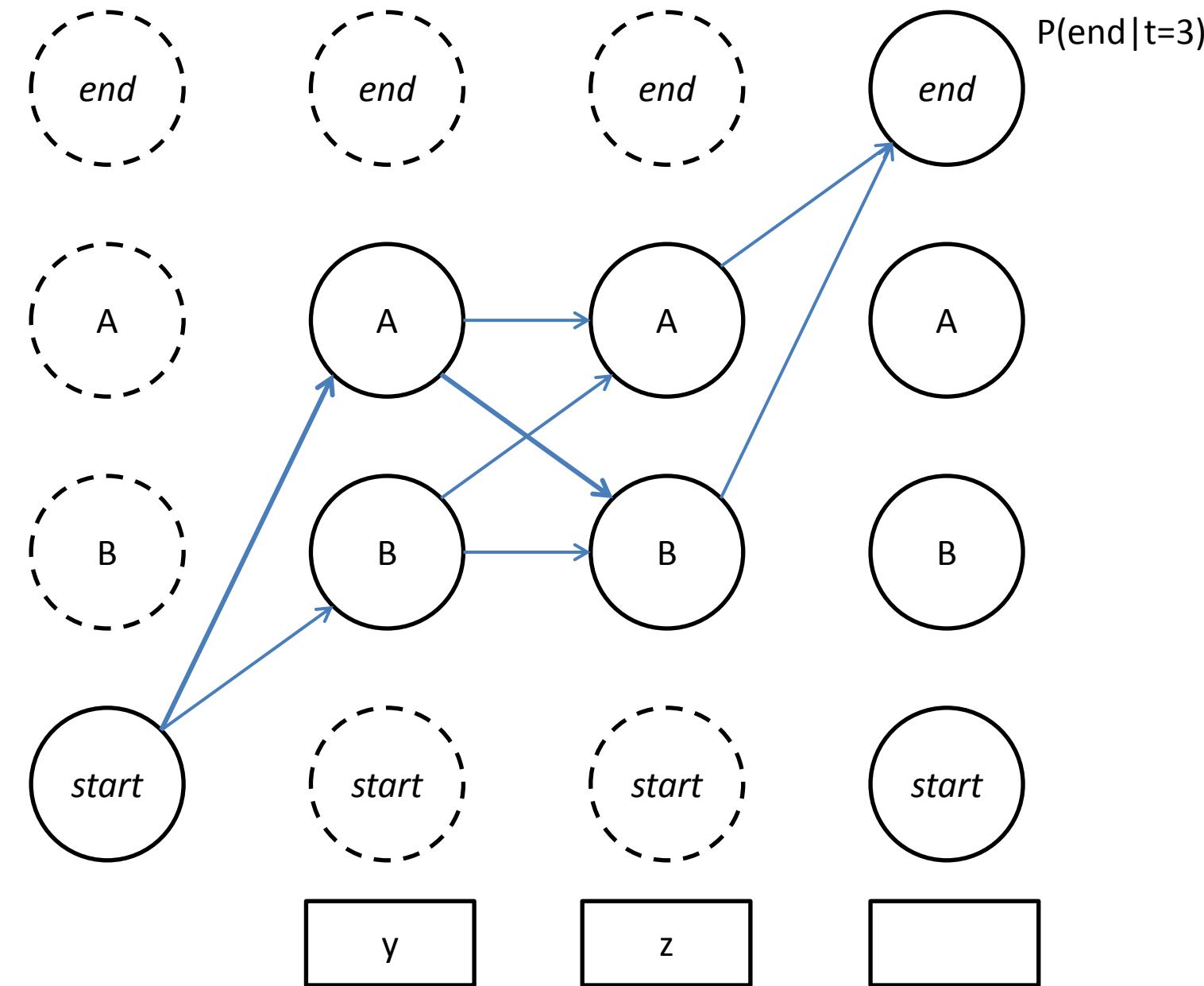


$$P(B|t=2) = \max(P(A|t=1) \times P(B|A) \times P(z|B), P(B|t=1) \times P(B|B) \times P(z|B))$$

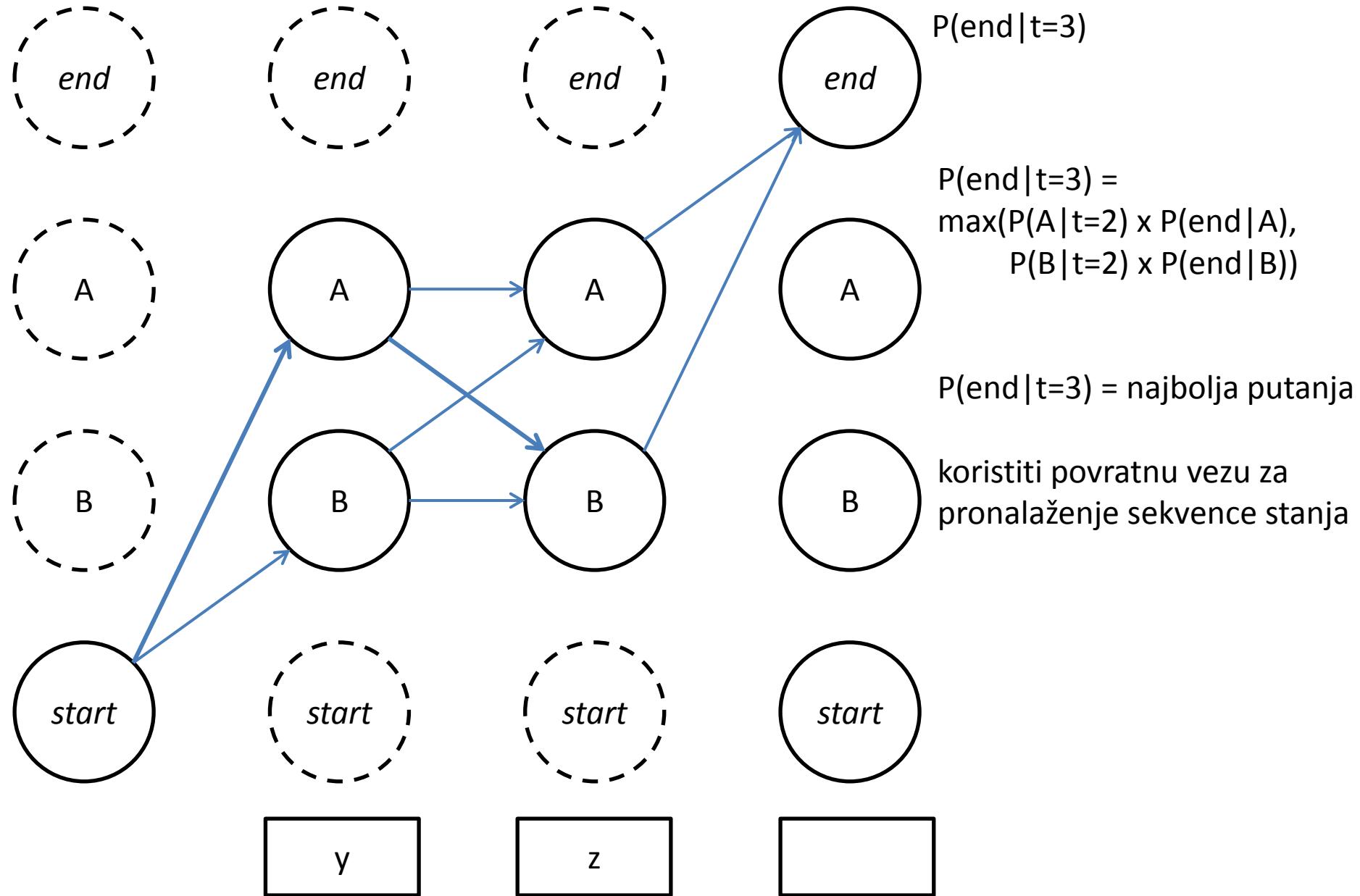
HMM rešetka (trellis)



HMM rešetka (trellis)



HMM rešetka (trellis)



HMM učenje

- Nadzirano
 - Sekvence za treniranje su označene
- Nenadzirano
 - Sekvence za treniranje nisu označene
 - Poznat broj stanja
- Polunadzirano
 - Neke sekvence za treniranje su označene

Nadzirano HMM učenje

- Procjeni vjerojatnosti tranzicija koristeći maksimalnu vjerodostojnost

$$a_{i,j} = \frac{\text{broj}(q_t = s_i, q_{t+1} = s_j)}{\text{broj}(q_t)}$$

- Procjeni vjerojatnosti opservacije koristeći maksimalnu vjerodostojnost

$$b_j(k) = \frac{\text{broj}(q_i = s_j, o_i = v_k)}{\text{broj}(q_i = s_j)}$$

- Koristi izglađivanje

Nenadzirano HMM učenje

- Dano
 - slijed opservacija
- Cilj
 - izgraditi HMM
- Koristiti metodu maksimizacije očekivanja
(EM - Expectation Maximization)
 - naprijed-nazad (forward-backward) (Baum-Welch) algoritam
 - Baum-Welch pronalazi približno rješenje za $P(O|\mu)$

Baum-Welch

- Algoritam
 - Postavi slučajnim izborom parametre za HMM
 - Dok parametri konvergiraju ponavljam
 - E korak (očekivanje) – odredi vjerojatnosti za razne sekvence stanja koje generiraju opservaciju (forward-backward)
 - M korak (maksimizacija) – ponovno procjeni parametre za HMM temeljem dobivenih vjerojatnosti
- Rezultati
 - algoritam garantira da će se kod svake iteracije vjerodostojnost od $P(O | \mu)$ povećavati
 - može se zaustaviti bilo kada i dobiti djelomično rješenje
 - konvergira prema lokalnom maksimumu

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

Nepromjenjive

Prilozi kada gdje kamo zašto

Prijedlozi od iza po

Veznici i ili ali

Uzvici ah oh eh

Čestice ne zar evo

Glagoli

Glavni

vidi
stajale

Pomoćni

će/ćemo/ćete

Pridjevi dobro/bolje/najbolje

Zamjenice on ona onaj

Brojevi jedan drugoj trima

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 višesmisljene
- I ljudi ponekad imaju problema s određivanjem vrste riječi.

Koliko je teško POS označavanje

- Oko 11% riječi u Brown korpusu su više značne 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 više značne

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% (Ijudski)

Kako poboljšati nadzirane rezultate?

- Izgraditi bolje osobine!

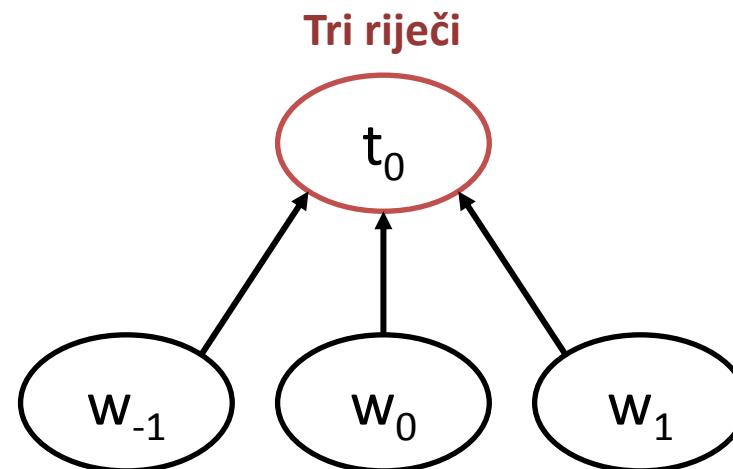
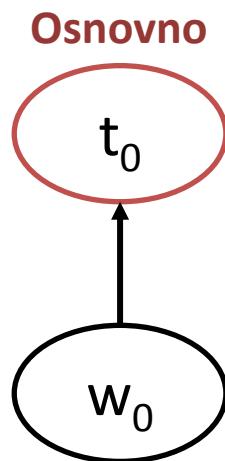
S
V P N A
Vrti se oko sunca

- Ovo smo mogli popraviti izgradnjom osobine koja gleda sljedeću riječ

A
NP N V A
Dobili ljudi ostaju dobri

- 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**.

Uvod u obradu prirodnog jezika

13.1. Sintaktičke strukture: sastavno vs. ovisnosno (Syntactic structures: constituency vs. dependency)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Dva pogleda na lingvističke strukture

1. Sastavni (constituency) - struktura fraze

– Struktura fraze organizira riječi u ugniježđene sastavnice

– Kako znamo što je **sastavnica**?

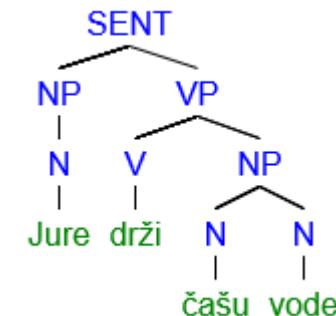
- Distribucija: sastavnica se ponaša kao jedinica koja se može pojaviti na različitim mjestima:

- Jure je govorio [svojoj djeci] o [štetnosti droga].
 - Jure je govorio o [štetnosti droga] [svojoj djeci].
 - *Jure je govorio droga svojoj štetnosti o djeci.

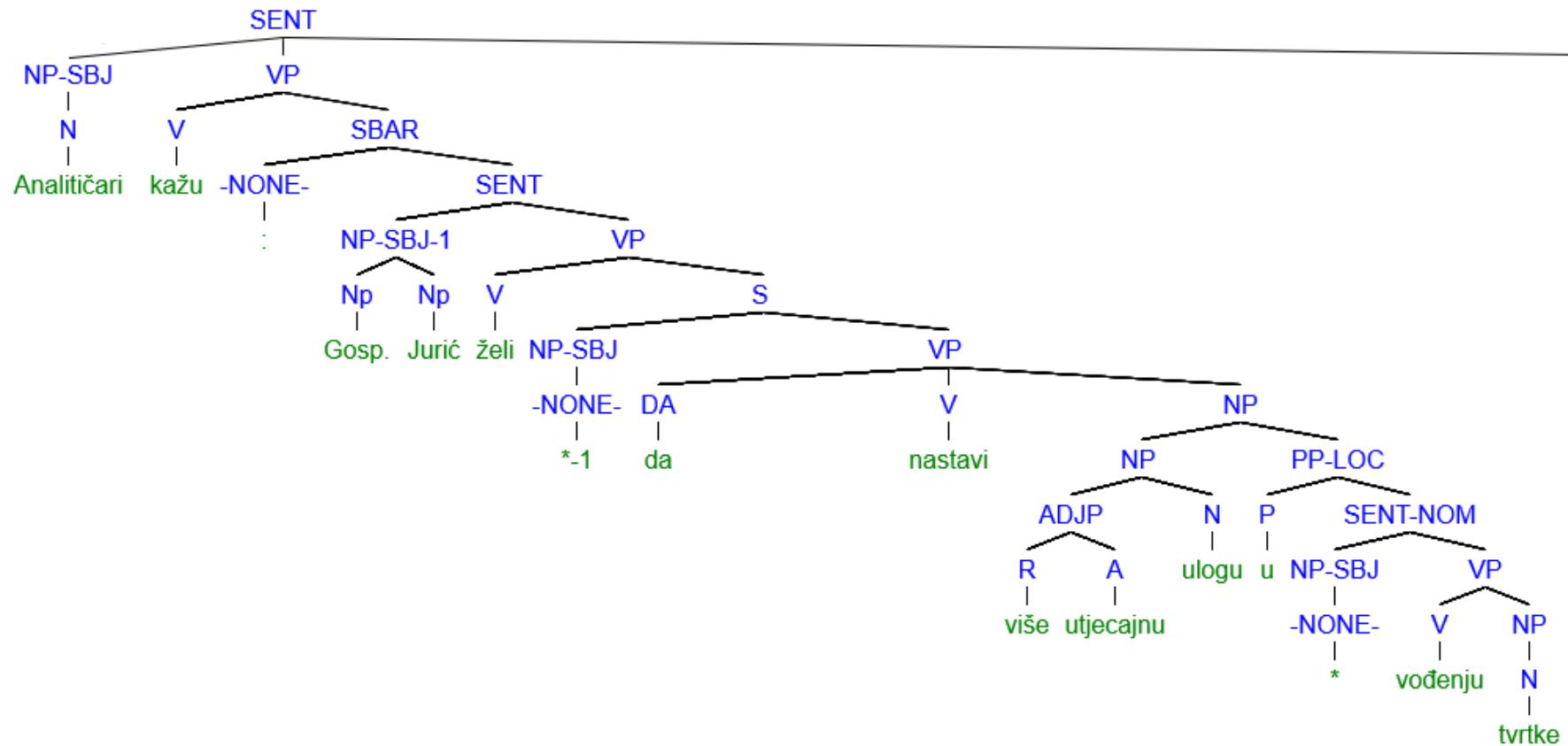
- Supstitucija/ekspanzija/zamjenične-forme

- Sjedio sam [na kutiji/baš na toj kutiji/tamo]

- Koordinacija, regularna interna struktura, nema upada, fragmenti, semantike, ...



Struktura fraze



Struktura fraze s glavom

- $\text{VP} \rightarrow \dots \text{V}^* \dots$
 - $\text{NP} \rightarrow \dots \text{N}^* \dots$
 - $\text{ADJP} \rightarrow \dots \text{A}^* \dots$
 - $\text{ADVP} \rightarrow \dots \text{Q}^* \dots$
- X-bar teorija**
- $\text{SBAR(Q)} \rightarrow \text{S} \mid \text{SINV} \mid \text{SQ} \rightarrow \dots \text{NP VP} \dots$
 - Još manjinski tipovi fraza
 - QP (kvantifikatorska fraza u NP)
 - CONJP (višeriječne konstrukcije kao "i tako dalje")
 - INTJ (uzvici ah, bravo, ...)

Ovisnost

- Ovisnost pokazuje koje riječi ovise (mijenjaju ili su argumenti) o drugim riječima

Taj dječak stavlja malu kornjaču na mekani tepih.

Ovisnost

- Ovisnost pokazuje koje riječi ovise (mijenjaju ili su argumenti) o drugim riječima

The diagram illustrates dependencies between words in the sentence "Taj dječak stavlja malu kornjaču na mekani tepih." using three colored arcs: red, purple, and green. The red arc connects the subject 'Taj dječak' to the verb 'stavlja'. The purple arc connects the verb 'stavlja' to the object 'malu kornjaču'. The green arc connects the prepositional phrase 'na mekani tepih.' to the verb 'stavlja'.

Taj dječak stavlja malu kornjaču na mekani tepih.

Uvod u obradu prirodnog jezika

13.2. Parsiranje (Parsing)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

"Klasično" parsiranje

- Napravljena gramatika (CFG) i leksikon

$S \rightarrow NP\ VP$	$NN \rightarrow interest$
$NP \rightarrow (DT)\ NN$	$NNS \rightarrow rates$
$NP \rightarrow NN\ NNS$	$NNS \rightarrow raises$
$NP \rightarrow NNP$	$VBP \rightarrow interest$
$VP \rightarrow V\ NP$	$VBZ \rightarrow rates$

- Korištenje gramatika/dokaza za dokazivanje stabla parsiranja iz riječi
- Skaliranje se pokazuje vrlo lošim, i ne obuhvaća opće primjere:

Fed raises interest rates 0.5% in effort to control inflation

- Minimalna gramatika: 36 stabala
- Jednostavna gramatika s 10 pravila: 592 stabala
- Realna sveobuhvatna gramatika: milioni stabala

"Klasično" parsiranje: problem i rješenje

- Dodavanje kategoričkih ograničenja u gramatici radi ograničavanja malo vjerojatnih parsiranja rečenice
 - ali gramatika gubi na robusnosti
 - oko 30% rečenica neće biti parsirane
- Manje ograničena gramatika može parsirati više rečenica
 - ali jednostavne rečenice dobivaju više različitih stabala bez načina da se izabere jedno od njih
- Potreban je mehanizam koji pronalazi najvjerojatnije stablo parsiranja rečenice
 - Statističko parsiranje omogućava rad s malim gramatikama koje dozvoljavaju milione stabala parsiranja rečenica i vrlo brzo pronalazi najbolje stablo.

Banka stabala

- Izgradnja banke stabala se čini mnogo sporijim i manje korisnim od izgradnje gramatike
- Ali banka stabala ima brojne koristi:
 - ponovna upotrebljivost
 - mnogi parseri, POS označavači, ...
 - Vrijedni resursi za lingviste
 - Široka pokrivenost
 - Frekvencije i distribucije
 - Način evaluacije sustava

Penn TreeBank

((S
 (NP-SBJ (DT The) (NN move))
 (VP (VBD followed)
 (NP
 (NP (DT a) (NN round))
 (PP (IN of)
 (NP
 (NP (JJ similar) (NNS increases))
 (PP (IN by)
 (NP (JJ other) (NNS lenders)))
 (PP (IN against)
 (NP (NNP Arizona) (JJ real) (NN estate) (NNS loans))))))
 (., ,)
 (S-ADV
 (NP-SBJ (-NONE- *))
 (VP (VBG reflecting)
 (NP
 (NP (DT a) (VBG continuing) (NN decline))
 (PP-LOC (IN in)
 (NP (DT that) (NN market))))))
 (. .)))

[Marcus et al. 1993, *Computational Linguistics*]

Uvod u obradu prirodnog jezika

13.3. Eksponencijalni problem parsiranja

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

"Klasično" parsiranje

- Ključna odluka kod parsiranja: na koji način "povezati" različite sastavnice
 - prijedložne, priložne fraze, fraze čestica, infinitive, koordinacije, ...

Odbor je potvrdio [svoja preuzimanja] [s vanjskom tvrtkom]

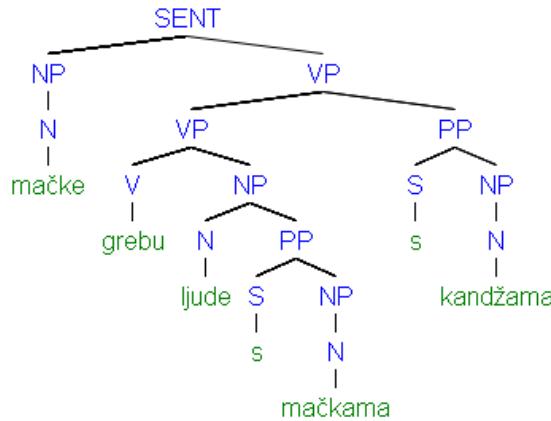
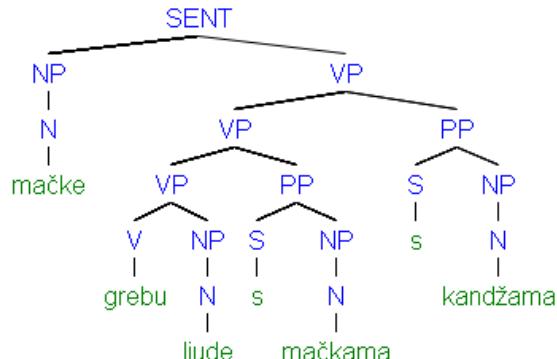
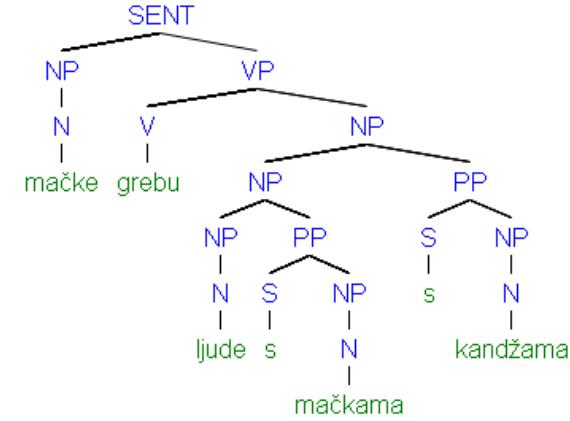
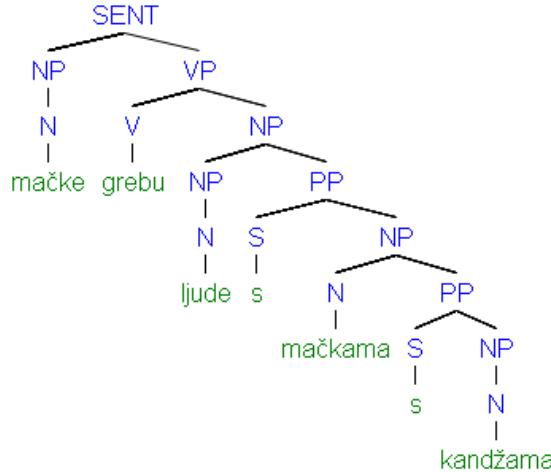
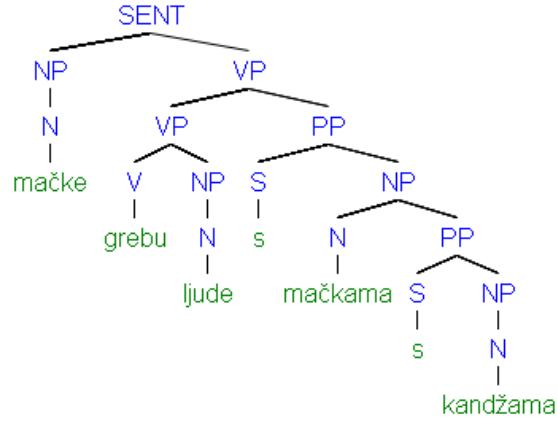
[iz Splita]

[za 25kn po dionici]

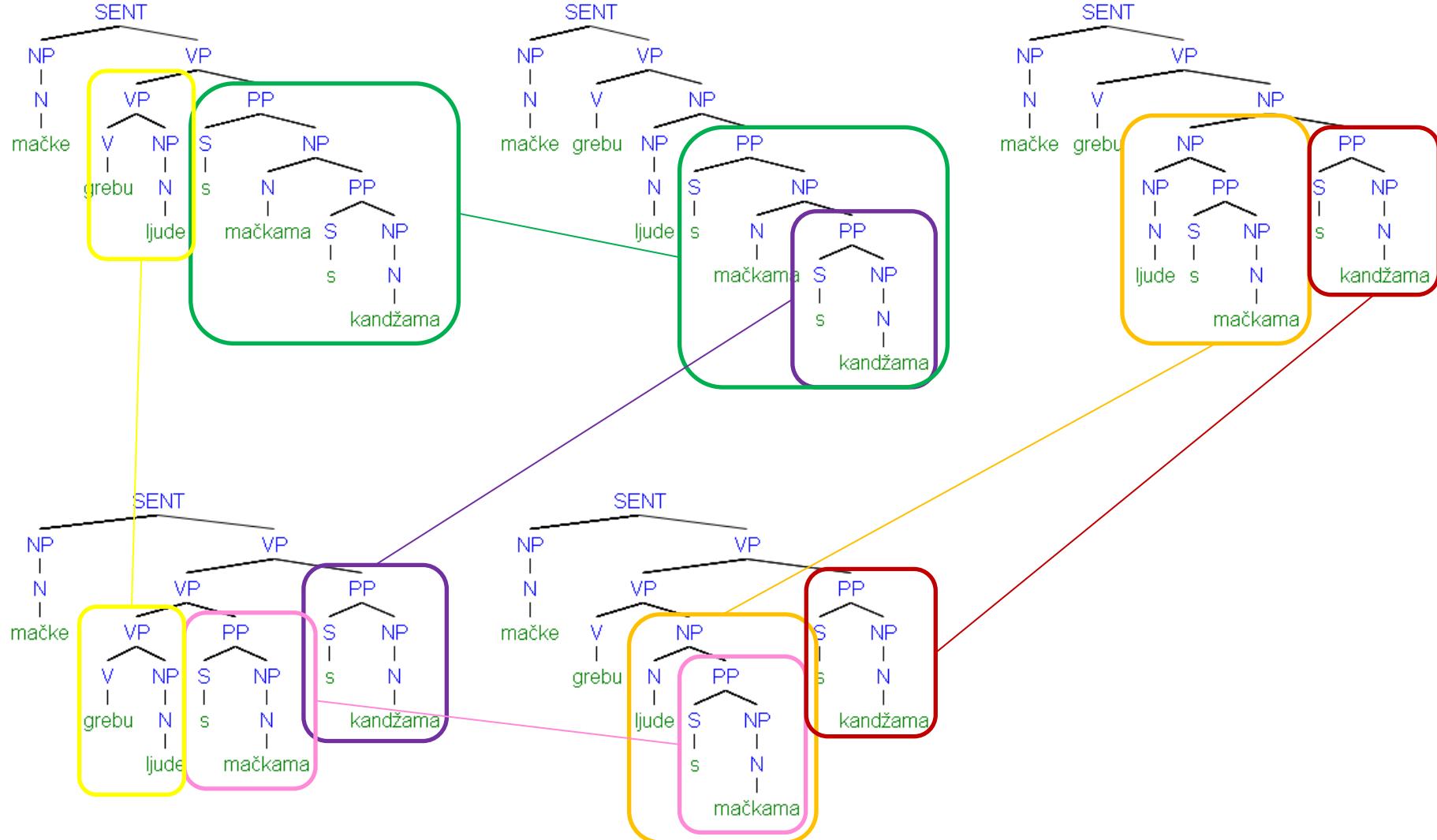
[na svom mjesecnom sastanku].

Catalanovi brojevi: $C_n = (2n)!/[(n+1)!n!]$ – eksponencijalni rast

Dva problema: 1. ponavljanje posla



Dva problema: 1. ponavljanje posla



Dva problema: 2. izbor stabla parsiranja

- Kako dobro povezati
[Ona vidi čovjeka s teleskopom](#)
- Riječi su dobri prediktori povezivanja
 - Čak i sa odsustvom potpunog razumijevanja

[Susjedna Saudijska Arabija je poslala više od 1000 vojnika u Bahrein ...](#)

[Izraelske vlasti su prekršile dogovor s washingtonskom administracijom ...](#)

- Statistički parseri će pokušati iskoristiti takvu statistiku

Uvod u obradu prirodnog jezika

14.1. Probabilističko parsiranje (Probabilistic parsing)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Gramatika strukture fraze

$S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$VP \rightarrow V\ NP\ PP$

$NP \rightarrow NP\ NP$

$NP \rightarrow NP\ PP$

$NP \rightarrow N$

$NP \rightarrow \epsilon$

$PP \rightarrow P\ NP$

$N \rightarrow primati$

$N \rightarrow kape$

$N \rightarrow nose$

$N \rightarrow glavi$

$V \rightarrow primati$

$V \rightarrow kape$

$V \rightarrow nose$

$P \rightarrow na$

primati kape nose

primati kape na glavi

Gramatike strukture fraze

- Kontekstno neovisne gramatike
Context Free Grammars (CFG)
- $G = (T, N, S, R)$
 - T – skup terminalnih simbola
 - N – skup neterminalnih simbola
 - S – početni simbol ($S \in N$)
 - R – skup pravila/produkcija oblika $X \rightarrow \gamma$
 $X \in N, \gamma \in (N \cup T)^*$
- Gramatika G generira jezik L

Gramatike strukture fraze u obradi prirodnog jezika

- $G = (T, C, N, S, L, R)$
 - T – skup terminalnih simbola
 - C – skup preterminalnih simbola
 - N – skup neterminalnih simbola
 - S – početni simbol ($S \in N$)
 - L – leksikon, skup elemenata obila $X \rightarrow x$
 $X \in N, x \in T$
 - R – skup pravila/produkcija oblika $X \rightarrow \gamma$
 $X \in N, \gamma \in (N \cup T)^*$
- Gramatika G generira jezik L

CFG

$S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$VP \rightarrow V\ NP\ PP$

$NP \rightarrow NP\ NP$

$NP \rightarrow NP\ PP$

$NP \rightarrow N$

$NP \rightarrow \epsilon$

$PP \rightarrow P\ NP$

$N \rightarrow primati$

$N \rightarrow kape$

$N \rightarrow nose$

$N \rightarrow glavi$

$V \rightarrow primati$

$V \rightarrow kape$

$V \rightarrow nose$

$P \rightarrow na$

primati kape nose

primati kape na glavi

Probabilistic CFG (PCFG)

- $G = (T, N, S, R, P)$
 - T – skup terminalnih simbola
 - N – skup neterminalnih simbola
 - S – početni simbol ($S \in N$)
 - R – skup pravila/produkcija oblika $X \rightarrow \gamma$
 $X \in N, \gamma \in (N \cup T)^*$
 - P – probabilistička funkcija
 $P : R \rightarrow [0, 1]$
 $\forall X \in N, \sum_{X \rightarrow \gamma \in R} P(X \rightarrow \gamma) = 1$
- Gramatika G generira model jezika L
$$\sum_{\gamma \in T^*} P(\gamma) = 1$$

PCFG

SENT → NP VP	1.0	N → primati	0.5
VP → V NP	0.6	N → kape	0.2
VP → V NP PP	0.4	N → nose	0.2
NP → NP NP	0.1	N → glavi	0.1
NP → NP PP	0.2	V → primati	0.1
NP → N	0.7	V → kape	0.6
NP → ε	0.0	V → nose	0.3
PP → P NP	1.0	S → na	1.0

primati kape nose

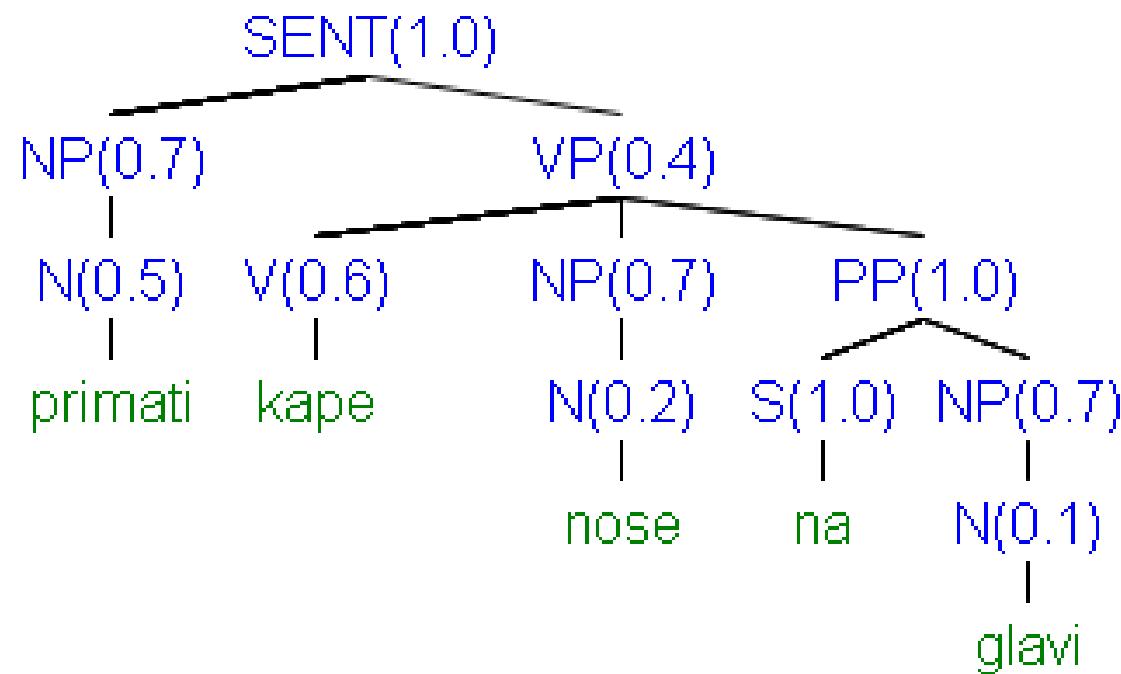
primati kape na glavi

Vjerojatnosti stabala i nizova riječi

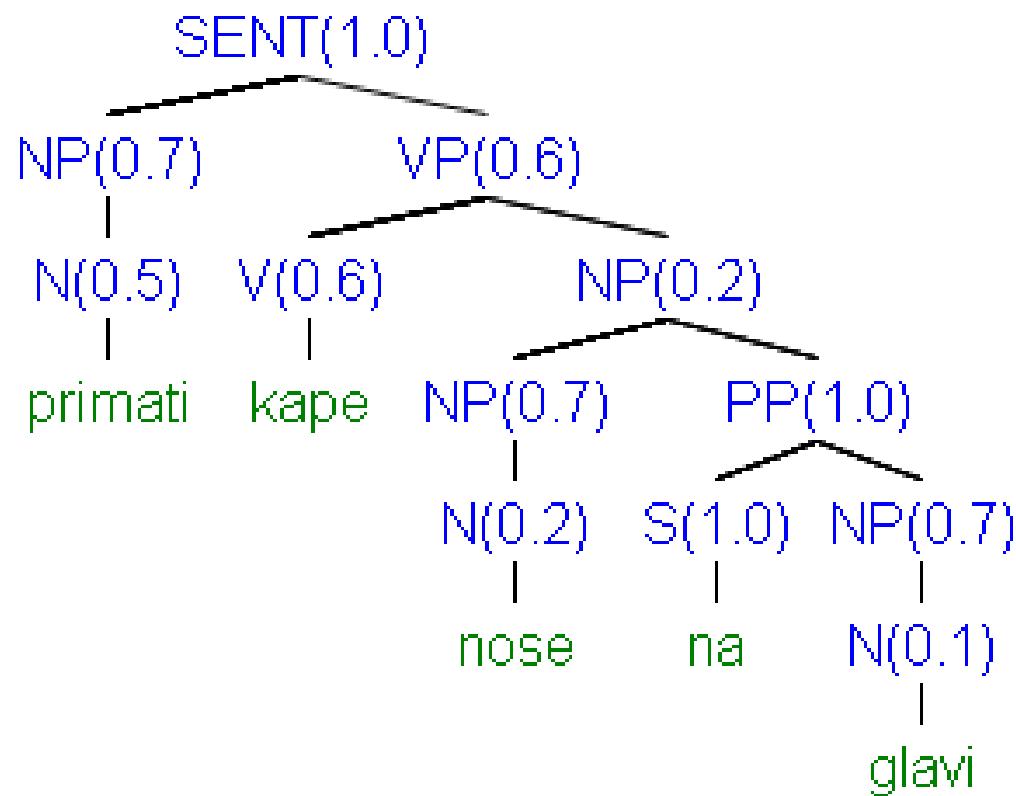
- $P(t)$ – vjerojatnost stabla t je umnožak vjerojatnosti pravila korištenih za generiranje stabla
- $P(s)$ – vjerojatnost niza riječi s je suma vjerojatnosti stabala koji generiraju s

$$\begin{aligned} P(s) &= \sum_t P(s, t) \text{ gdje je } t \text{ stablo parsiranja od } s \\ &= \sum_t P(t) \end{aligned}$$

t_1



t_2



Vjerojatnosti stabla i niza riječi

- $s = \text{primati kape nose na glavi}$
- $P(t_1) = 1.0 \times 0.7 \times 0.4 \times 0.5 \times 0.6 \times 0.7 \times 1.0 \times 0.2 \times 1.0 \times 0.7 \times 0.1$
 $= 0.0008232$
- $P(t_2) = 1.0 \times 0.7 \times 0.6 \times 0.5 \times 0.6 \times 0.2 \times 0.7 \times 1.0 \times 0.2 \times 1.0 \times 0.7 \times 0.1$
 $= 0.00024696$
- $P(s) = P(t_1) + P(t_2)$
 $= 0.0008232 + 0.00024696$
 $= 0.00107016$

Uvod u obradu prirodnog jezika

14.2. Transformacija gramatike

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Chomsky normalna forma

- Sva pravila su oblika $X \rightarrow YZ$ ili $X \rightarrow w$
 $X, Y, Z \in N, w \in T$
- Transformacija u ovu formu ne mijenja slabo generativno svojstvo CFG
 - Odnosno, prepoznae isti jezika, ali možda s različitim stablima
- Prazna i unarna pravila se rekurzivno izbacuju
- n-arna pravila se dijele uvođenjem novih neterminala ($n > 2$)

CFG

$S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$VP \rightarrow V\ NP\ PP$

$NP \rightarrow NP\ NP$

$NP \rightarrow NP\ PP$

$NP \rightarrow N$

$NP \rightarrow \epsilon$

$PP \rightarrow P\ NP$

$N \rightarrow primati$

$N \rightarrow kape$

$N \rightarrow nose$

$N \rightarrow glavi$

$V \rightarrow primati$

$V \rightarrow kape$

$V \rightarrow nose$

$P \rightarrow na$

Chomsky – eliminacija ϵ

$S \rightarrow NP VP$

$VP \rightarrow V NP$

$VP \rightarrow V NP PP$

$NP \rightarrow NP NP$

$NP \rightarrow NP PP$

$NP \rightarrow N$

~~$NP \rightarrow \epsilon$~~

$PP \rightarrow P NP$

$S \rightarrow NP VP$

$S \rightarrow VP$

$VP \rightarrow V NP$

$VP \rightarrow NP$

$VP \rightarrow V NP PP$

$VP \rightarrow V PP$

$NP \rightarrow NP NP$

$NP \rightarrow NP$

$NP \rightarrow NP PP$

$NP \rightarrow PP$

$PP \rightarrow NP PP$

$PP \rightarrow P$

$N \rightarrow primati$

$N \rightarrow kape$

$N \rightarrow nose$

$N \rightarrow glavi$

$V \rightarrow primati$

$V \rightarrow kape$

$V \rightarrow nose$

$P \rightarrow na$

Chomsky – eliminacija unarnih pravila

$S \rightarrow NP\ VP$

~~$S \rightarrow VP$~~

$VP \rightarrow V\ NP$

$VP \rightarrow V$

$VP \rightarrow V\ NP\ PP$

$VP \rightarrow V\ PP$

$NP \rightarrow NP\ NP$

$NP \rightarrow NP$

$NP \rightarrow NP\ PP$

$NP \rightarrow PP$

$NP \rightarrow N$

$PP \rightarrow P\ NP$

$PP \rightarrow P$

$S \rightarrow V\ NP$
 $S \rightarrow V$
 $S \rightarrow V\ NP\ PP$
 $S \rightarrow V\ PP$

$N \rightarrow primati$

$N \rightarrow kape$

$N \rightarrow nose$

$N \rightarrow glavi$

$V \rightarrow primati$

$V \rightarrow kape$

$V \rightarrow nose$

$P \rightarrow na$

Chomsky – eliminacija unarnih pravila

$S \rightarrow NP VP$

$VP \rightarrow V NP$

$S \rightarrow V NP$

$VP \rightarrow V$

~~$S \rightarrow V$~~

$VP \rightarrow V NP PP$

$S \rightarrow V NP PP$

$VP \rightarrow V PP$

$S \rightarrow V PP$

$NP \rightarrow NP NP$

$NP \rightarrow NP$

$NP \rightarrow NP PP$

$NP \rightarrow PP$

$NP \rightarrow N$

$PP \rightarrow P NP$

$PP \rightarrow P$

$N \rightarrow primati$

$N \rightarrow kape$

$N \rightarrow nose$

$N \rightarrow glavi$

$S \rightarrow primati$

$S \rightarrow kape$

$S \rightarrow nose$

$V \rightarrow primati$

$V \rightarrow kape$

$V \rightarrow nose$

$P \rightarrow na$

Chomsky – eliminacija unarnih pravila

$S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$S \rightarrow V\ NP$

~~$VP \rightarrow V$~~

$VP \rightarrow V\ NP\ PP$

$S \rightarrow V\ NP\ PP$

$VP \rightarrow V\ PP$

$S \rightarrow V\ PP$

$NP \rightarrow NP\ NP$

$NP \rightarrow NP$

$NP \rightarrow NP\ PP$

$NP \rightarrow PP$

$NP \rightarrow N$

$PP \rightarrow P\ NP$

$PP \rightarrow P$

$VP \rightarrow primati$

$VP \rightarrow kape$

$VP \rightarrow nose$

$N \rightarrow primati$

$N \rightarrow kape$

$N \rightarrow nose$

$N \rightarrow glavi$

$V \rightarrow primati$

$S \rightarrow primati$

$V \rightarrow kape$

$S \rightarrow kape$

$V \rightarrow nose$

$S \rightarrow nose$

$P \rightarrow na$

Chomsky – eliminacija unarnih pravila

$S \rightarrow NP VP$	$N \rightarrow primati$
$VP \rightarrow V NP$	$N \rightarrow kape$
$S \rightarrow V NP$	$N \rightarrow nose$
$VP \rightarrow V NP PP$	$N \rightarrow glavi$
$S \rightarrow V NP PP$	$V \rightarrow primati$
$VP \rightarrow V PP$	$S \rightarrow primati$
$S \rightarrow V PP$	$VP \rightarrow primati$
$NP \rightarrow NP NP$	$V \rightarrow kape$
$NP \rightarrow NP$	$S \rightarrow kape$
$NP \rightarrow NP PP$	$VP \rightarrow kape$
$NP \rightarrow PP$	$V \rightarrow nose$
$NP \rightarrow N$	$S \rightarrow nose$
$PP \rightarrow P NP$	$VP \rightarrow nose$
$PP \rightarrow P$	$P \rightarrow na$

Chomsky – eliminacija unarnih pravila

$S \rightarrow NP VP$
 $VP \rightarrow V NP$
 $S \rightarrow V NP$
 $VP \rightarrow V NP PP$
 $S \rightarrow V NP PP$
 $VP \rightarrow V PP$
 $S \rightarrow V PP$
 $NP \rightarrow NP NP$
 $NP \rightarrow NP PP$
 ~~$NP \rightarrow PP$~~
 $NP \rightarrow N$
 $PP \rightarrow P NP$
 $PP \rightarrow P$

$PP \rightarrow P NP$
 $PP \rightarrow P$

$N \rightarrow primati$
 $N \rightarrow kape$
 $N \rightarrow nose$
 $N \rightarrow glavi$
 $V \rightarrow primati$
 $S \rightarrow primati$
 $VP \rightarrow primati$
 $V \rightarrow kape$
 $S \rightarrow kape$
 $VP \rightarrow kape$
 $V \rightarrow nose$
 $S \rightarrow nose$
 $VP \rightarrow nose$
 $P \rightarrow na$

Chomsky – eliminacija unarnih pravila

$S \rightarrow NP VP$

$VP \rightarrow V NP$

$S \rightarrow V NP$

$VP \rightarrow V NP PP$

$S \rightarrow V NP PP$

$VP \rightarrow V PP$

$S \rightarrow V PP$

$NP \rightarrow NP NP$

$NP \rightarrow NP PP$

~~$NP \rightarrow N$~~

$PP \rightarrow P NP$

$NP \rightarrow P NP$

$PP \rightarrow P$

$NP \rightarrow P$

$NP \rightarrow primati$

$NP \rightarrow kape$

$NP \rightarrow nose$

$NP \rightarrow glavi$

~~$N \rightarrow primati$~~

~~$N \rightarrow kape$~~

~~$N \rightarrow nose$~~

~~$N \rightarrow glavi$~~

$V \rightarrow primati$

$S \rightarrow primati$

$VP \rightarrow primati$

$V \rightarrow kape$

$S \rightarrow kape$

$VP \rightarrow kape$

$V \rightarrow nose$

$S \rightarrow nose$

$VP \rightarrow nose$

$P \rightarrow na$

Chomsky – eliminacija unarnih pravila

$S \rightarrow NP VP$	$NP \rightarrow primati$
$VP \rightarrow V NP$	$NP \rightarrow kape$
$S \rightarrow V NP$	$NP \rightarrow nose$
$VP \rightarrow V NP PP$	$NP \rightarrow glavi$
$S \rightarrow V NP PP$	$V \rightarrow primati$
$VP \rightarrow V PP$	$S \rightarrow primati$
$S \rightarrow V PP$	$VP \rightarrow primati$
$NP \rightarrow NP NP$	$V \rightarrow kape$
$NP \rightarrow NP PP$	$S \rightarrow kape$
$PP \rightarrow P NP$	$VP \rightarrow kape$
$NP \rightarrow P NP$	$V \rightarrow nose$
$PP \rightarrow P$	$S \rightarrow nose$
$NP \rightarrow P$	$VP \rightarrow nose$
	$P \rightarrow na$
	$PP \rightarrow na$

Chomsky – eliminacija unarnih pravila

$S \rightarrow NP\ VP$	$NP \rightarrow primati$
$VP \rightarrow V\ NP$	$NP \rightarrow kape$
$S \rightarrow V\ NP$	$NP \rightarrow nose$
$VP \rightarrow V\ NP\ PP$	$NP \rightarrow glavi$
$S \rightarrow V\ NP\ PP$	$V \rightarrow primati$
$VP \rightarrow V\ PP$	$S \rightarrow primati$
$S \rightarrow V\ PP$	$VP \rightarrow primati$
$NP \rightarrow NP\ NP$	$V \rightarrow kape$
$NP \rightarrow NP\ PP$	$S \rightarrow kape$
$PP \rightarrow P\ NP$	$VP \rightarrow kape$
$NP \rightarrow P\ NP$	$V \rightarrow nose$
$NP \rightarrow P$	$S \rightarrow nose$
	$VP \rightarrow nose$
	$NP \rightarrow na$
	$P \rightarrow na$
	$PP \rightarrow na$

Chomsky – binarizacija

$S \rightarrow NP VP$
 $VP \rightarrow V NP$
 $S \rightarrow V NP$
 ~~$VP \rightarrow V NP PP$~~
 ~~$S \rightarrow V NP PP$~~
 $VP \rightarrow V PP$
 $S \rightarrow V PP$
 $NP \rightarrow NP NP$
 $NP \rightarrow NP PP$
 $PP \rightarrow P NP$
 $NP \rightarrow P NP$

$VP \rightarrow V @VP-V$
 $@VP-V \rightarrow NP PP$

$S \rightarrow V @S-V$
 $@S-V \rightarrow NP PP$

$NP \rightarrow primati$
 $NP \rightarrow kape$
 $NP \rightarrow nose$
 $NP \rightarrow glavi$
 $V \rightarrow primati$
 $S \rightarrow primati$
 $VP \rightarrow primati$
 $V \rightarrow kape$
 $S \rightarrow kape$
 $VP \rightarrow kape$
 $V \rightarrow nose$
 $S \rightarrow nose$
 $VP \rightarrow nose$
 $P \rightarrow na$
 $PP \rightarrow na$
 $NP \rightarrow na$

Chomsky normalna forma

$S \rightarrow NP\ VP$	$NP \rightarrow primati$
$VP \rightarrow V\ NP$	$NP \rightarrow kape$
$S \rightarrow V\ NP$	$NP \rightarrow nose$
$VP \rightarrow V @VP-V$	$NP \rightarrow glavi$
$@VP-V \rightarrow NP\ PP$	$V \rightarrow primati$
$S \rightarrow V @S-V$	$S \rightarrow primati$
$@S-V \rightarrow NP\ PP$	$VP \rightarrow primati$
$VP \rightarrow V\ PP$	$V \rightarrow kape$
$S \rightarrow V\ PP$	$S \rightarrow kape$
$NP \rightarrow NP\ NP$	$VP \rightarrow kape$
$NP \rightarrow NP\ PP$	$V \rightarrow nose$
$PP \rightarrow P\ NP$	$S \rightarrow nose$
$NP \rightarrow P\ NP$	$VP \rightarrow nose$
	$P \rightarrow na$
	$PP \rightarrow na$
	$NP \rightarrow na$

CFG i Chomsky normalna forma

$S \rightarrow NP VP$	$N \rightarrow primati$
$VP \rightarrow V NP$	$N \rightarrow kape$
$VP \rightarrow V NP PP$	$N \rightarrow nose$
$NP \rightarrow NP NP$	$N \rightarrow glavi$
$NP \rightarrow NP PP$	$V \rightarrow primati$
$NP \rightarrow N$	$V \rightarrow kape$
$NP \rightarrow \epsilon$	$V \rightarrow nose$
$PP \rightarrow P NP$	$P \rightarrow na$

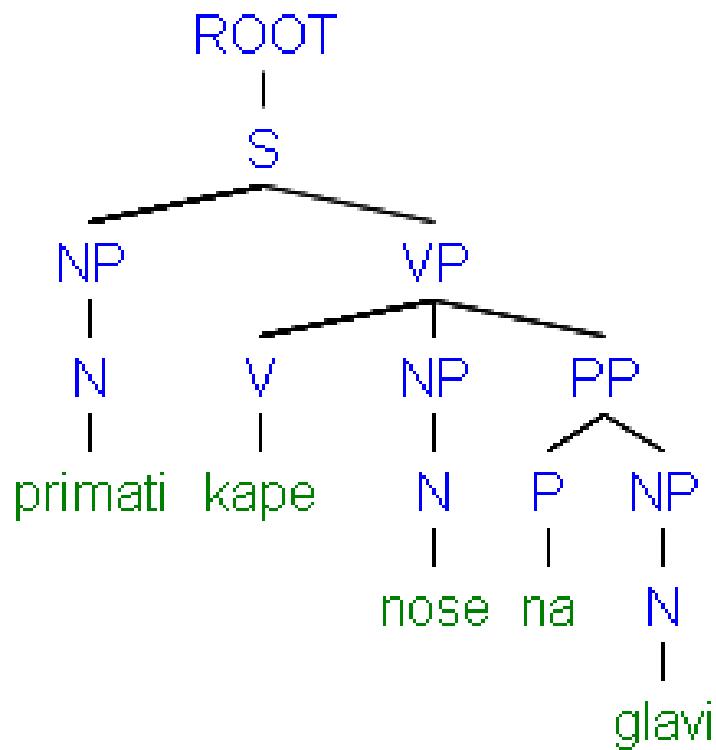
$S \rightarrow NP VP$	$NP \rightarrow primati$
$VP \rightarrow V NP$	$NP \rightarrow kape$
$S \rightarrow V NP$	$NP \rightarrow nose$
$VP \rightarrow V @VP-V$	$NP \rightarrow glavi$
$@VP-V \rightarrow NP PP$	$V \rightarrow primati$
$S \rightarrow V @S-V$	$S \rightarrow primati$
$@S-V \rightarrow NP PP$	$VP \rightarrow primati$
$VP \rightarrow V PP$	$V \rightarrow kape$
$S \rightarrow V PP$	$S \rightarrow kape$
$NP \rightarrow NP NP$	$VP \rightarrow kape$
$NP \rightarrow NP PP$	$V \rightarrow nose$
$PP \rightarrow P NP$	$S \rightarrow nose$
$NP \rightarrow P NP$	$VP \rightarrow nose$
	$P \rightarrow na$
	$PP \rightarrow na$
	$NP \rightarrow na$

Chomsky normalna forma

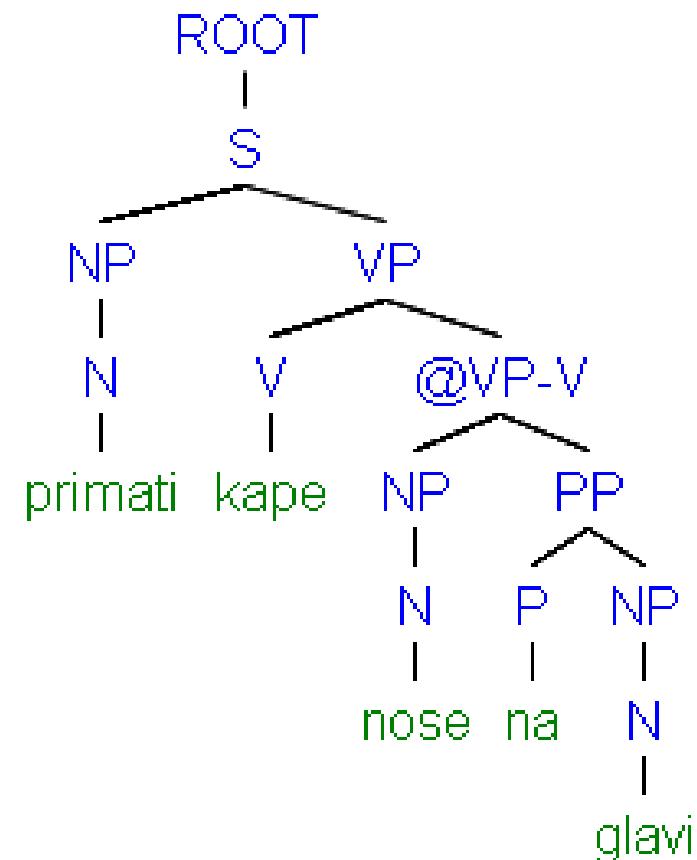
- Transformacija radi efikasnog parsiranja
- S pažljivim izborom neterminala moguće je rekonstruirati ista stabla detransformacijom
- Chomsky normalizacija nije lagana
 - rekonstrukcija n-arnih pravila je laka
 - rekonstrukcija unarnih i praznih pravila je složenija
- **Binarizacija** je krucijalna za $O(n^3)$ parsiranje CFG

Primjer: prije i poslije binarizacije

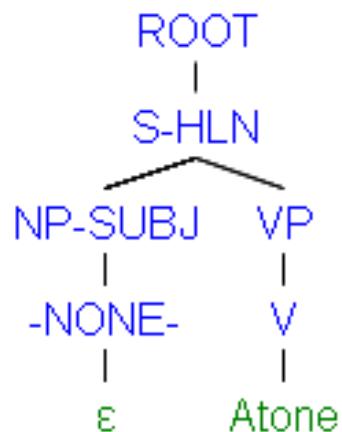
CFG



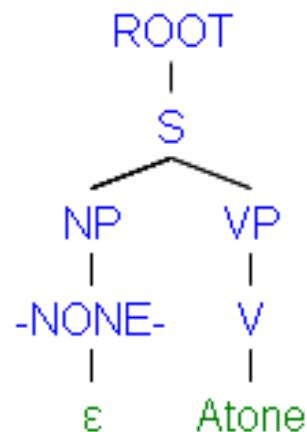
Normalizirani CFG



Banka stabala: unarna i prazna pravila



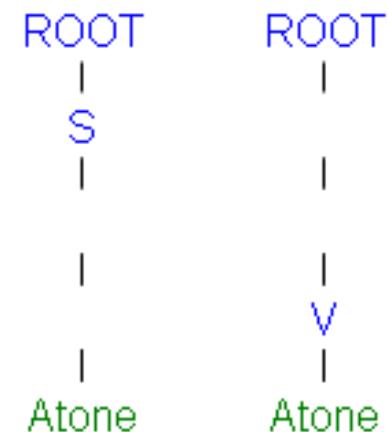
PTB Stablo



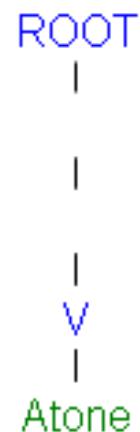
Bez
funkcijskih
oznaka



Bez
praznih
pravila



Visoko
Bez
unarnih
pravila



Nisko

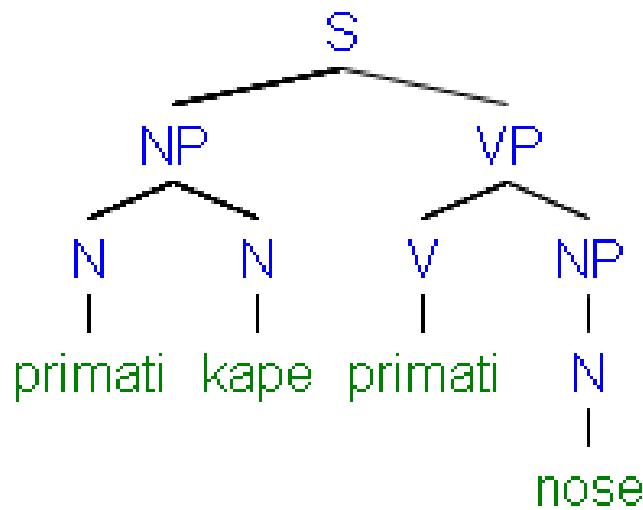
Uvod u obradu prirodnog jezika

14.3. CKY parsiranje

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Strukturno parsiranje

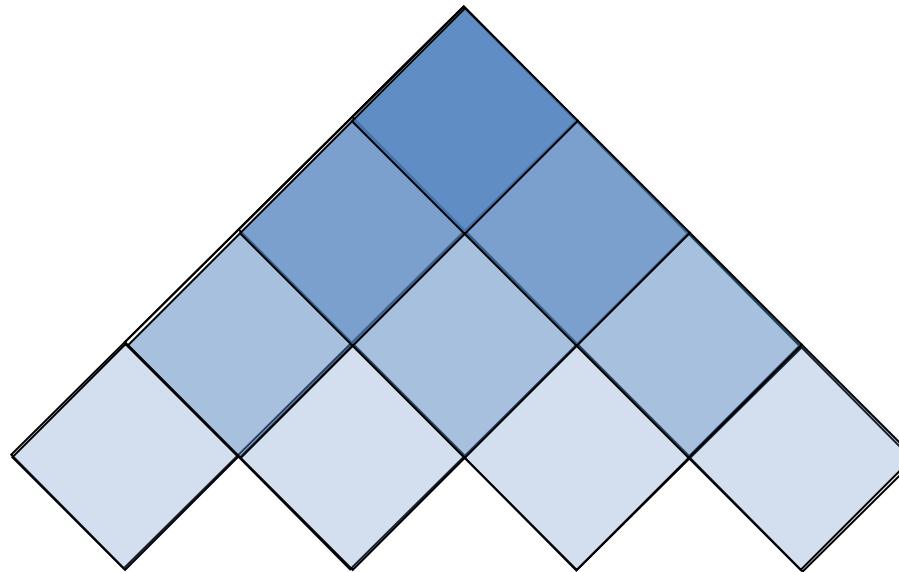


PCFG

Vjerojatnost pravila θ_i

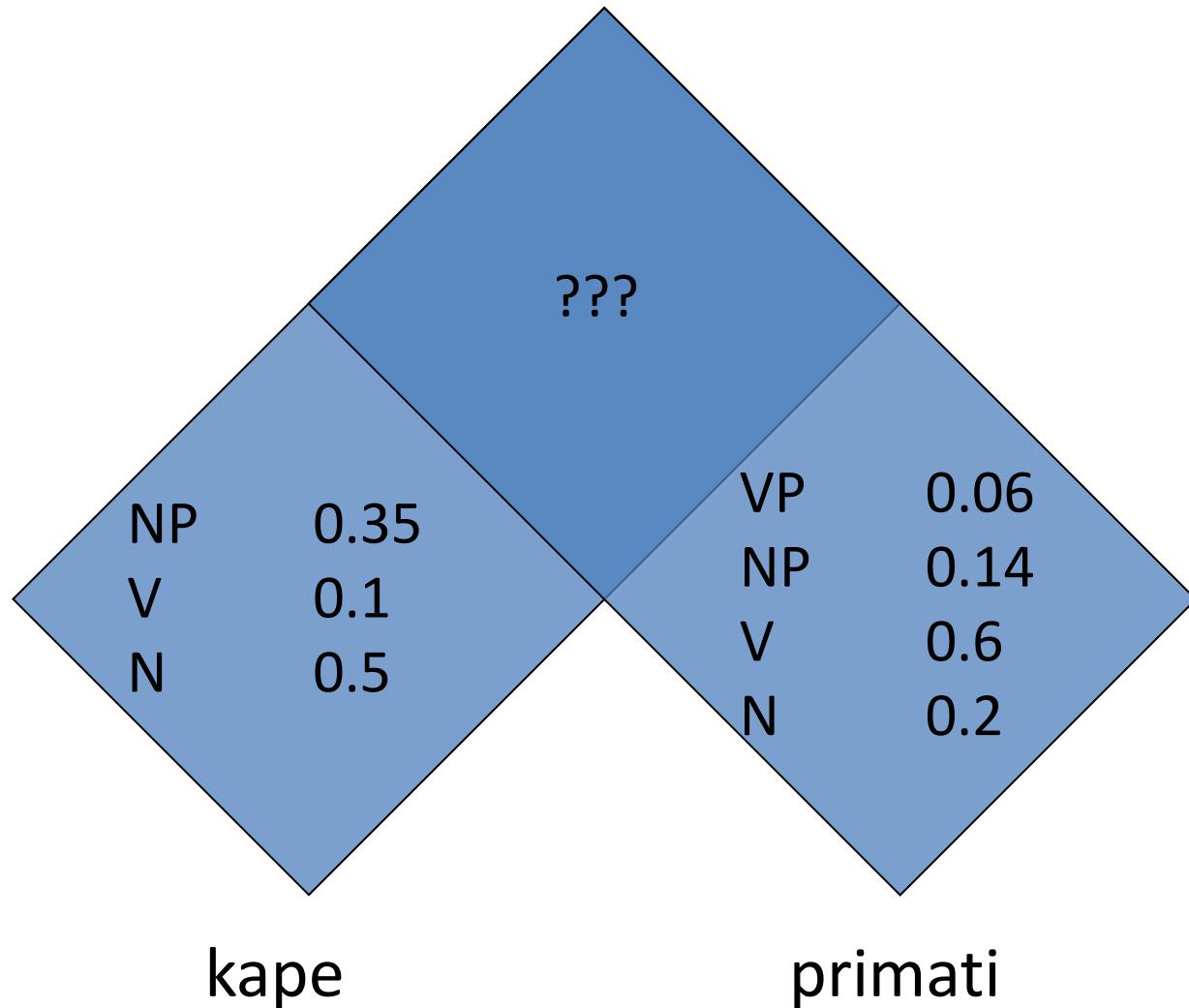
$S \rightarrow NP\ VP$	θ_1
$VP \rightarrow V\ NP$	θ_2
...	
$N \rightarrow primati$	θ_{42}
$N \rightarrow kape$	θ_{43}
$V \rightarrow primati$	θ_{44}
...	

Cocke-Kasami-Younger (CKY) parsiranje



cape primati kape nose

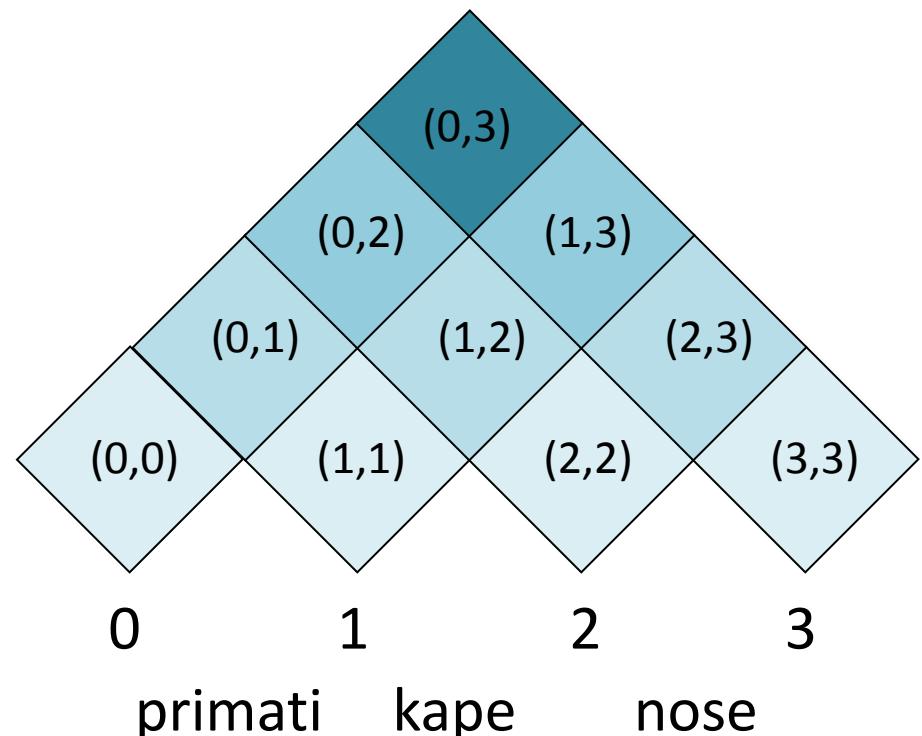
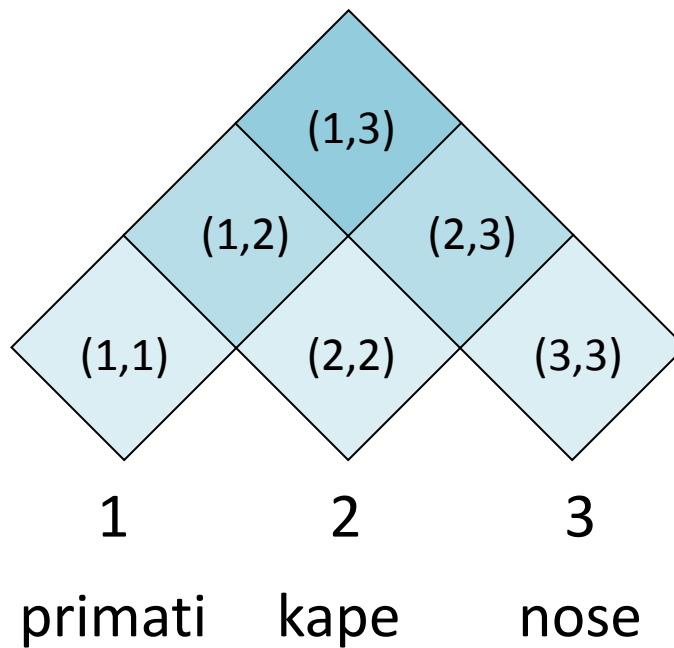
Viterbi (maksimalno bodovanje)



$S \rightarrow NP\ VP$	0.9
$S \rightarrow VP$	0.1
$VP \rightarrow V\ NP$	0.5
$VP \rightarrow V$	0.1
$VP \rightarrow V @VP_V$	0.3
$VP \rightarrow V\ PP$	0.1
$@VP_V \rightarrow NP\ PP$	1.0
$NP \rightarrow NP\ NP$	0.1
$NP \rightarrow NP\ PP$	0.2
$NP \rightarrow N$	0.7
$PP \rightarrow P\ NP$	1.0

Prošireno CKY parsiranje

- Unarna pravila se mogu uključiti u algoritam
 - malo neredno, ali ne povećava složenost algoritma
- Prazna pravila se mogu uključiti u algoritam
 - korištenje praznih ćelija
 - ne povećava složenost algoritma; slično kao unarna pravila



Prošireno CKY parsiranje

- Binarizacija je vitalna
 - bez binarizacije se ne dobiva kubično vrijeme parsiranja u odnosu na duljinu rečenice i broja neterminala u gramatici
 - Binarizacija može biti eksplisitna ili implicitna kod rada algoritma (kao Earley-ev algoritam), ali je uvijek prisutna

CKY algoritam

```
function CKY(rijeci, gramatika)
    # inicializacija
    bod = realna matrica dimenziye |rijeci|+1 x |rijeci|+1 x |neterminali|
    nazad = matrica parova dimenziye |rijeci|+1 x |rijeci|+1 x |neterminali|

    # prvi red
    for pocetak = 0 to |rijeci|-1 do
        kraj = pocetak + 1
        for A -> rijeci[pocetak] in gramatika do
            bod[pocetak][kraj][A] = P(A -> rijeci[pocetak])

    # unarna pravila za prvi red
    dodan = True
    while dodan do
        dodan = False
        for A -> B in gramatika do
            if bod[pocetak][kraj][B] > 0 then
                prob = P(A->B) * bod[pocetak][kraj][B]
                if prob > bod[pocetak][kraj][A] then
                    bod[pocetak][kraj][A] = prob
                    nazad[pocetak][kraj][A] = B
                    dodan = True
```

CKY algoritam

```
# ostali redovi
for raspon = 2 to |rijeci| do
    for pocetak = 0 to |rijeci|-raspon do
        kraj = pocetak + raspon
        for podjela = pocetak+1 to kraj-1 do
            for A -> B C in gramatika do
                prob = bod[pocetak] [podjela] [B] * bod[podjela] [kraj] [C] *
                    P(A -> B C)
                if prob > bod[pocetak] [kraj] [A] then
                    bod[pocetak] [kraj] [A] = prob
                    nazad[pocetak] [kraj] [A] = (podjela, B, C)

# unarna pravila za ostale redove
dodan = True
while dodan do
    dodan = False
    for A -> B in gramatika do
        prob = P(A -> B) * bod[pocetak] [kraj] [B]
        if prob > bod[pocetak] [kraj] [A] then
            bod[pocetak] [kraj] [A] = prob
            nazad[pocetak] [kraj] [A] = B
            dodan = True

return NapraviStablo(bod, nazad)
```

Uvod u obradu prirodnog jezika

14.4. CKY parsiranje: primjer

Branko Žitko

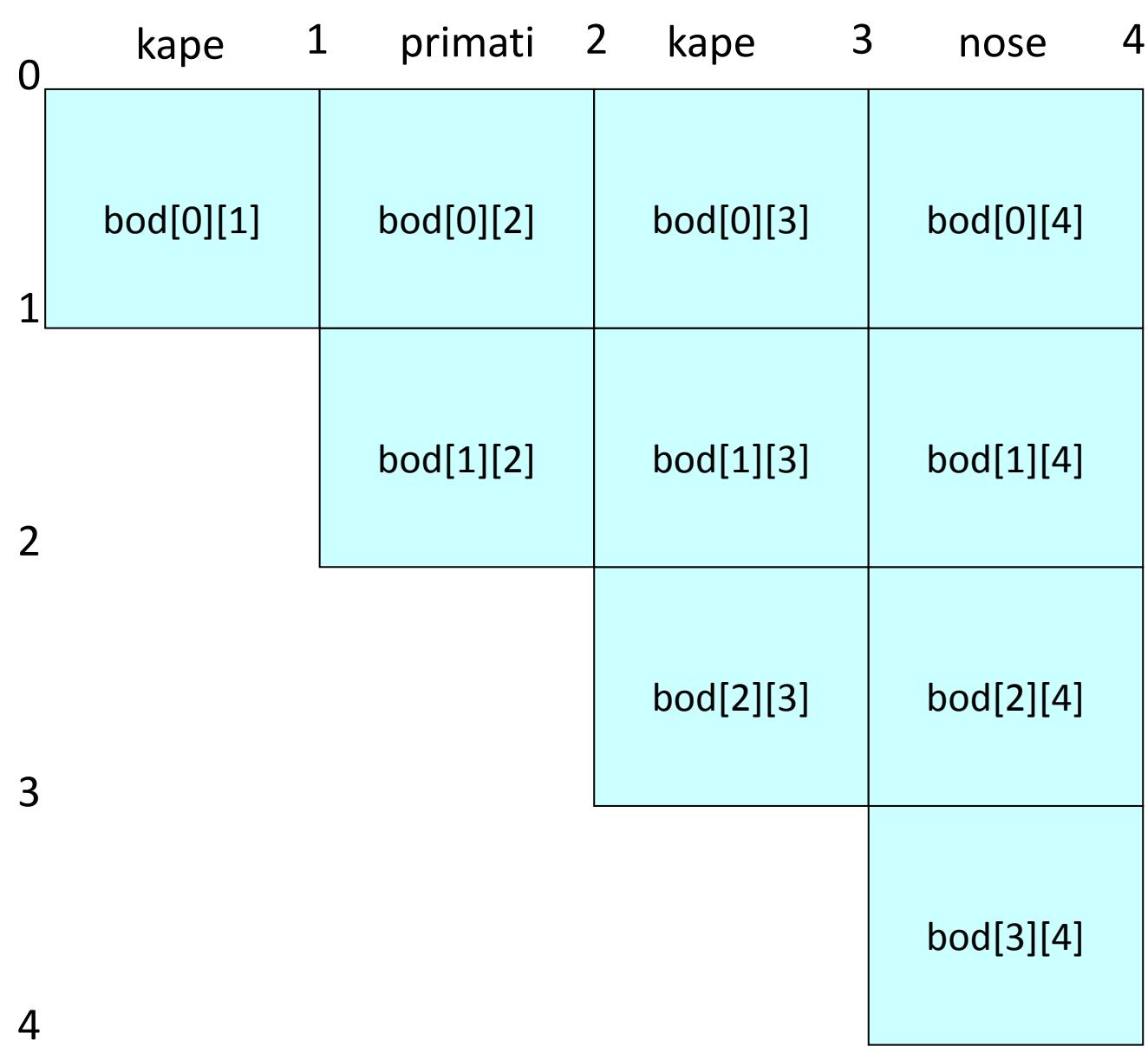
prevedeno od: Dan Jurafsky, Chris Manning

Binarna gramatika bez praznih pravila

$S \rightarrow NP VP$	0.9	$N \rightarrow primati$	0.5
$S \rightarrow VP$	0.1	$N \rightarrow kape$	0.2
$VP \rightarrow V NP$	0.5	$N \rightarrow nose$	0.2
$VP \rightarrow V$	0.1	$N \rightarrow glavi$	0.1
$VP \rightarrow V @VP_V$	0.3	$V \rightarrow primati$	0.1
$VP \rightarrow V PP$	0.1	$V \rightarrow kape$	0.6
$@VP_V \rightarrow NP PP$	1.0	$V \rightarrow nose$	0.3
$NP \rightarrow NP NP$	0.1	$P \rightarrow na$	1.0
$NP \rightarrow NP PP$	0.2		
$NP \rightarrow N$	0.7		
$PP \rightarrow P NP$	1.0		

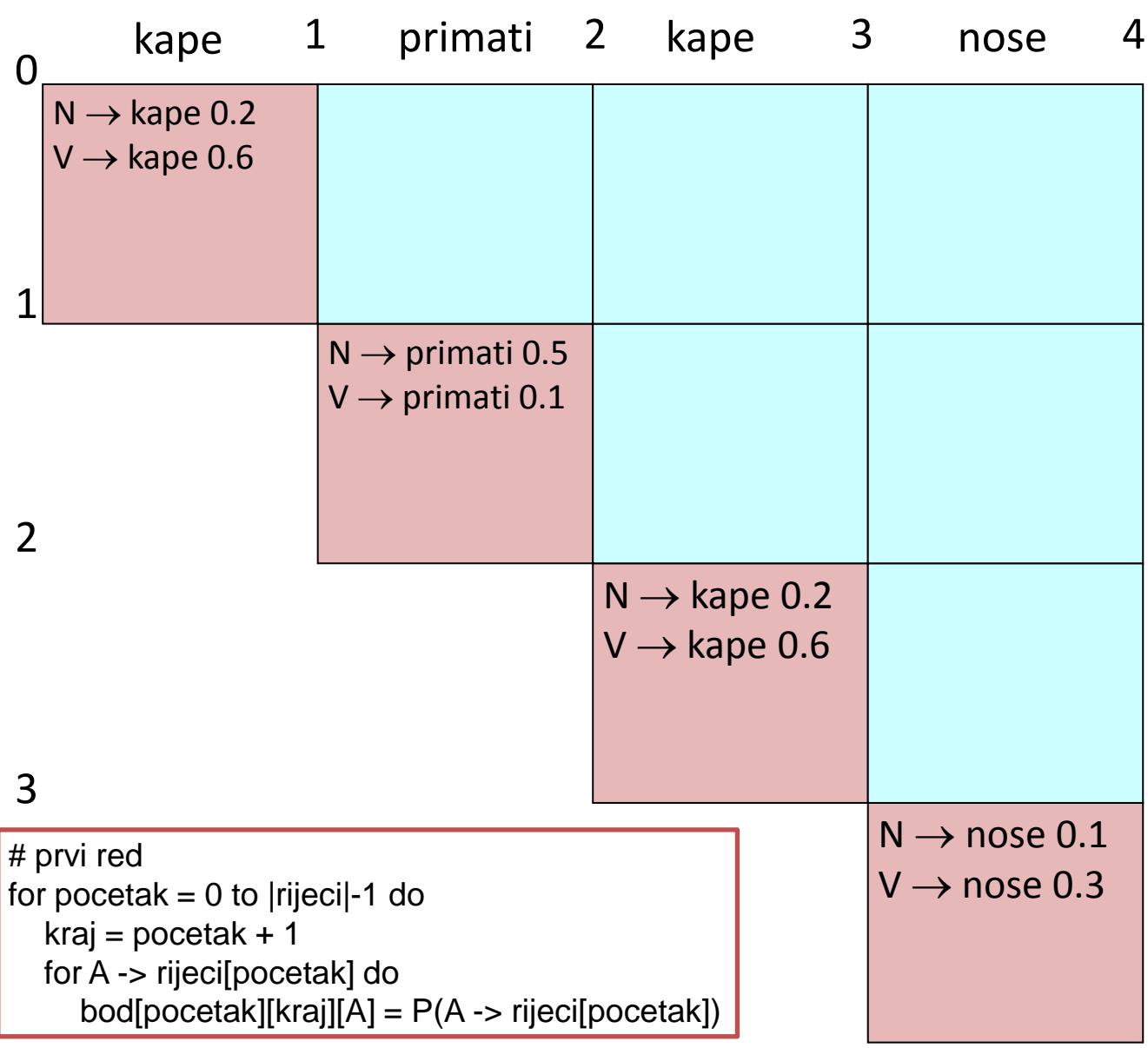
CKY

$S \rightarrow NP VP$	0.9
$S \rightarrow VP$	0.1
$VP \rightarrow V NP$	0.5
$VP \rightarrow V$	0.1
$VP \rightarrow V @VP_V$	0.3
$VP \rightarrow V PP$	0.1
$@VP_V \rightarrow NP PP$	1.0
$NP \rightarrow NP NP$	0.1
$NP \rightarrow NP PP$	0.2
$NP \rightarrow N$	0.7
$PP \rightarrow P NP$	1.0
$N \rightarrow primati$	0.5
$N \rightarrow kape$	0.2
$N \rightarrow nose$	0.2
$N \rightarrow glavi$	0.1
$V \rightarrow primati$	0.1
$V \rightarrow kape$	0.6
$V \rightarrow nose$	0.3
$P \rightarrow na$	1.0



CKY – leksička pravila

$S \rightarrow NP VP$	0.9
$S \rightarrow VP$	0.1
$VP \rightarrow V NP$	0.5
$VP \rightarrow V$	0.1
$VP \rightarrow V @VP_V$	0.3
$VP \rightarrow V PP$	0.1
$@VP_V \rightarrow NP PP$	1.0
$NP \rightarrow NP NP$	0.1
$NP \rightarrow NP PP$	0.2
$NP \rightarrow N$	0.7
$PP \rightarrow P NP$	1.0
$N \rightarrow primati$	0.5
$N \rightarrow kape$	0.2
$N \rightarrow nose$	0.2
$N \rightarrow glavi$	0.1
$V \rightarrow primati$	0.1
$V \rightarrow kape$	0.6
$V \rightarrow nose$	0.3
$P \rightarrow na$	1.0



CKY – unarna pravila

		0	1	2	3	4
		cape	primati	cape	nose	
S → NP VP	0.9					
S → VP	0.1					
VP → V NP	0.5	N → kape 0.2				
VP → V	0.1	V → kape 0.6				
VP → V @VP_V	0.3	NP → N 0.14				
VP → V PP	0.1	VP → V 0.06				
@VP_V → NP PP	1.0	S → VP 0.006				
NP → NP NP	0.1		N → primati 0.5			
NP → NP PP	0.2		V → primati 0.1			
NP → N	0.7		NP → N 0.35			
PP → P NP	1.0		VP → V 0.01			
		2	S → VP 0.001			
N → primati	0.5			N → kape 0.2		
N → kape	0.2	# unarna pravila za prvi red		V → kape 0.6		
N → nose	0.2	dodan = True		NP → N 0.14		
N → glavi	0.1	while dodan do		VP → V 0.06		
V → primati	0.1	dodan = False		S → VP 0.006		
V → kape	0.6	for A -> B in gramatika do			N → nose 0.1	
V → nose	0.3	if bod[pocetak][kraj][B] > 0 then			V → nose 0.3	
P → na	1.0	prob = P(A->B) * bod[pocetak][kraj][B]			NP → N 0.14	
		if prob > bod[pocetak][kraj][A] then			VP → V 0.03	
		bod[pocetak][kraj][A] = prob			S → VP 0.003	
		nazad[pocetak][kraj][A] = B				
		dodan = True				

CKY – binarna pravila

$S \rightarrow NP VP$	0.9	0	cape	1	primati	2	cape	3	nose	4
$S \rightarrow VP$	0.1	0								
$VP \rightarrow V NP$	0.5	0	$N \rightarrow kape$ 0.2	$NP \rightarrow NP NP$ 0.0049						
$VP \rightarrow V$	0.1	0	$V \rightarrow kape$ 0.6	$VP \rightarrow V NP$ 0.105						
$VP \rightarrow V @VP_V$	0.3	0	$NP \rightarrow N$ 0.14	$S \rightarrow NP VP$ 0.00126						
$VP \rightarrow V PP$	0.1	0	$VP \rightarrow V$ 0.06							
$@VP_V \rightarrow NP PP$	1.0	1	$S \rightarrow VP$ 0.006							
$NP \rightarrow NP NP$	0.1	1		$N \rightarrow primati$ 0.5	$NP \rightarrow NP NP$ 0.0049					
$NP \rightarrow NP PP$	0.2	1		$V \rightarrow primati$ 0.1	$VP \rightarrow V NP$ 0.007					
$NP \rightarrow N$	0.7	1		$NP \rightarrow N$ 0.35	$S \rightarrow NP VP$ 0.0189					
$PP \rightarrow P NP$	1.0	1		$VP \rightarrow V$ 0.01						
$N \rightarrow primati$	0.5	2		$S \rightarrow VP$ 0.001						
$N \rightarrow kape$	0.2	2								
$N \rightarrow nose$	0.2	2								
$N \rightarrow glavi$	0.1	2								
$V \rightarrow primati$	0.1	3								
$V \rightarrow kape$	0.6	3								
$V \rightarrow nose$	0.3	3								
$P \rightarrow na$	1.0	4								

```

prob = bod[pocetak][podjela][B] * bod[podjela][kraj][C] * P(A -> B C)
if prob > bod[pocetak][kraj][A] then
    bod[pocetak][kraj][A] = prob
    nazad[pocetak][kraj][A] = (podjela, B, C)

```

CKY – unarna pravila

$S \rightarrow NP VP$	0.9	0	cape	1	primati	2	cape	3	nose	4
$S \rightarrow VP$	0.1	0								
$VP \rightarrow V NP$	0.5	0	$N \rightarrow kape$ 0.2	$NP \rightarrow NP NP$ 0.0049						
$VP \rightarrow V$	0.1	0	$V \rightarrow kape$ 0.6	$VP \rightarrow V NP$ 0.105						
$VP \rightarrow V @VP_V$	0.3	0	$NP \rightarrow N$ 0.14	$S \rightarrow VP$ 0.0105						
$VP \rightarrow V PP$	0.1	0	$VP \rightarrow V$ 0.06							
$@VP_V \rightarrow NP PP$	1.0	1	$S \rightarrow VP$ 0.006							
$NP \rightarrow NP NP$	0.1	1		$N \rightarrow primati$ 0.5	$NP \rightarrow NP NP$ 0.0049					
$NP \rightarrow NP PP$	0.2	1		$V \rightarrow primati$ 0.1	$VP \rightarrow V NP$ 0.007					
$NP \rightarrow N$	0.7	1		$NP \rightarrow N$ 0.35	$S \rightarrow NP VP$ 0.0189					
$PP \rightarrow P NP$	1.0	1		$VP \rightarrow V$ 0.01						
$N \rightarrow primati$	0.5	2		$S \rightarrow VP$ 0.001						
$N \rightarrow kape$	0.2	2	# unarna pravila za ostale redove	dodan = True	$N \rightarrow kape$ 0.2	$NP \rightarrow NP NP$ 0.00196				
$N \rightarrow nose$	0.2	2		while dodan do	$V \rightarrow kape$ 0.6	$VP \rightarrow V NP$ 0.042				
$N \rightarrow glavi$	0.1	2		dodan = False	$NP \rightarrow N$ 0.14	$S \rightarrow VP$ 0.0042				
$V \rightarrow primati$	0.1	2		for A -> B in gramatika do	$VP \rightarrow V$ 0.06					
$V \rightarrow kape$	0.6	2		prob = P(A -> B) * bod[pocetak][kraj][B]	$S \rightarrow VP$ 0.006					
$V \rightarrow nose$	0.3	2		if prob > bod[pocetak][kraj][A] then						
$P \rightarrow na$	1.0	2		bod[pocetak][kraj][A] = prob						
				nazad[pocetak][kraj][A] = B						
				dodan = True						
						$N \rightarrow nose$ 0.1				
						$V \rightarrow nose$ 0.3				
						$NP \rightarrow N$ 0.14				
						$VP \rightarrow V$ 0.03				
						$S \rightarrow VP$ 0.003				

CKY – binarna i unarna pravila

$S \rightarrow NP VP$	0.9	0	cape	1	primati	2	cape	3	nose	4
$S \rightarrow VP$	0.1	0								
$VP \rightarrow V NP$	0.5	0	$N \rightarrow kape$ 0.2	$NP \rightarrow NP NP$ 0.0049	$NP \rightarrow NP NP$ 0.0000686	0				
$VP \rightarrow V$	0.1	0	$V \rightarrow kape$ 0.6	$VP \rightarrow V NP$ 0.105	$VP \rightarrow V NP$ 0.00147	0				
$VP \rightarrow V @VP_V$	0.3	0	$NP \rightarrow N$ 0.14	$S \rightarrow VP$ 0.0105	$S \rightarrow NP VP$ 0.000882	0				
$VP \rightarrow V PP$	0.1	0	$VP \rightarrow V$ 0.06			1				
$@VP_V \rightarrow NP PP$	1.0	0	$S \rightarrow VP$ 0.006			1				
$NP \rightarrow NP NP$	0.1	1		$N \rightarrow primati$ 0.5	$NP \rightarrow NP NP$ 0.0049	1				
$NP \rightarrow NP PP$	0.2	1		$V \rightarrow primati$ 0.1	$VP \rightarrow V NP$ 0.007	1				
$NP \rightarrow N$	0.7	1		$NP \rightarrow N$ 0.35	$S \rightarrow NP VP$ 0.0189	1				
$PP \rightarrow P NP$	1.0	1	2	$VP \rightarrow V$ 0.01		2				
$N \rightarrow primati$	0.5	2		$S \rightarrow VP$ 0.001		2				
$N \rightarrow kape$	0.2	2	# ostali redovi			2				
$N \rightarrow nose$	0.2	2	for raspon = 2 to rijeci do			2				
$N \rightarrow glavi$	0.1	2	for pocetak = 0 to rijeci -raspon do			2				
$V \rightarrow primati$	0.1	2	kraj = pocetak + raspon			2				
$V \rightarrow kape$	0.6	2	for podjela = pocetak+1 to kraj-1 do			2				
$V \rightarrow nose$	0.3	2	for A -> B C in gramatika do			2				
$P \rightarrow na$	1.0	2	prob = bod[pocetak][podjela][B] * bod[podjela][kraj][C]			2				
		2	* P(A -> B C)			2				
		2	if prob > bod[pocetak][kraj][A] then			2				
		2	bod[pocetak][kraj][A] = prob			2				
		2	nazad[pocetak][kraj][A] = (podjela, B, C)			2				

CKY – binarna i unarna pravila

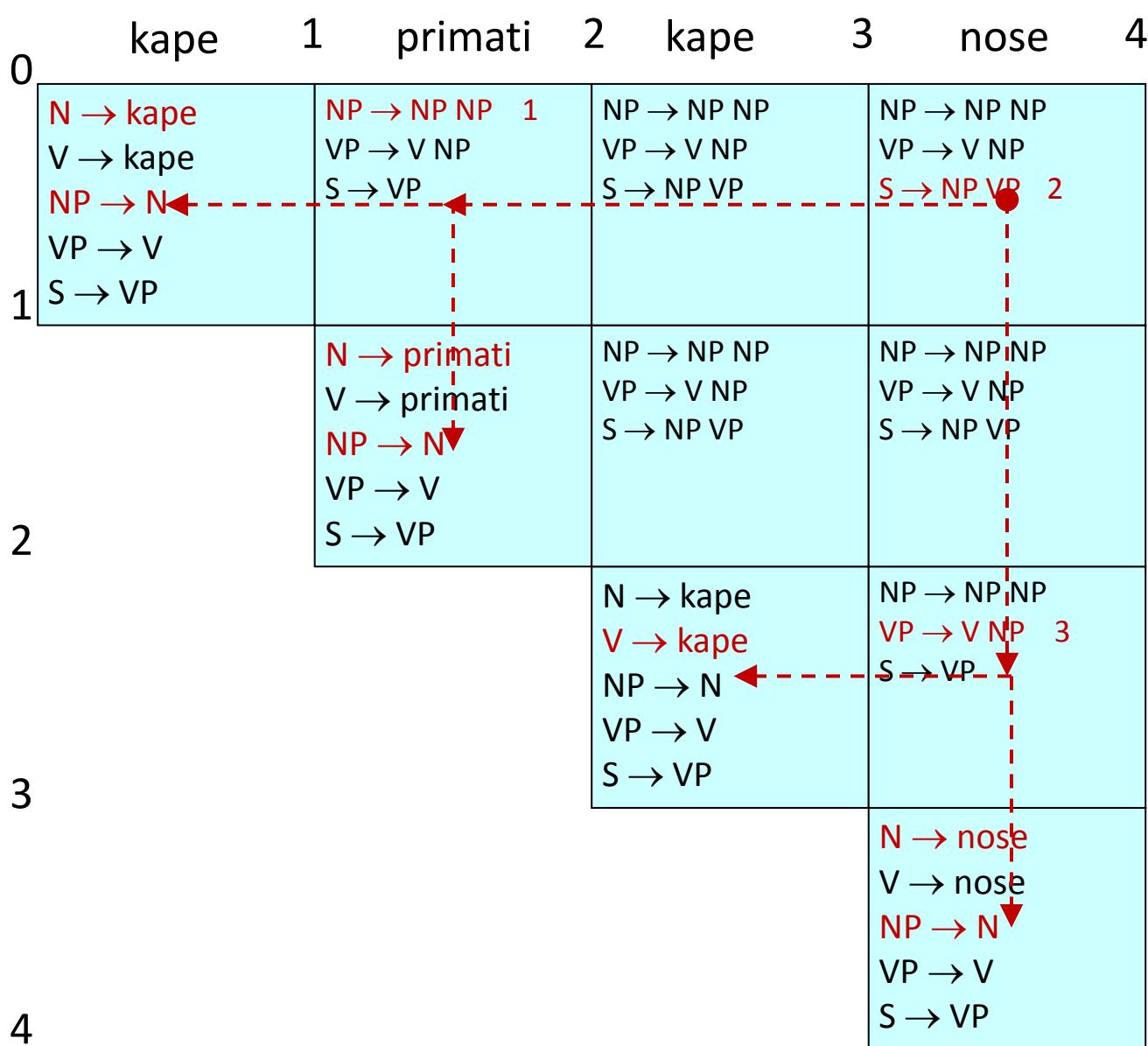
$S \rightarrow NP VP$	0.9	0	cape	1	primati	2	cape	3	nose	4
$S \rightarrow VP$	0.1	0								
$VP \rightarrow V NP$	0.5	0	$N \rightarrow kape$ 0.2	$NP \rightarrow NP NP$ 0.0049	$NP \rightarrow NP NP$ 0.0000686	$VP \rightarrow V NP$ 0.00147	$S \rightarrow NP VP$ 0.000882			
$VP \rightarrow V$	0.1	0	$V \rightarrow kape$ 0.6	$VP \rightarrow V NP$ 0.105	$VP \rightarrow V NP$ 0.00147					
$VP \rightarrow V @VP_V$	0.3	0	$NP \rightarrow N$ 0.14	$S \rightarrow VP$ 0.0105	$S \rightarrow NP VP$ 0.000882					
$VP \rightarrow V PP$	0.1	0	$VP \rightarrow V$ 0.06							
$@VP_V \rightarrow NP PP$	1.0	1	$S \rightarrow VP$ 0.006							
$NP \rightarrow NP NP$	0.1	1		$N \rightarrow primati$ 0.5	$NP \rightarrow NP NP$ 0.0049	$NP \rightarrow NP NP$ 0.0000686				
$NP \rightarrow NP PP$	0.2	1		$V \rightarrow primati$ 0.1	$VP \rightarrow V NP$ 0.007	$VP \rightarrow V NP$ 0.000098				
$NP \rightarrow N$	0.7	1		$NP \rightarrow N$ 0.35	$S \rightarrow NP VP$ 0.0189	$S \rightarrow NP VP$ 0.01323				
$PP \rightarrow P NP$	1.0	1		$VP \rightarrow V$ 0.01						
		2		$S \rightarrow VP$ 0.001						
$N \rightarrow primati$	0.5				$N \rightarrow kape$ 0.2	$NP \rightarrow NP NP$ 0.00196				
$N \rightarrow kape$	0.2				$V \rightarrow kape$ 0.6	$VP \rightarrow V NP$ 0.042				
$N \rightarrow nose$	0.2		# ostali redovi		$NP \rightarrow N$ 0.14	$S \rightarrow VP$ 0.0042				
$N \rightarrow glavi$	0.1		for raspon = 2 to rijeci do							
$V \rightarrow primati$	0.1		for pocetak = 0 to rijeci -raspon do							
$V \rightarrow kape$	0.6		kraj = pocetak + raspon							
$V \rightarrow nose$	0.3		for podjela = pocetak+1 to kraj-1 do							
$P \rightarrow na$	1.0		for A -> B C in gramatika do							
			prob = bod[pocetak][podjela][B] * bod[podjela][kraj][C]							
			* P(A -> B C)							
			if prob > bod[pocetak][kraj][A] then							
			bod[pocetak][kraj][A] = prob							
			nazad[pocetak][kraj][A] = (podjela, B, C)							

CKY – binarna i unarna pravila

$S \rightarrow NP VP$	0.9	0	cape	1	primati	2	cape	3	nose	4
$S \rightarrow VP$	0.1	0								
$VP \rightarrow V NP$	0.5	0	$N \rightarrow kape$ 0.2	$NP \rightarrow NP NP$ 0.0049	$NP \rightarrow NP NP$ 0.0000686	$NP \rightarrow NP NP$ 0.0000009604				
$VP \rightarrow V$	0.1	0	$V \rightarrow kape$ 0.6	$VP \rightarrow V NP$ 0.105	$VP \rightarrow V NP$ 0.00147	$VP \rightarrow V NP$ 0.00002058				
$VP \rightarrow V @VP_V$	0.3	0	$NP \rightarrow N$ 0.14	$S \rightarrow VP$ 0.0105	$S \rightarrow NP VP$ 0.000882	$S \rightarrow NP VP$ 0.00018522				
$VP \rightarrow V PP$	0.1	0	$VP \rightarrow V$ 0.06							
$@VP_V \rightarrow NP PP$	1.0	1	$S \rightarrow VP$ 0.006							
$NP \rightarrow NP NP$	0.1	1		$N \rightarrow primati$ 0.5	$NP \rightarrow NP NP$ 0.0049	$NP \rightarrow NP NP$ 0.0000686				
$NP \rightarrow NP PP$	0.2	1		$V \rightarrow primati$ 0.1	$VP \rightarrow V NP$ 0.007	$VP \rightarrow V NP$ 0.000098				
$NP \rightarrow N$	0.7	1		$NP \rightarrow N$ 0.35	$S \rightarrow NP VP$ 0.0189	$S \rightarrow NP VP$ 0.01323				
$PP \rightarrow P NP$	1.0	1		$VP \rightarrow V$ 0.01						
		2		$S \rightarrow VP$ 0.001						
$N \rightarrow primati$	0.5				$N \rightarrow kape$ 0.2	$NP \rightarrow NP NP$ 0.00196				
$N \rightarrow kape$	0.2				$V \rightarrow kape$ 0.6	$VP \rightarrow V NP$ 0.042				
$N \rightarrow nose$	0.2		# ostali redovi		$NP \rightarrow N$ 0.14	$S \rightarrow VP$ 0.0042				
$N \rightarrow glavi$	0.1		for raspon = 2 to rijeci do							
$V \rightarrow primati$	0.1		for pocetak = 0 to rijeci -raspon do							
$V \rightarrow kape$	0.6		kraj = pocetak + raspon							
$V \rightarrow nose$	0.3		for podjela = pocetak+1 to kraj-1 do							
$P \rightarrow na$	1.0		for A -> B C in gramatika do							
			prob = bod[pocetak][podjela][B] * bod[podjela][kraj][C]							
			* P(A -> B C)							
			if prob > bod[pocetak][kraj][A] then							
			bod[pocetak][kraj][A] = prob							
			nazad[pocetak][kraj][A] = (podjela, B, C)							

CKY – vraćanje unatrag

$S \rightarrow NP VP$	0.9
$S \rightarrow VP$	0.1
$VP \rightarrow V NP$	0.5
$VP \rightarrow V$	0.1
$VP \rightarrow V @VP_V$	0.3
$VP \rightarrow V PP$	0.1
$@VP_V \rightarrow NP PP$	1.0
$NP \rightarrow NP NP$	0.1
$NP \rightarrow NP PP$	0.2
$NP \rightarrow N$	0.7
$PP \rightarrow P NP$	1.0
$N \rightarrow primati$	0.5
$N \rightarrow kape$	0.2
$N \rightarrow nose$	0.2
$N \rightarrow glavi$	0.1
$V \rightarrow primati$	0.1
$V \rightarrow kape$	0.6
$V \rightarrow nose$	0.3
$P \rightarrow na$	1.0



Uvod u obradu prirodnog jezika

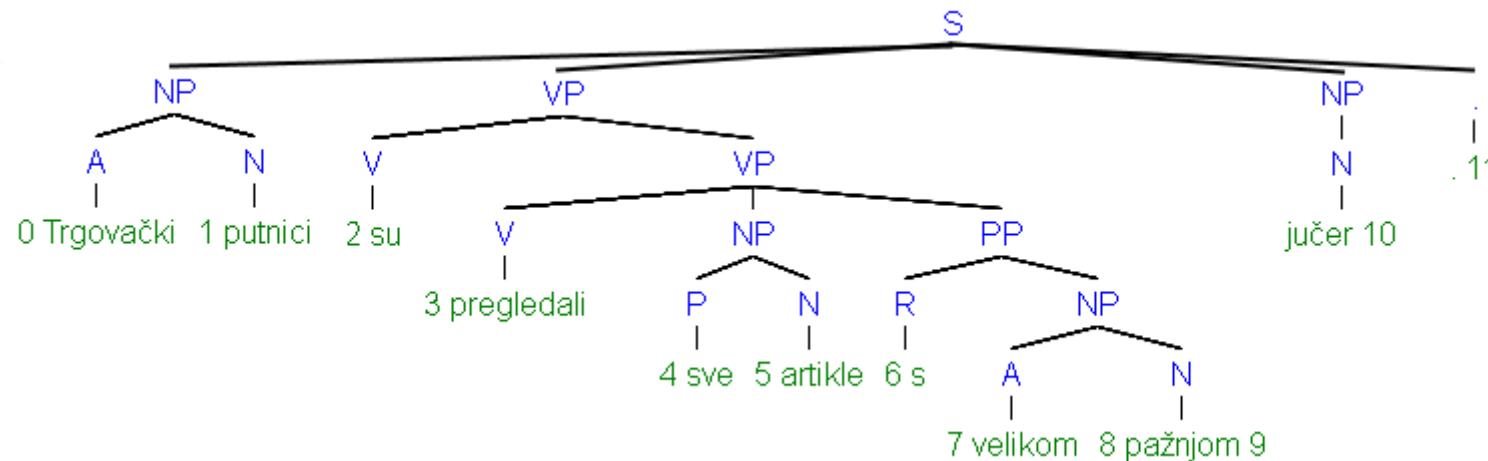
14.5. Evaluacija strukturnog parsiranja

Branko Žitko

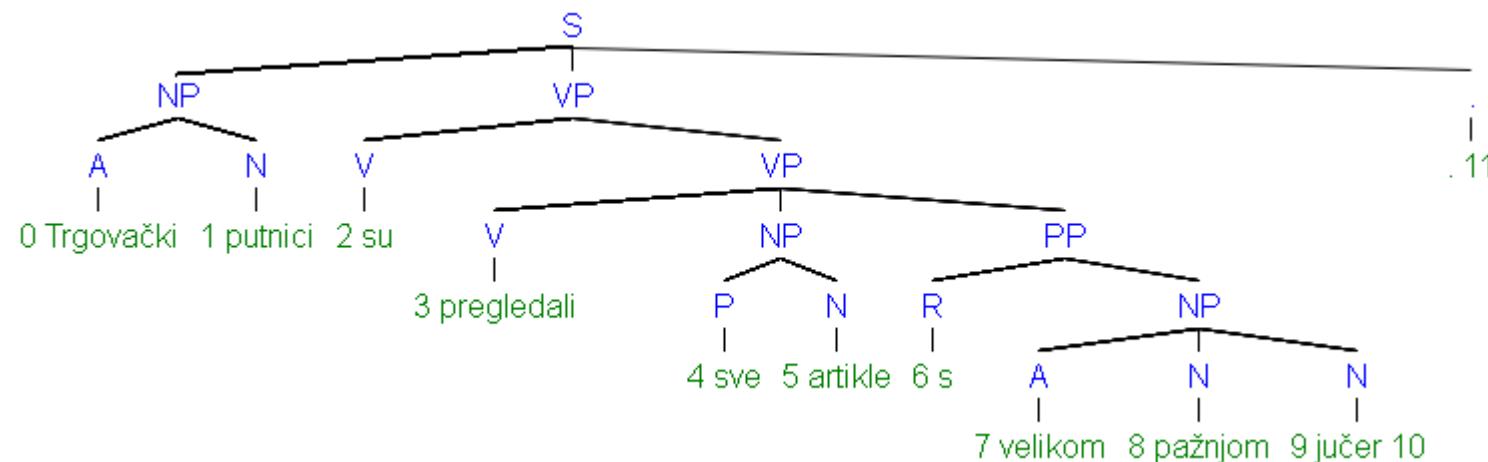
prevedeno od: Dan Jurafsky, Chris Manning

Evaluacija struktturnog parsiranja

Zlatni standard: **S(0:11) NP(0:2) VP(2:9) VP(3:9) NP(4:6) PP(6:9) NP(7:9) NP(9:10)**



Kandidat: **S(0:11) NP(0:2) VP(2:10) VP(3:10) NP(4:6) PP(6:10) NP(7:10)**



Evaluacija struktturnog parsiranja

Zlatni standard:

S(0:11) NP(0:2) VP(2:9) VP(3:9) NP(4:6) PP(6:9) NP(7:9) NP(9:10)

Kandidat:

S(0:11) NP(0:2) VP(2:10) VP(3:10) NP(4:6) PP(6:10) NP(7:10)

Preciznost oznake (PO): $3/7 = 42.9\%$

Odziv oznake (OO): $3/8 = 37.5\%$

F1 oznake: 40%

POS točnost: $11/11 = 100\%$

Koliko su dobre PCFG?

- Točnost parsiranja Penn WSJ: oko 73% F1
- Robusno
 - Obično prihvaca sve, ali s malom vjerojatnošću
- Parcijalno rješenje za višeznačnost gramatike
 - PCFG daje neke ideje vjerojatnosti parsiranja
 - ali ne toliko dobre jer su pretpostavke nezavisnosti previše jake
- Daju probabilistički model jezika
 - ali kod jednostavnih slučajeva daje lošije rezultate od trigram modela
- Izgleda da je problem PCFG-a u nedostatku leksikalizacije trigram modela