OBRADA PRIRODNOG JEZIKA (NLP)

Uvodno predavanje

2. listopada 2012.

Literatura

- Foundations of Statistical Natural Language Processing;Ch. D. Manning; H. Schütze;1999;MIT Press
- Speech and Language Processing;D. Jurafsky; J.H. Martin;2000;Prentice Hall
- Text to Speech Synthesis: New Paradigms and Advances;Sh. Narayanan; A. Alwan;2004;Prentice Hall
- The Oxford Handbook of Computational Linguistics;R. Mitkov (ed.);2005;Oxford University Press

Čemu služi NLP

- NLP je multidisciplinarno područje istraživanja i razvoja čiji je cilj produkcija aplikacija za svakodnevni život.
- Aplikacije su:
  - Pretraživanje informacija
  - Inteligentno pretraživanje Weba
  - Strojno sažimanje teksta
  - Spell checking
  - Grammar checking
  - Strojna pretvorba teksta u govor (TTS)
  - Strojna pretvorba govora u tekst (ASR)
  - Dijaloški sustavi zasnovani na razumijevanju prirodnog jezika
  - Strojno prevođenje (MT)

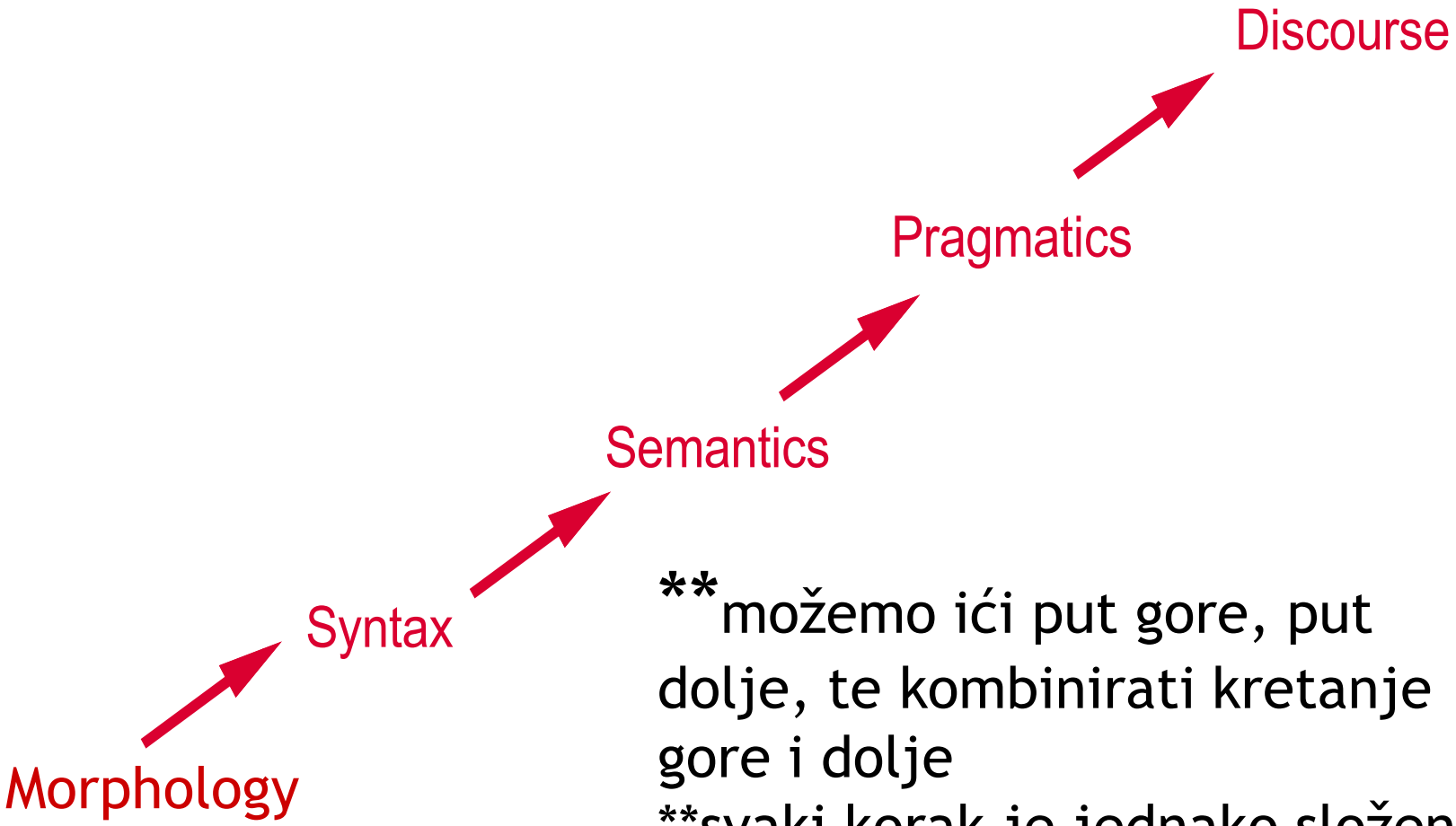
i još mnoge druge ....

www.itu.int/newsarchive/press/PP98/Documents/Statement_Gore.html , Digitalna deklaracija međuovisnosti, Al Gore, 1998

Alati NLP-a

- Teorija vjerojatnosti
- Teorija informacije
- Algoritmi
- Strukture podataka
- Umjetna inteligencija
- Process Modeling

Hijerarhija u NLP-u



**možemo ići put gore, put dolje, te kombinirati kretanje gore i dolje
**svaki korak je jednako složen!

Elementi hijerarhije

- Morfologija: bavi se načinom na koji se tvore riječi
- Sintaksa: bavi se načinom na koji niz riječi tvori rečenicu i strukturalnom ulogom riječi u rečenici
- Semantika: bavi se značenjem riječi i kako ta značenja složena u rečenicu tvore smisao rečenice
- Pragmatika: bavi se uporabom rečenica u različitim situacijama, te kao uporaba utječe na interpretaciju rečenice
- Diskurs: bavi se pitanjem kako prethodna rečenica utječe na interpretaciju sljedeće rečenice

Važnost statističkih metoda u NLP-u

- Do otprilike 1990. godine NLP se temeljio na metodama zasnovanim na pravilima (rule-based approach)
- Međutim, ovakav pristup pokazao se "pretvrdim" za opis načinâ na koji ljudi koriste prirodni jezik
- Ljudi znaju "rastezati" i "svijati" pravila da bi zadovoljili svoje komunikacijske potrebe
- Stoga su se za potrebe modeliranja prirodnoga jezika počele rabiti statističke metode (SNLP), koje pokazale dovoljno fleksibilnim u nizu aplikacija

Hijerarhija u SNLP-u

- POS (Part-of-Speech) i morfologija: riječi, njihova uloga i redoslijed u rečenici i oblici koje poprimaju
- PS (Phrase Structure) i sintaksa: pravila i ograničenja u redanju riječi i struktura fraze kao dijela rečenice
- Semantika: istraživanje značenja riječi (leksička semantika) i kako se ono ostvaruje u rečenici
- Pragmatika: istraživanje kako znanje o riječima i jezičnim konvencijama utječe na interpretaciju rečenice (implicitno uključuje diskurs)

Zašto je NLP zahtjevan

- Višeznačnost riječi
- Složenost strukture rečenice
- Riječi mogu značiti više nego njihov zbroj ili dijelovi (Fakultet elektrotehnike i računarstva)
- Suvisle riječi mogu tvoriti nesuvislu rečenicu (zeleni tangens pjeva)
- Problem određivanja značenja (ljudi vole sladoled; znači li to da svi ljudi vole sladoled?)
- Kompleksni načini interakcije između različitih stupnjeva u hijerarhiji NLP-a, odnosno SNLP-a

Što je korpus?

- Korpus je skupina označenih ili neoznačenih tekstova nad kojom primjenjuju metode NLP-a, odnosno SNLP-a, sve u cilju otkrivanja novih teorija o prirodnom jeziku, odnosno zanimljivih i korisnih načina organizacije znanja o jeziku
- Bez korpusa je danas bilo koji oblik obrade prirodnoga jezika nezamisliv
- Hrvatski korpusi:
- http://riznica.ihjj.hr, Hrvatska jezična riznica
- http://www.hnk.ffzg.hr/korpus.html, Hrvatski nacionalni korpus

Što se iz korpusa dade ekstrahirati

- Na razini riječi
  - Učestalost pojavljivanje riječi
  - Kolokacije (slijed riječi koje se učestalo pojavljuju zajedno)
  - Značenje riječi
  - N-grami (pojavljivanje dvije, tri i više riječi u fiksnom poretku)
  - Akvizicija riječi
- Na razini rečenice
  - Gramatička funkcija riječi
  - Sintaktičke strukture
- Na razini teksta (potkorpusa)
  - Značenja rečenica
- Kao sredstva na višim razinama obrade rabe se HMM (skriveni Markovljevi modeli), POS Tagging i CFG (Context Free Grammars, uključujući i probabilističke) i drugo.

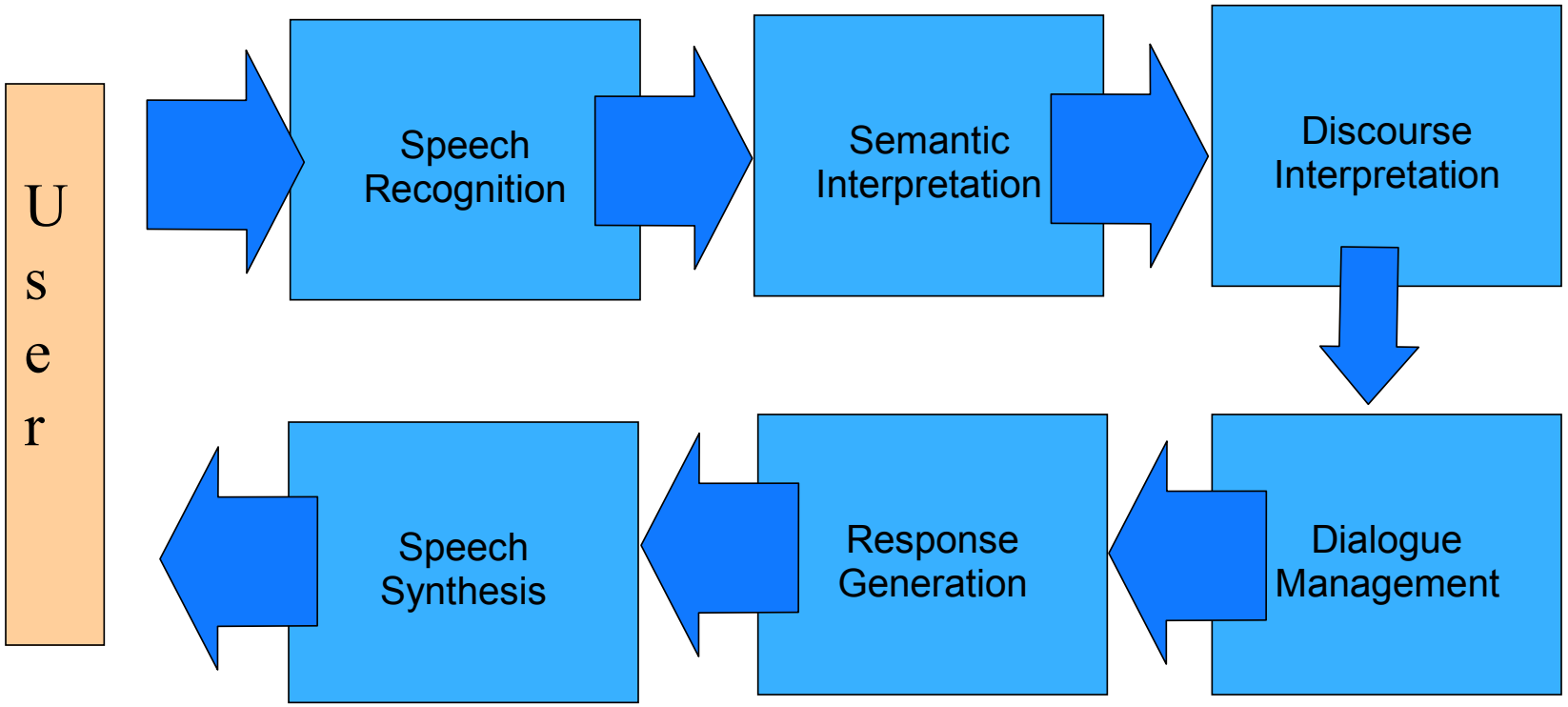
“Jednostavne” aplikacije

- Prebrajanje riječi (wc u Unix-u)
- Konkordancije
- Pravopisni provjernici (spell checkeri)
- Gramatički provjernici (grammar checkeri)
- Kontekstualni provjernici (contextual spell checking)
- Prediktori riječi na mobitelima (T9-like systems)
- .....

“Složene” aplikacije

- Inteligentni kompjutorski sustavi
- NLP sučelja prema bazama podataka
- Računalom potpomognuto poučavanje (CAI)
- Inteligentno pretraživanje Weba
- Data mining
- Strojna tvorba govora (TTS)
- Strojno prepoznavanje govora (ASR)
- Strojno prevođenje (MT, odnosno MAT, uključujuću Speech-to-Speech Translation (SST)
- Dijaloški sustavi
- ....

Dijaloški sustav – primjer složene aplikacije



Ocjenjivanje na kolegiju

- Sudjelovanje u nastavi: 10 bodova
- Međuispiti: 2x15 bodova, svaki se međuispit sastoji od 15 teorijskih pitanja od 1 bod
- Projekt: 40 bodova
- Pretprojektni izvještaj: 20 bodova

- Za postići prolaznu ocjenu potrebno je skupiti

50 bodova

Područja projektnih zadataka

- Prepoznavanje govora temeljeno na CMU Sphinxu
- Paralelno pretraživanje n-gramskih leksičkih baza
- Može i drugo, prema vlastitoj ideji

OBRADA PRIRODNOG JEZIKA (NLP)

Matematičke osnove NLP-a

9. listopada 2012.

Sadržaj

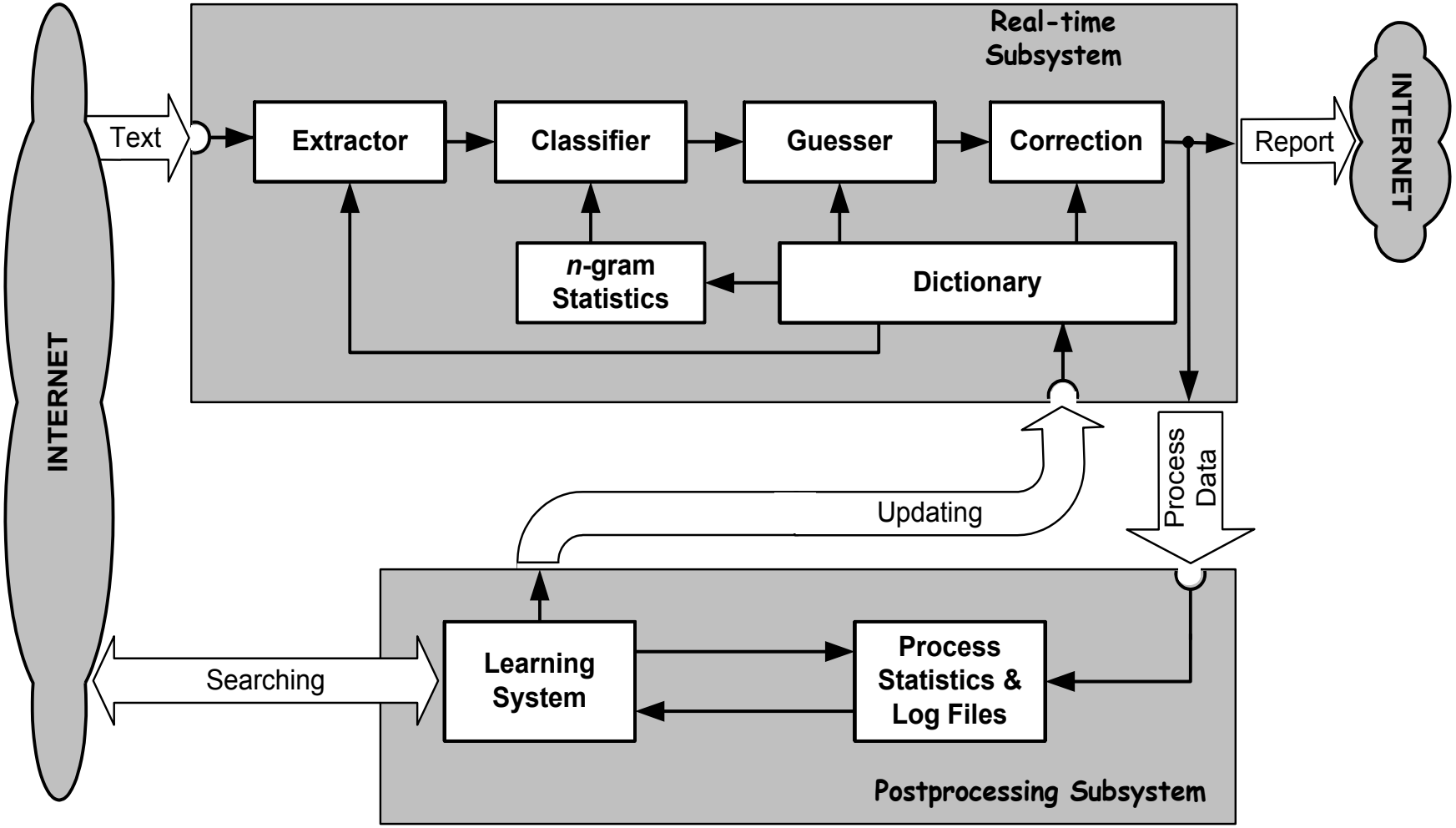
- Teorija vjerojatnosti i statistika
- Teorija informacije

- Poznavanje temeljnih postavki teorije vjerojatnosti i statistike nužno je kako za razumijevanje metoda koje se koriste u NLP-u tako i za uvođenje novih metoda i postupaka u područje
- Pored teorije vjerojatnosti matematičke osnove NLP-a obuhvaćaju i teoriju skupova, matematičku analizu (teorija funkcija, relacije), te matrični i vektorski račun
- Poznavanje temelja teorije informacija nužno je jer je prirodni jezik jedan od oblika komuniciranja s ciljem prijenosa ili pohranjivanja informacije

Zašto statistika

- Statistički pristup NLP-u jest način zaključivanja o nepoznatome
- Primjer zaključivanja: kako predvidjeti sljedeće slovo (ili sjedeću riječ) ako je poznato koja su mu slova (ili riječi) prethodili
- Da bismo to mogli napraviti potreban nam je odgovarajući jezični model
- Teorija vjerojatnosti nam pomaže da izgradimo takve modele
- Primjer: http://hacheck.tel.fer.hr/

Blok-shema Haschecka



Teorija vjerojatnosti

- Govori kolika je vjerojatnost da se nešto dogodi
- Skup Ω sadrži sve moguće ishode nekog eksperiment ili procesa (npr. sva slova nekog alfabeta ili sve riječi nekog jezika)
- Događaj A podskup je od Ω
- Nad događajem se definira vjerojatnosna funkcija
- Ona ne pretpostavlja da išta znamo o onome što je događaju A prethodilo

$$P : \Omega \rightarrow [0,1]$$

Uvjetna vjerojatnost

- Pretpostavlja izvjesno znanje o tomu što je prethodilo nekom događaju
- Neka je događaj B istinit (B se dogodio)
- Vjerojatnost da se A dogodi uz uvjet da se B dogodio piše se

$$P(A \mid B)$$

- Vrijedi

$$P(A,B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

- Uvjetne vjerojatnosti opisuju se dvodimenzionalnom tablicom

Proširenje uvjetne vjerojatnosti

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

- Može se proširiti na n događaja

$$P(A,B,C,D\ldots) = P(A)P(B|A)P(C|A,B)P(D|A,B,C..)$$

Uvjetna neovisnost događaja

- Događaji A i B su neovisni ako vrijedi

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

- Primjer neovisnih događaja: bacanje novčića
- Događaji A i B su uvjetno neovisni ako vrijedi

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

tj. ako vjerojatnost događanja A uz uvjet C ne ovisi o događaju B koji je prethodio događaju C

- Uvjetna vjerojatnost česta je pretpostavka u NLP-u (Markovljevi modeli)

Bayesov teorem

- Omogućuje da zamijenimo redoslijed ovisnosti

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- Koristi se kada je traženu uvjetnu vjerojatnost teško odrediti
- Teorem je trivijalna posljedica definicije uvjetne vjerojatnosti
- Primjenjuje se vrlo često u NLP-u

Primjer primjene Bayesova teorema

- Događa S: ukočenost vrata
- Događaj M: meningitis
- Klinički podaci: P(S|M) =0.5, P(M) = 1/50000, P(S)=1/20
- Trebam li se bojati da imam meningitis ako mi je vrat ukočen?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)}$$

$$= \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Slučajna varijabla

- Prostor događaja Ω razlikuje se za svaki problem kojega možemo razmatrati
- Kako bismo "ujednačili" prostore mogućih događaja uvodimo slučajnu varijablu X:

$$X : \Omega \rightarrow R$$

- Diskretna slučajna varijabla:

$$X : \Omega \rightarrow S$$

- (*S* je prebrojivi podskup od *R*)
- Binarna slučajna varijabla (BSV):

$$X : \Omega \rightarrow \{0,1\}$$

- http://hacheck.tel.fer.hr/
- BSV se još naziva i Bernoullijevim pokusom uz koji je povezana Bernoullijeva ili binomna razdioba

Notacija i parametri

$$p(x) = p(X = x) = p(A_x)$$

$$A_x = \{\omega \in \Omega : X(\omega) = x\}$$

$$\sum_x p(x) = 1 \qquad 0 \le p(x) \le 1$$

Očekivanje je srednja vrijednost slučajne varijable X

$$E(x) = \sum_x xp(x) = \mu$$

Varijanca

- Varijanca slučajne varijable kazuje koliko se događaju u nekom pokusu grupiraju, odnosno raspršuju

$$Var(X) = E((X - E(X))^2)$$
$$= E(X^2) - E^2(X) = \sigma^2$$

- σ je standardna devijacija
- σ^2 je varijanca

Jezični modeli

- Općenito, ako je Ω skup događaja u prirodnom jeziku (npr. alfabet ili glasovi nekoga jezika, ili pak riječi u jeziku), razdioba P je nepoznata
- Zadaća modeliranja (dobivanje modela M), je određivanje neke razdiobe P
- Razdiobu ćemo dobiti nakon uvida u statističke zakonitosti sadržane u korpusu
- http://hacheck.tel.fer.hr/

Kako se radi modeliranje

- Statistika učestalosti (prebrajanje događaja, jednodimenzionalna)
- Bayesova statistika (matrica uvjetnih vjerojatnosti, minimalno dvodimenzionalna)
- Bayesova statistika uvijek je poduprta statistikom učestalosti

Statistika učestalosti

- Prebrajamo koliko se puta neki događaj *u* dogodio u korpusu i izračunavamo relativnu frekvenciju

$$f_u = \frac{C(u)}{N}$$

- C(u) je broj pojava događaja *u* u korpusu, dok je N broj svih događaja u korpusu
- Kada N pustimo da teži prema beskonačnosti (u praksi prema nekom velikom broju) uočavamo da se relativne frekvencije stabiliziraju
- Na ovaj način dobivamo razdiobu P (statistički model M) o događajima u korpusu

Pristupi modeliranju

- Postoje dva osnovna pristupa jezičnom modeliranju u slučaju modela temeljenog na statistici učestalosti:

  Parametarsko modeliranje

  Neparametarsko modeliranje (slobodne distribucije)

Parametarsko modeliranje

- Polazi od pretpostavke da se jezični fenomen koji istražujemo i želimo modelirati ravna po nekoj od poznatih razdioba, npr. prema binomnoj (Bernoullijevoj) ili normalnoj (Gaussovoj) razdiobi
- Kada nam je razdioba poznata (ili pretpostavljiva) iz jezičnih podataka (korpusa) trebamo dobiti samo parametre distribucije (u pravilu μ i σ)
- Parametarsko modeliranje u pravilu zahtijeva manji opseg podataka (manji korpus)

Neparametarsko modeliranje

- Rabi se kada radimo jezični model za kojega unaprijed ne znamo kakvu razdiobu u tome modelu možemo očekivati
- Razdioba P utvrđuje se prebrajanjem svih događaja u velikom korpusu jezičnih događaja
- Po dobivanju razdiobe P možemo pristupiti izračunu njezinih parametara
- Kod neparametarskog modeliranja, jer nemamo nikakvo prethodno znanje ili pretpostavku o jezičnom modelu, treba puno više empirijskih podataka nego u slučaju parametarskog modeliranja

Osnovne postavke TI-a

- Teoriju informacije utemeljio je Shannon prije sedamdesetak godina
- Osnovne zadaće:

1. Maksimizirati količinu informacije koja se može prenijeti putem nesavršenog komunikacijskog kanala
2. Sažimanje (kompresija) podataka bez gubitka informacije
3. Određivanje moguće brzine prijenosa podataka po nesavršenom kanalu

Entropija

- Nužno je znati razdiobu slučajne varijable X, tj. p(x) za svaki x iz X. Tada se entropija izračunava ovako:

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

- Entropijom se mjeri količina informacije (u bitima ako je $\log_2$) potrebna za opis slučajne varijable X, odnosno njoj korespondentnog skupa slučajnih događaja Ω

- Praktički, entropija govori koliku najmanju količinu informacije u prosjeku treba prenijeti da bi nekomu priopćili da se dogodio slučajni događaj x iz X

Združena entropija

- U slučaju kada imamo dvije slučajne varijable X i Y opisane razdiobom P(X,Y) združena entropija je prosječna količina informacije potrebna da bi se nekomu priopćilo da su se dogodili i x i y iz (X,Y)

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(X,Y)$$

Uvjetna entropija

- Ukoliko prijemna strana zna da se x iz X dogodio, tada uvjetna entropija govori koliko još u prosjeku dodatne informacije moramo prenijeti kako bi se nekomu priopćilo da se dogodio i y iz Y

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X=x)$$

$$= -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log p(y|x)$$

$$= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y|x) = -E(\log p(Y|X))$$

Ulančavanje združene i uvjetne entropije

$$H(X,Y) = H(X) + H(Y \mid X)$$

$$H(Y \mid X) \leq H(Y)$$

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + \ldots + H(X_n \mid X_1, \ldots X_{n-1})$$

Srednji uzajamni sadržaj informacije

- Srednji uzajamni sadržaj informacije označava se s I(X,Y) i definira se ovako:

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$
$$H(X) - H(X|Y) = H(Y) - H(Y|X) = I(X,Y)$$

- Može se interpretirati kao smanjenje neodređenosti jedne slučajne varijable zbog poznavanja druge
- Druga interpretacija je da je I(X,Y) količina informacije koju jedna slučajna varijabla posjeduje o drugoj

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Svojstva uzajamnog sadržaja informacije

- $I(X,Y) \leq H(X)$
- $I(X,Y) \leq H(Y)$
- $I(X,Y) = 0$ onda i samo onda ako su X i Y nezavisne varijable; tada je $H(X|Y)=H(X)$, odnosno $H(Y|X)=H(Y)$
- $H(X)=H(X)-H(X|X)=I(X,X)$, što znači da je entropija "samoinformacija"

Entropija u NLP-u

- Entropija je mjera neodređenosti, što znači da se s porastom našega znanja o fenomenu kojega istražujemo (jeziku) smanjuje entropija sustava
- Kako s primjerenim jezičnim modeliranjem povećavamo naše znanje o strukturi prirodnoga jezika tako pada entropija u jeziku
- Zbog toga entropiju možemo koristiti kao mjeru uspješnosti našega modeliranja prirodnoga jezika

OBRADA PRIRODNOG JEZIKA (NLP)

Lingvističke osnove NLP-a

16. listopada 2012.

Vrste riječi

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

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

Morfologija i POS

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

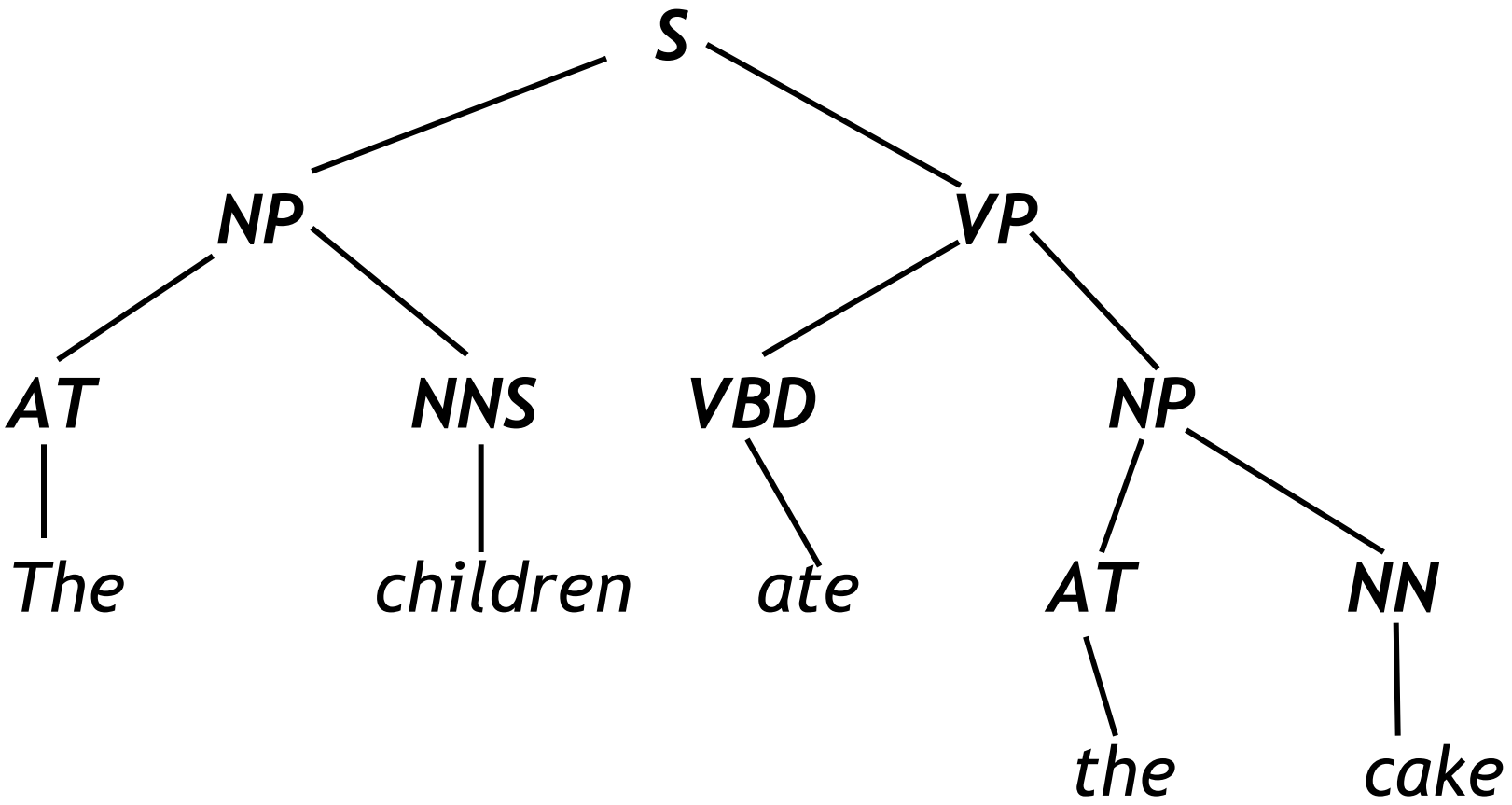
Označavanje POS-ova

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

Označavanje u hrvatskome

- Radi složenosti hrvatske morfologije označavanje u hrvatskome izuzetno je složeno
- Hrvatski morfološki leksikon http://www.hnk.ffzg.hr/hml/ koristi MSD sustav označavanja http://nl.ijs.si/ME/V2/msd/html/node15.html#SECTION03810000000000000000 koji je razvijen za neke istočnoeuropske jezike + engleski http://nl.ijs.si/ME/V2/msd/html/
- Zato ćemo se zadržati na engleskim primjerima i njihovom konvencionalnom označavanja

Rečenično stablo (parsing)



CFG (Context Free Grammars)

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

Primjer jednostavne CF gramatike

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

Sustavi označavanja

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

Stohastičke gramatike

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

Svojstva stohastičkih gramatika

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

Primjer stohastičke CF gramatike

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

Prednosti stohastičkih gramatika

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

Semantika

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

Leksička semantika

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

WordNet

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

Semantičko parsiranje rečenice

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

Pragmatika

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

I nema *sestre ni brata*

I nema *oca ni majke*

I nema *drage ni druga*

(Tin Ujević: „Svakidašnja jadikovka”)

Kreiranje hrvatske digramske baze

- Izvedena je iz korpusa od od 533,057.697 pojavnica (srpanj 2007. - rujan 2011.);
- Zahtijevala je 61 obradu ukupnoga trajanja 5:57:29;

```
BAZA-DIGRAMA+          15,927.518          378,532.290
BAZA-DIGRAMA-0          1,885.611           12,212.693
BAZA-DIGRAMA1          19,165.975           19,165.975
BAZA-DIGRAMA2           8,419.686           16,839.372
=============================================================
BAZA-DIGRAMA           45,398.790          426,750.330
```

- PRVI-BD (prvi članovi BD-a) s 1,411.638 riječi
- DRUGI-BD (drugi članovi BD-a) s 1,298.968 riječi

OBRADA PRIRODNOG JEZIKA (NLP)

O korpusima u NLP-u

23. listopada 2012.

Što su korpusi

- Velike baze podataka čiji je sadržaj tekst ili govor, ili paralelno tekst i govor
- Najznačajnije vrste tekstualnih korpusa: “plain text” korpusi, označeni (tagged) korpusi, “domain specific” korpusi, paralelni bilingvalni korpusi...
- Korpus mora biti reprezentativan da bi se na njemu mogle primijeniti statističke tehnike u cilju dobivanja pouzdanih razdioba, odnosno statističkih modela jezika

Problemi formatiranja

- Čišćenje: iz korpusa treba odstraniti primjerice HTML oznake, slike, tablice i slične sadržaje
- Veliko/malo slovo: treba li zadržati “case sensitive” pisanje? Primjerice, je li AKO identično s “ako”? Nadalje, kako razlučiti “Ivana Vlak” (ime i prezime) i “vlak za Ivana”
- Tokenizacija: što su graničnici koji određuju pojavnicu (token)
- Rečenica: što ograničava rečenicu (kratice tipa dr., dipl. ing. i slične mogu unositi zabunu)

Problemi formatiranja (2)

- Skraćenice: treba li “dr.” proširiti u “doktor”?
- Lematizacija: treba li od pojavnica ostaviti samo lemu (korijen) i odbaciti nastavke? Važno za pretraživanje korpusa (http://www.google.hr/), ali vrlo problematično u hrvatskome
- Pitanje naslova: treba li zadržati informaciju o fontu, jer naslovi i podnaslovi se obično pišu drugom fontom od ostatka teksta, tako da je font ponekad važan pri pretraživanju
- Ravnanje (alignment): izuzetno važno kod bilingvalnih korpusa

Korištenje korpusa

- U svakom dobrom korpusa velika je količina informacije o odnosu između riječi
- SNLP teži da se iz korpusa nauče ti odnosi na razini leksičkih i strukturnih preferencija koje postoje na razini pojavnica u svakom jeziku
- Kolokacije (n-grami riječi koje se učestalo pojavljuju zajedno) mogu poslu kao dobar primjer
- Primjeri: bit će uzeti iz engleskoga, jer hrvatski, radi morfološkoga bogatstva, prethodno traži lematizaciju

Kolokacije

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Osim digrama *New York*, svi učestali digrami tzv. "funkcijske riječi"

Filtriranje kolokacija

Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

POS tagging je preduvjet za uspješno filtriranje kolokacija

Rezultat filtriranja

$C(w^1\ w^2)$	w^1	w^2	tag pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

Pregled najučestalijih digramskih kolokacija nakon primjene sintaksnog filtriranja

Osnovni oblici korištenja korpusa

- Prebrajanje riječi:

1. Učestalost riječi
2. Odnos pojavnica (tokens) i različnica (types) u korpusu
3. Razvrstavanje riječi u klase (općejezični fond, imenski fond, brojevi, kratice itd.)
4. Izračun relativne učestalosti riječi po klasama

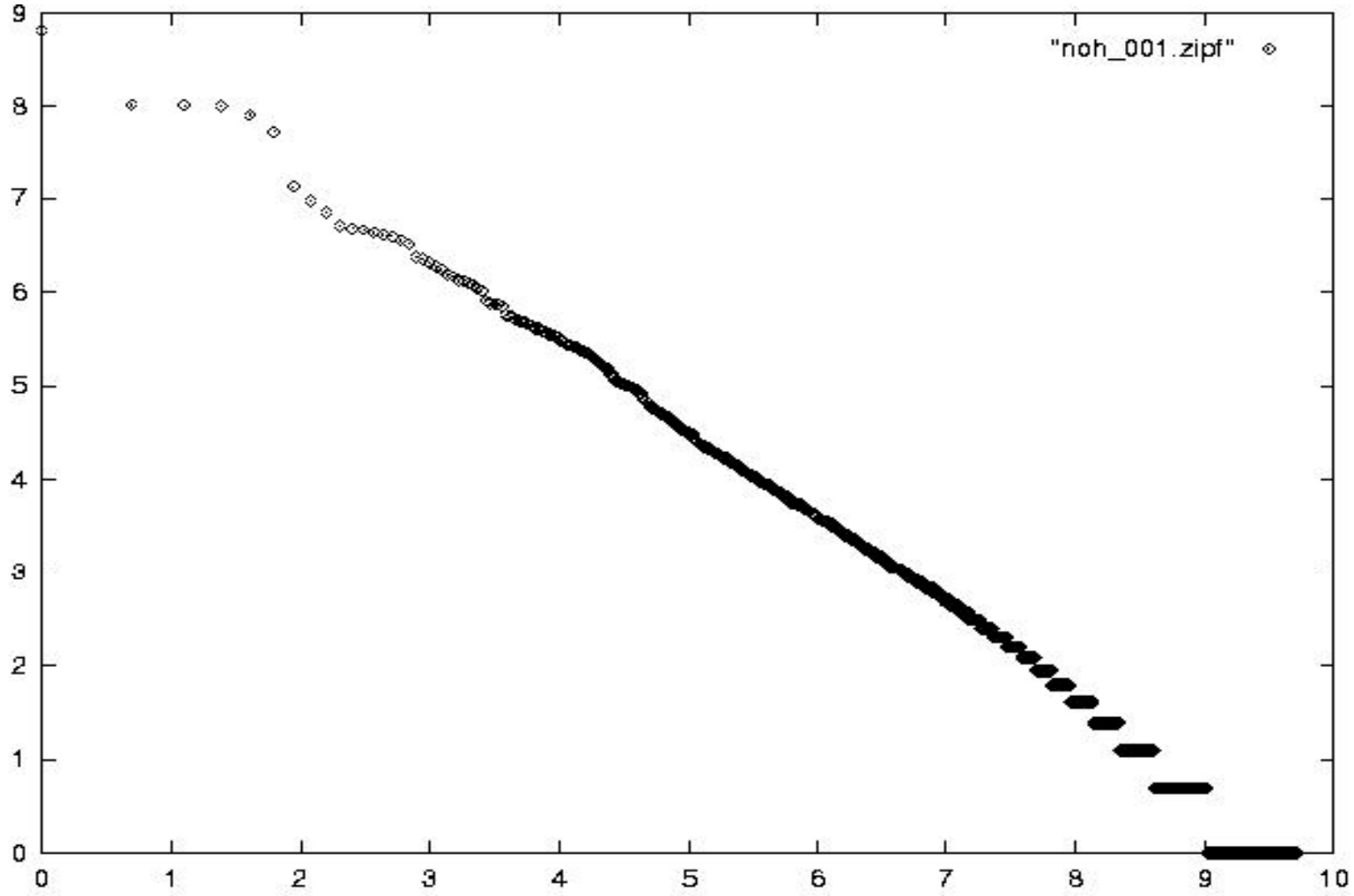
.....

Primjer: http://riznica.ihjj.hr/

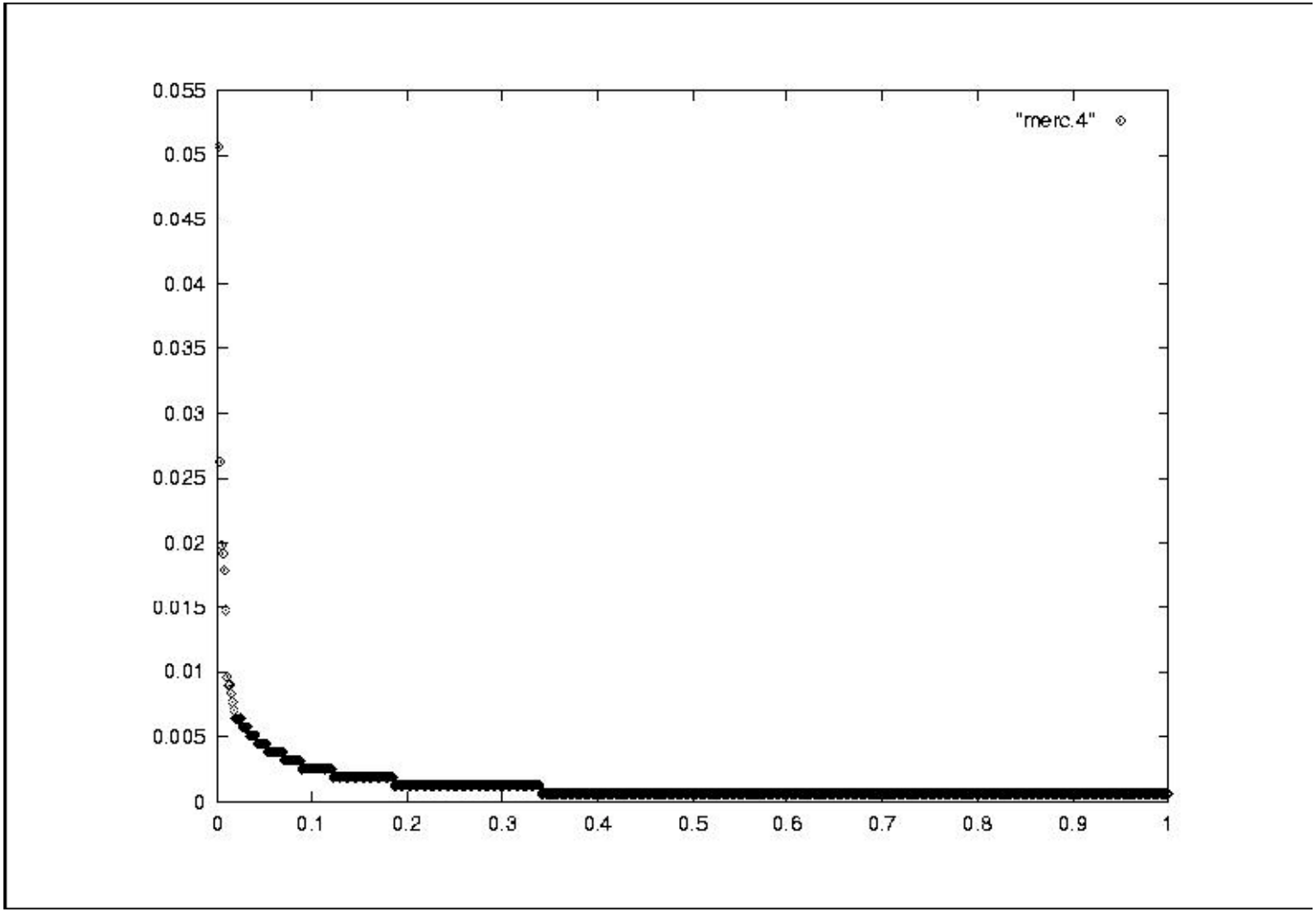
Zipfov zakon

- Zipfov zakon tvrdi da je: $f \propto 1/r$
- Zipfov zakon tvrdi da je učestalost riječi (f) obrnuto proporcionalna njezinom rankingu (r); ranking je redno mjesto pojavljivanja riječi na listi svih različnica u korpusu
- Mandelbrot je matematički doradio Zipfov zakon, tako da je on poznat i pod nazivom Zipf-Mandelbrotov zakon (**Zipf-Mandelbrot law)**
- Posljedica Zipfova zakona: za većinu riječi naše znanje o njima bit će vrlo skromno; 50-60% različnica i u opsežnim korpusima su tzv. hapax legomene

Zipfov zakon u log-log prikazu



Zipfov zakon za visoke frekvencije



Heapsov zakon - *Wikipedia*

- In linguistics, **Heaps' law** is an empirical law which describes the portion of a vocabulary which is represented by an instance document (or set of instance documents) consisting of words chosen from the vocabulary. This can be formulated as
  - $$q_V = at^c$$
- Where q_V is the subset of the vocabulary V represented by the instance text of size t [tokens]. **a** and **c** are free parameters determined empirically.
- With English text corpora, typically a is between 10 and 100, and **c** is between 0.4 and 0.6.

Heapsov zakon -komentar

- Heapsov zakon nije empirički zakon kako tvrdi Wikipedija
- Heapsov zakon je matematička posljedica Zipfova zakona
- Dokaz: A. Kornai: “How many words are there?”, http://www.metacarta.com/Collateral/Documents/English-US/how-many-words-are-there-Kornai.pdf
- Heapsov zakon je integralni oblik Zipfova zakona
- Prednost: robusniji je od Zipfova zakona

Primjena na Haschecka

$$V_i(t) = \alpha_i \cdot (t - K_i)^{\beta_i}$$

$$V_{WT}(t) + V_{NT}(t) = V(t)$$

Područje A: početni desetomilijunski korpus;
Područje B: ukupni korpus u rasponu od 50 do 100 milijuna pojavnica;
Područje C: ukupni korpus iznad 100 milijuna pojavnica.

Modeliranje Heapsova zakona

- Heapsov zakon stalno podliježe promjeni, odnosno izračunu optimalnih parametara za željeni opseg korpusa
- Kod modeliranja je moguće olabaviti drugi uvjet s prethodnog *slidea*, tj. zadovoljiti se s

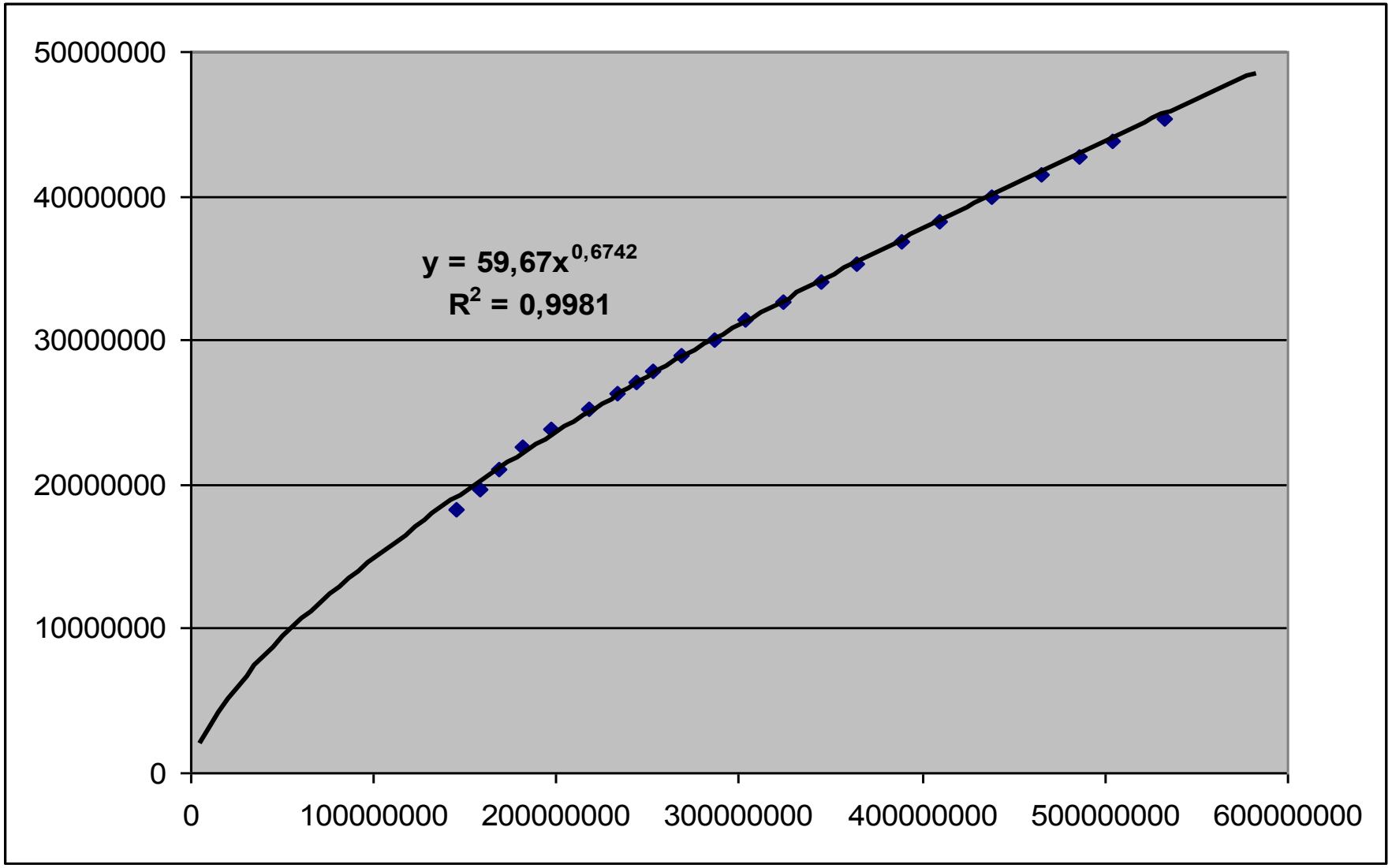
$$V_{WT}(t) + V_{NT}(t) \cong V(t)$$

- Tada se svaka komponenta može modelirati neovisno o drugim komponentama i dobiti optimalni parametri za pojedinu funkciju

Heapsov zakon za hrvatski

	Heapsov zakon	*α*	*β*	*K*
Područje A	V	207,24	0,4578	-778.825
	V_{WT}	662,67	0,3716	-766.928
	V_{NT}	0,4310	0,7467	123.821
Područje B	V	145,13	0,4791	
	V_{WT}	4.281,1	0,2722	
	V_{NT}	0,006014	0,9701	
Područje C	V	819,9	0,3852	
	V_{WT}	5.398,1	0,2592	
	V_{NT}	6,8616	0,5885	

Heapsov zakon primijenjen na digrame



Primjena Heapsova zakona

- Indeks učenja

$$LIi(t) = 100 \cdot \frac{dVi}{dt}$$

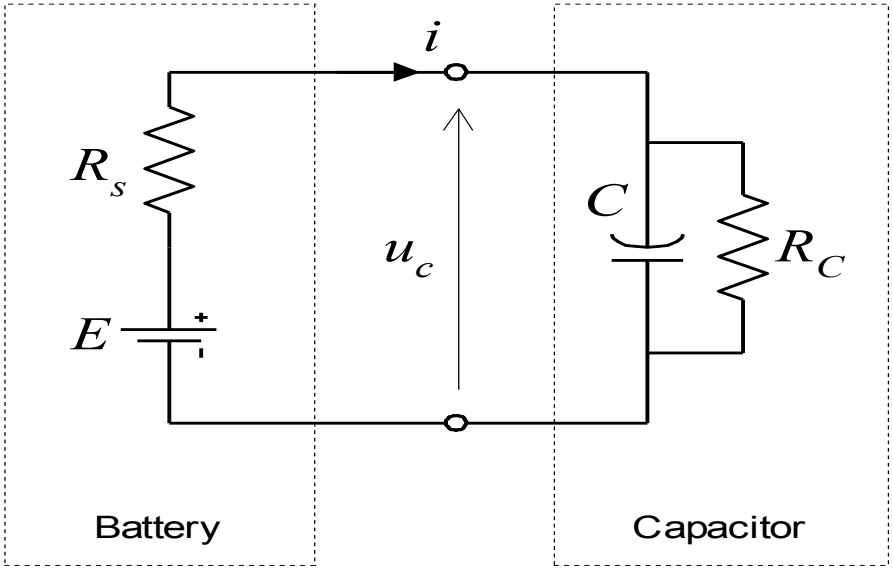
- Na početku korpusa treba voditi računa o eksponencijalnom karakteru Zipfova zakona

$$LI(t) = 100 \cdot \left[\; a + (1-a) \cdot e^{-\frac{t-T}{\tau}} \; \right]$$

- Indeks učenja se na početku ponaša kao kapacitivna struja

Model učenja temeljen na Heapsovu zakonu

- Kognoelektrička analogija



- Omogućava energetsko bilanciranje učenja; daje zapanjujuće dobre rezultate

LI = struja, TC = napon

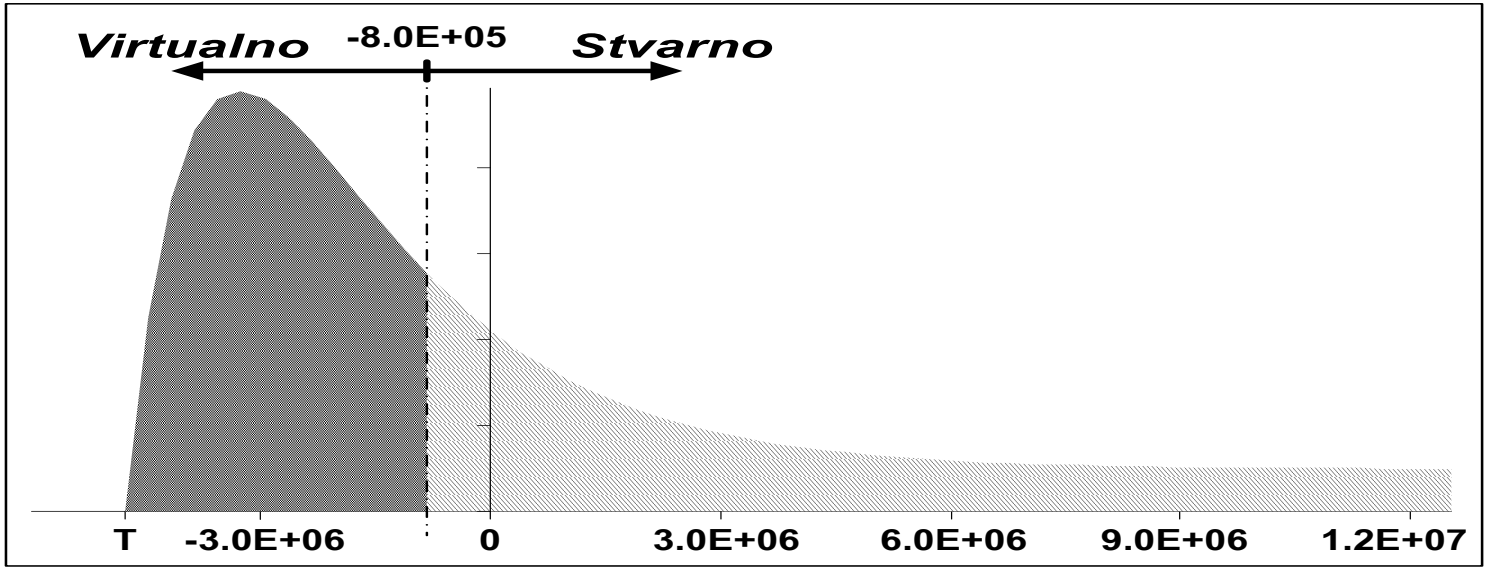
$$li(t) = a + (1-a) \cdot e^{-\frac{t-T}{\tau}}$$

$$tc(t) = b \cdot \left(1 - e^{-\frac{t-T}{\tau}} \right)$$

$$PoL(t) = \left[a + (1-a) \cdot e^{-\frac{t-T}{\tau}} \right] \cdot b \cdot \left(1 - e^{-\frac{t-T}{\tau}} \right)$$

a = 0,025855; b = 0,952595; τ = 2,035.042; T = –4,774.792

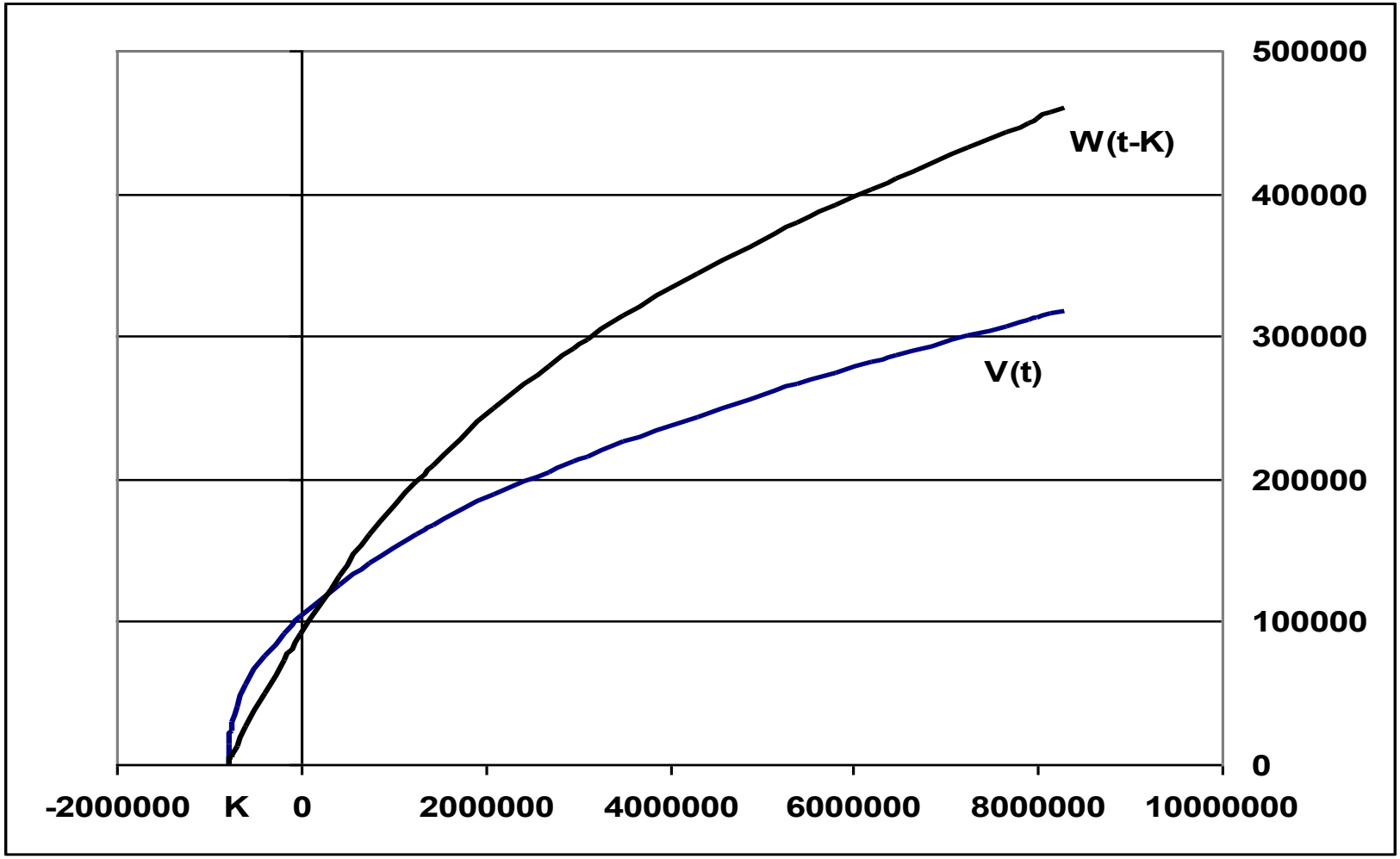
Snaga učenja



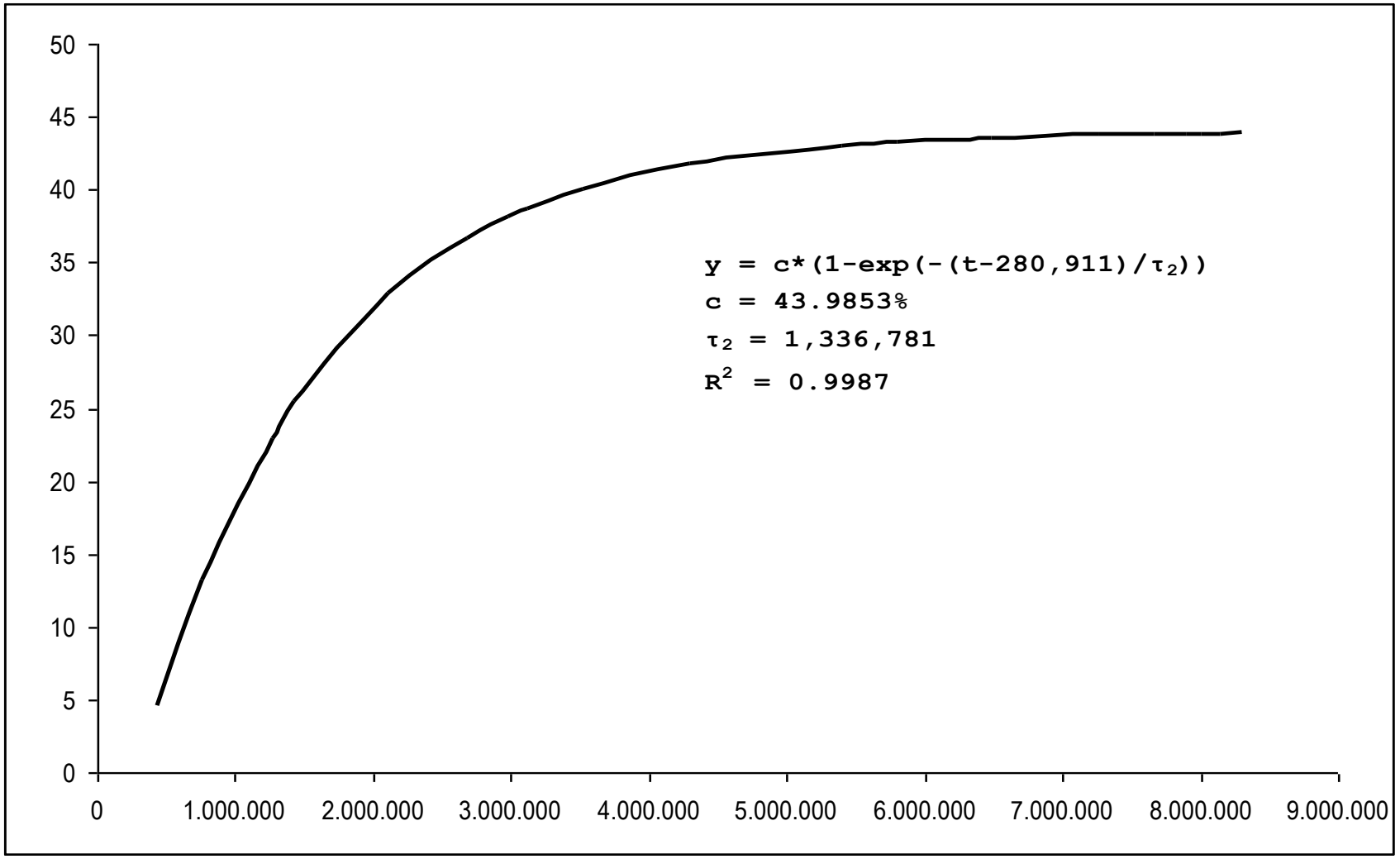
$$W(t) = \int_{T}^{t} PoL(t) \cdot dt = \tau \cdot b \cdot \left[p \cdot a + (1-2a) \cdot (1-e^{-p}) - \frac{1-a}{2} \cdot (1-e^{-2p}) \right]$$

$p = (t - T)/\tau$; $W_{\text{šrafirano}}$ = 750.000 različnica

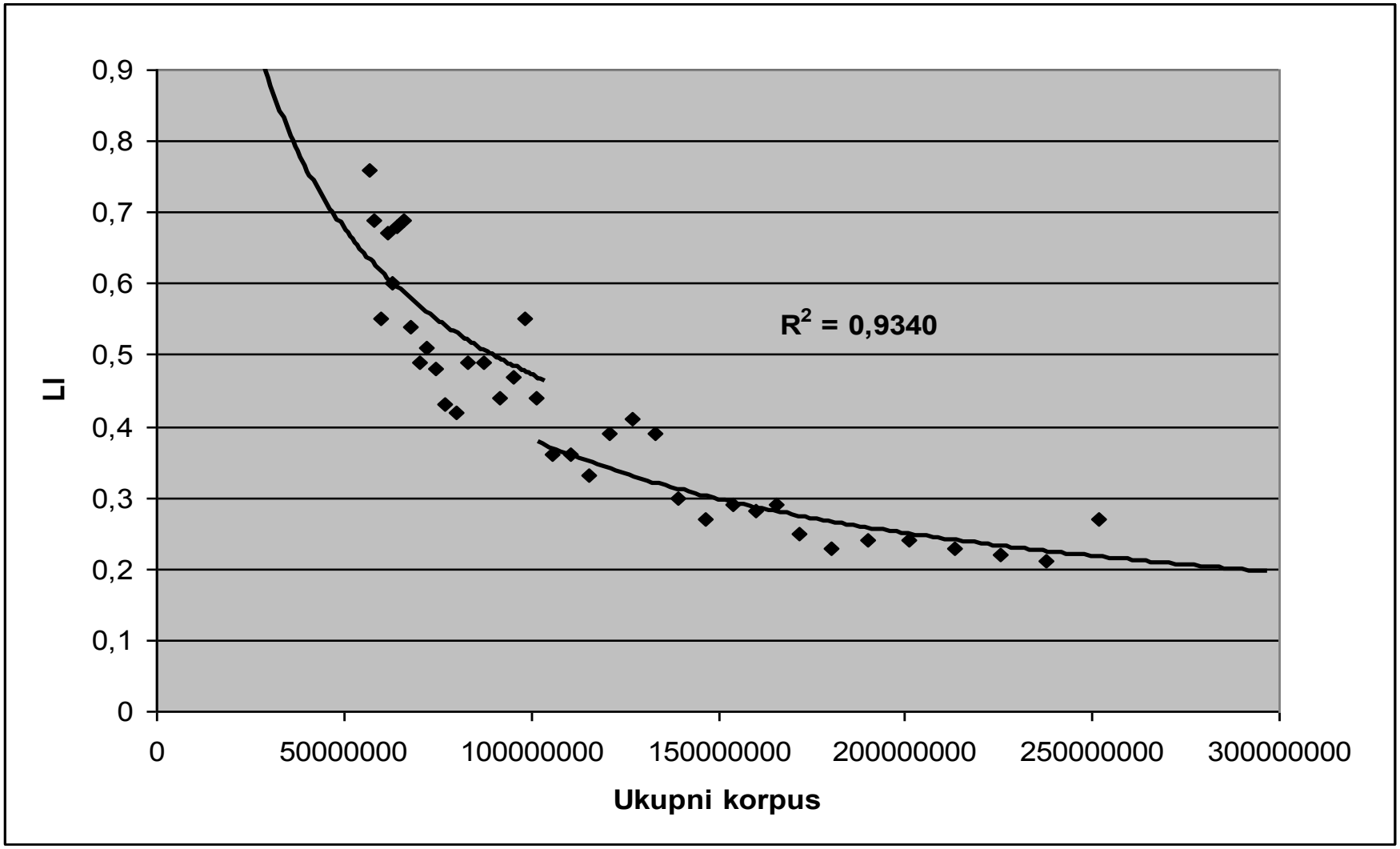
Prikaz W(t-K) i V(t) funkcije u području A



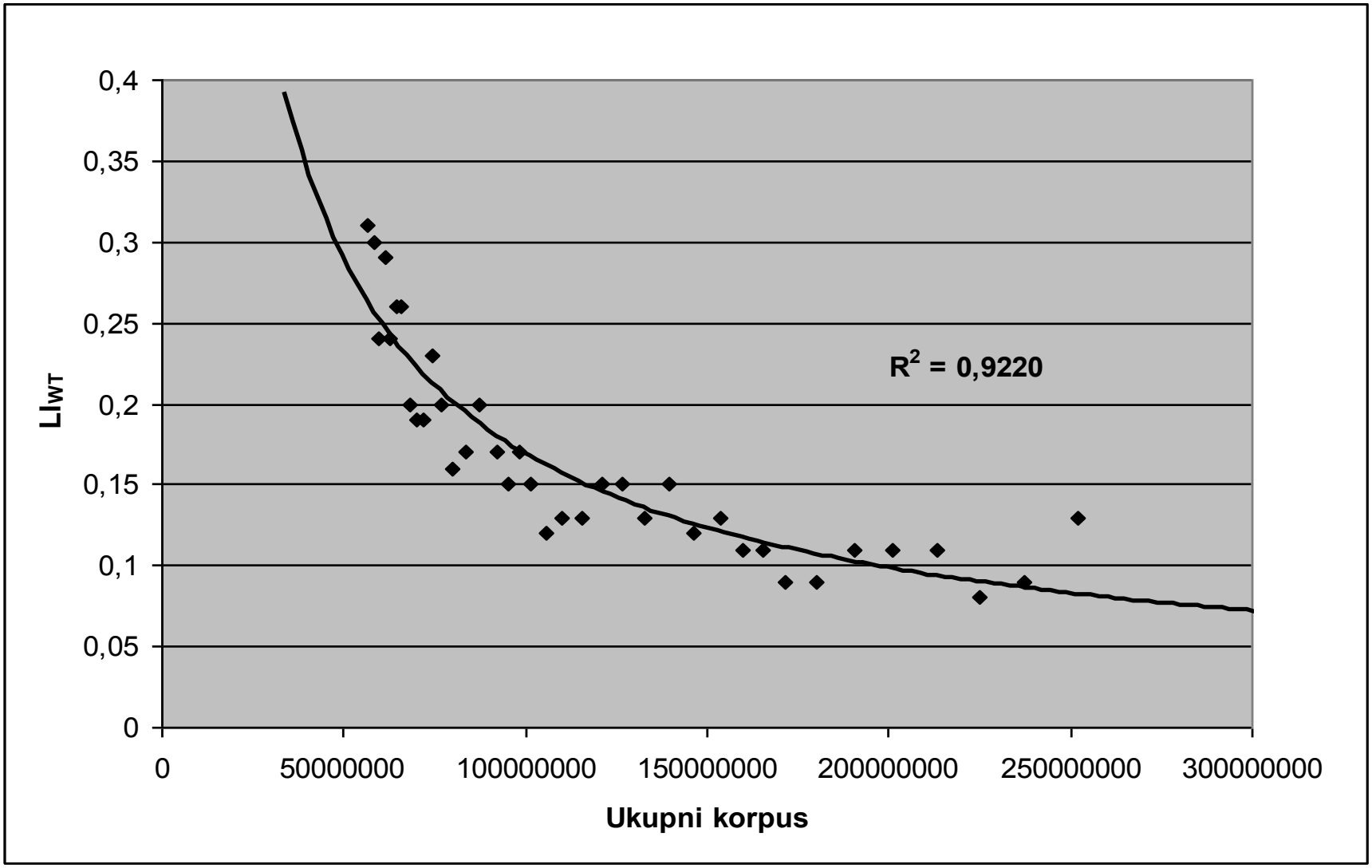
Ponašanje W(t-K)-V(t) funkcije u području A



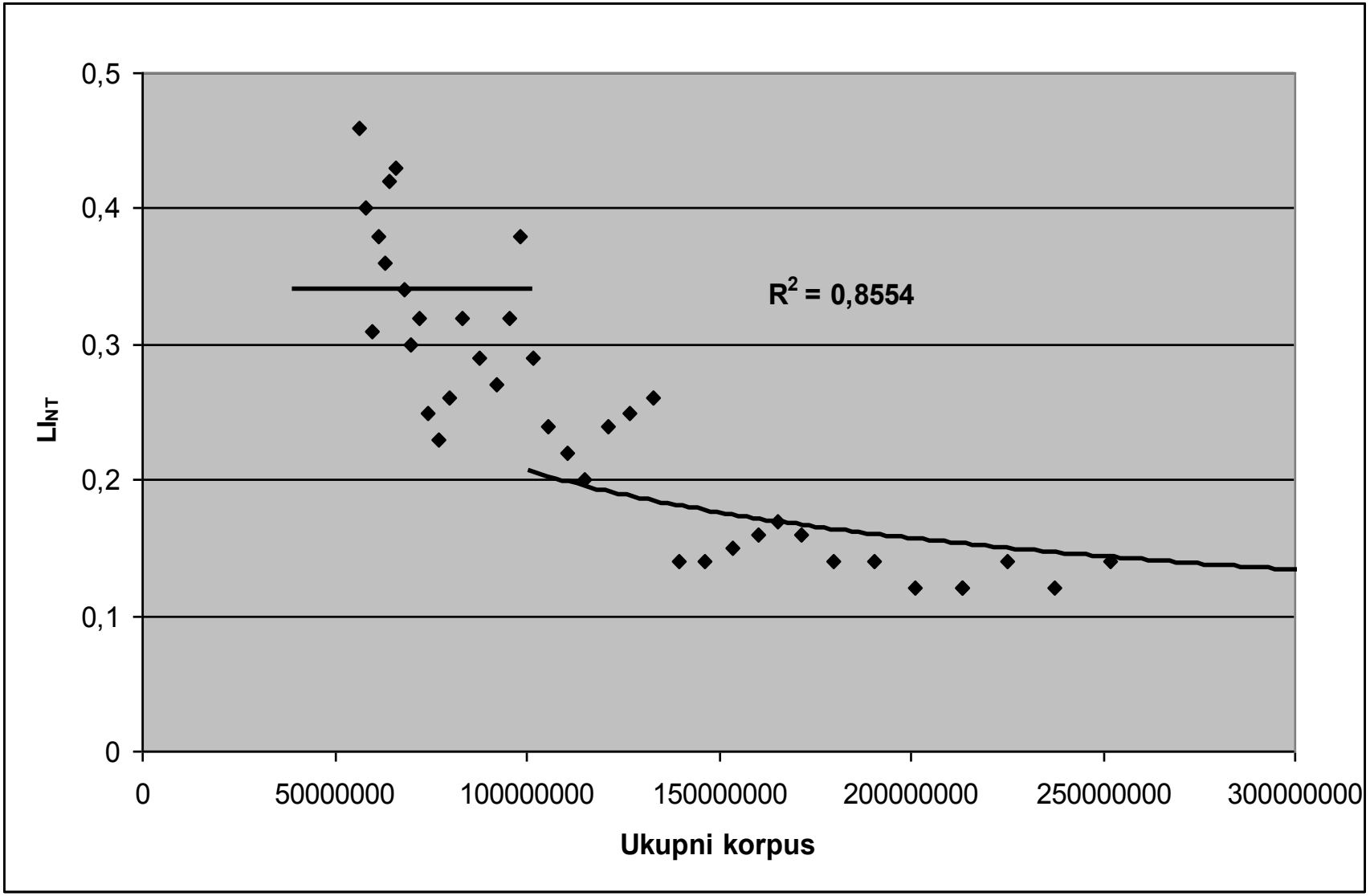
Empirijski podaci i funkcija LI



Empirijski podaci i funkcija LI_{WT}



Empirijski podaci i funkcija LI_{NT}



Zaključak

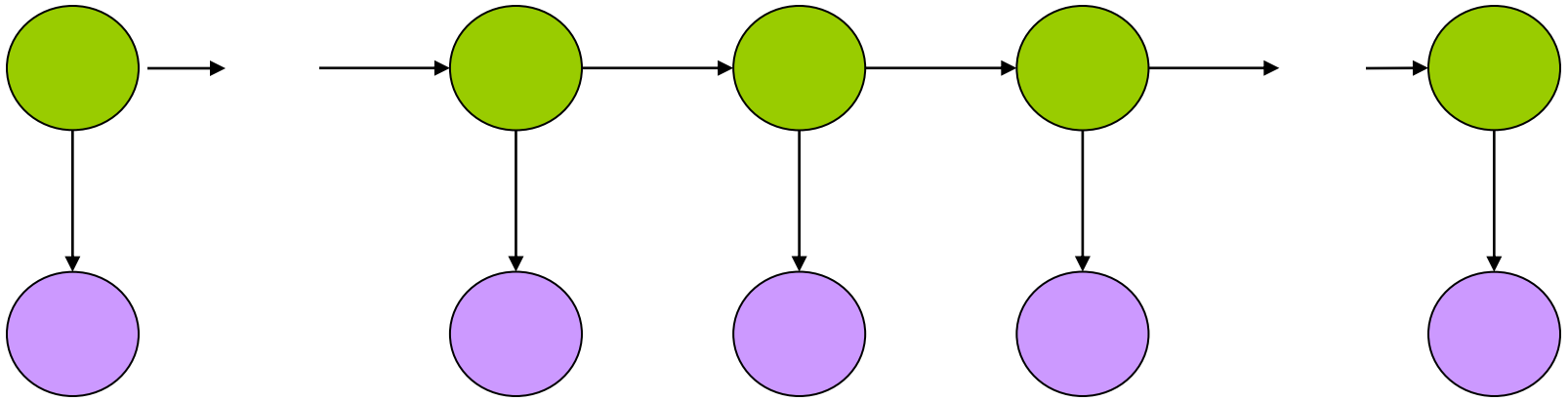
- Prometni modeli i Zipfov, odnosno Heapsov zakon omogućuju planiranje rada nad korpusom
- Sredstvo su predviđanja, jer u NLP-u svaki novi radni zahvat traži i pripremu u programiranju, odnosno reprogramiranju, što iziskuje vrijeme, kako u zahvatima nad programskom podrškom tako i u testiranju novih rješenja
- Važno je pratiti promjene kako prometnih tako i prirodnojezičnih parametara u funkciji "vremena" (opseg obrađenoga korpusa) da bi planiranje bilo dovoljno precizno, time i pouzdano

OBRADA PRIRODNOG JEZIKA (NLP)

Hidden Markov Models (HMM)

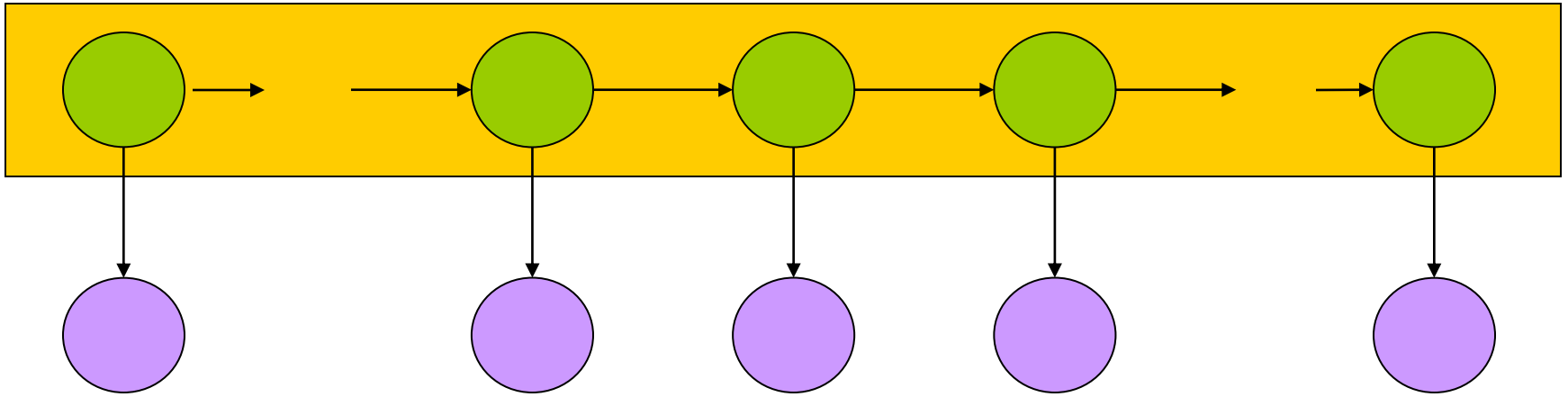
30. listopada 2012.

What is an HMM?



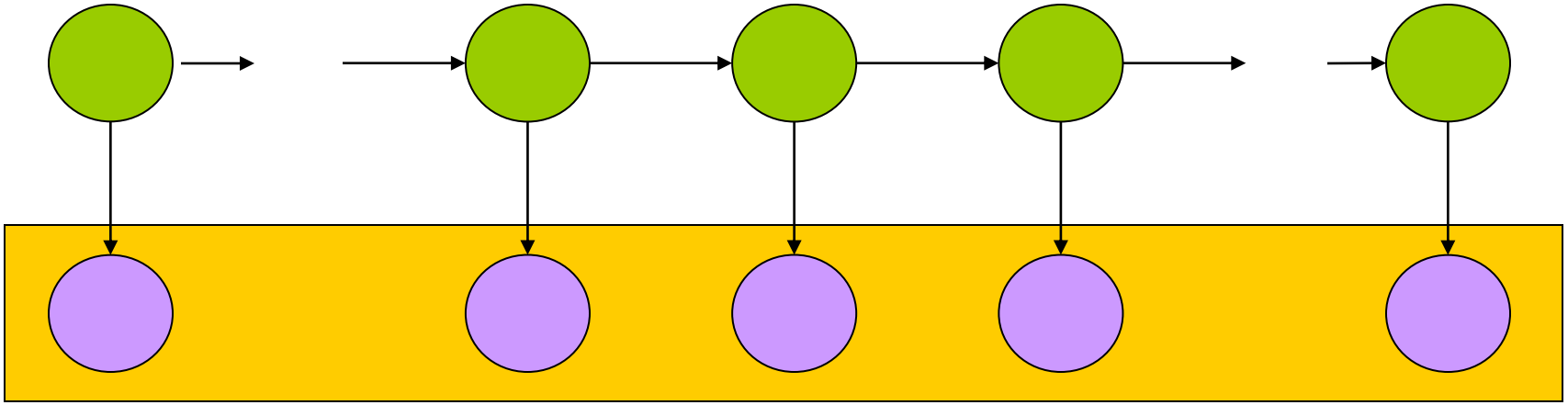
- Graphical Model
- Circles indicate states
- Arrows indicate probabilistic dependencies between states

What is an HMM?



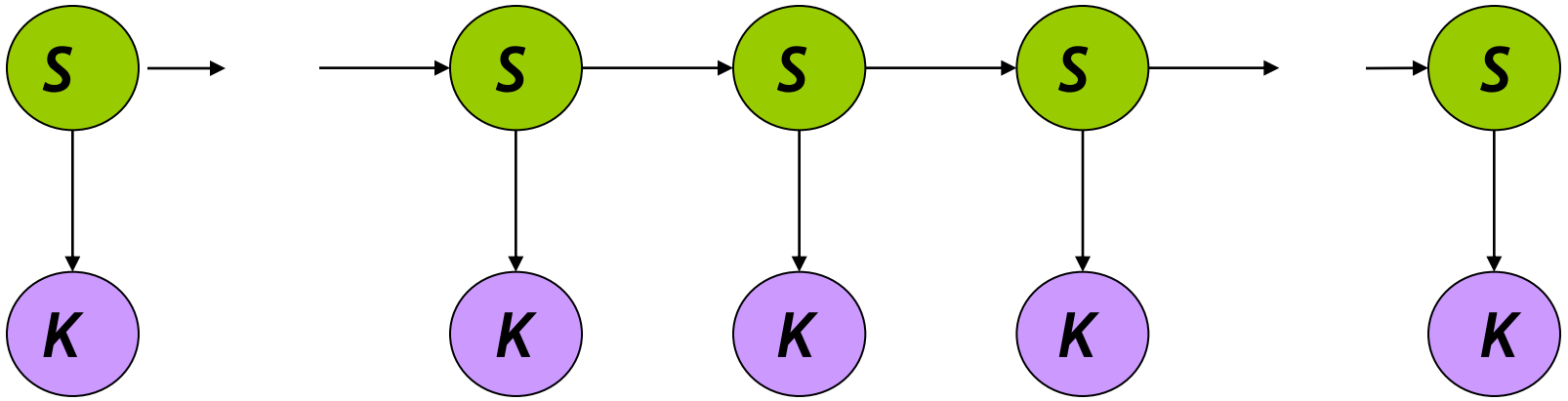
- Green circles are ***hidden states***
- Dependent only on the previous state
- "The past is independent of the future given the present."

What is an HMM?



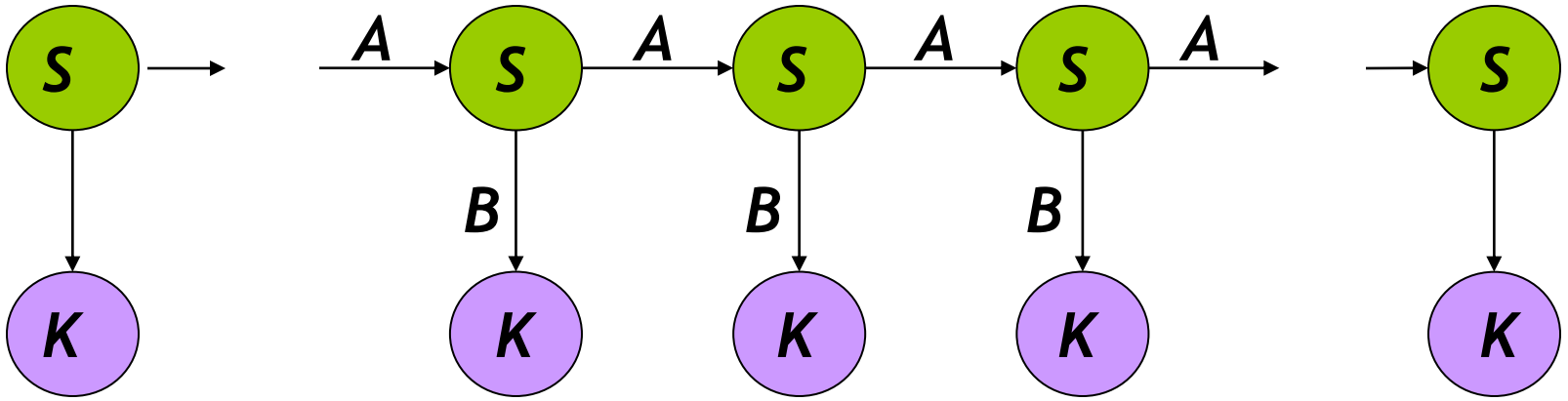
- Purple nodes are ***observed states***
- Dependent only on their corresponding hidden state

HMM Formalism



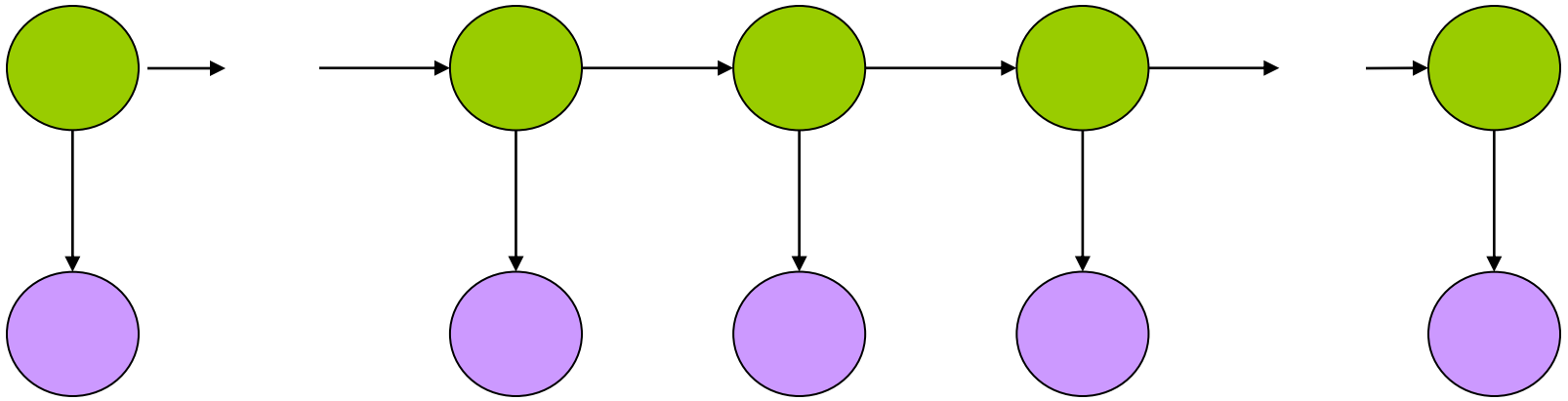
- $\{S, K, \Pi, A, B\}$
- S : $\{s_1 \ldots s_N\}$ are the values for the hidden states
- K : $\{k_1 \ldots k_M\}$ are the values for the observations

HMM Formalism



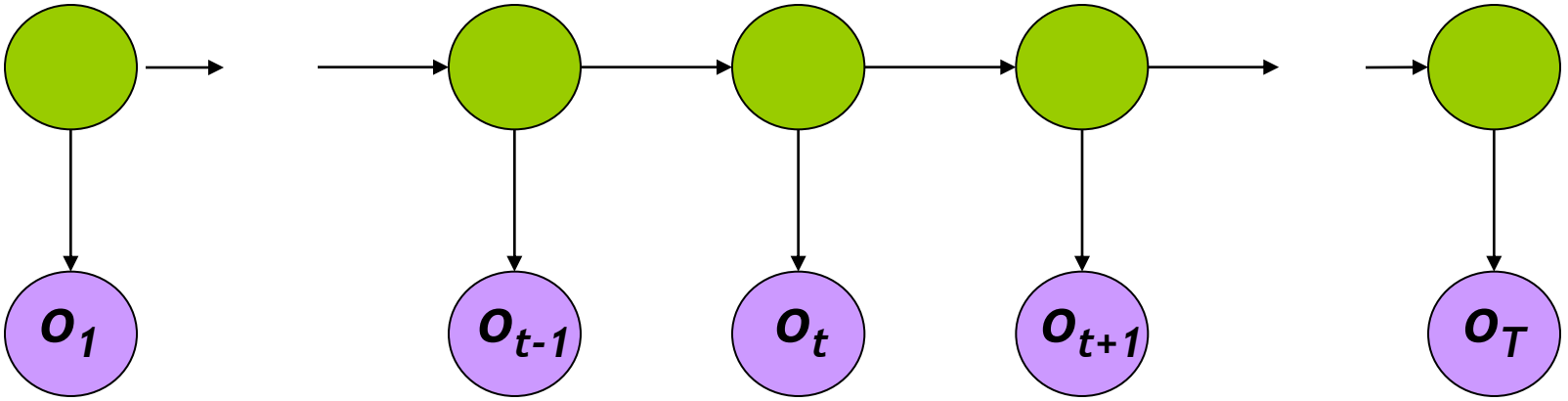
- $\{S, K, \Pi, A, B\}$
- $\Pi = \{\pi_i\}$ are the initial state probabilities
- $A = \{a_{ij}\}$ are the state transition probabilities
- $B = \{b_{ik}\}$ are the observation state probabilities

Inference in an HMM



- Compute the probability of a given observation sequence
- Given an observation sequence, compute the most likely hidden state sequence
- Given an observation sequence and set of possible models, which model most closely fits the data?

Decoding

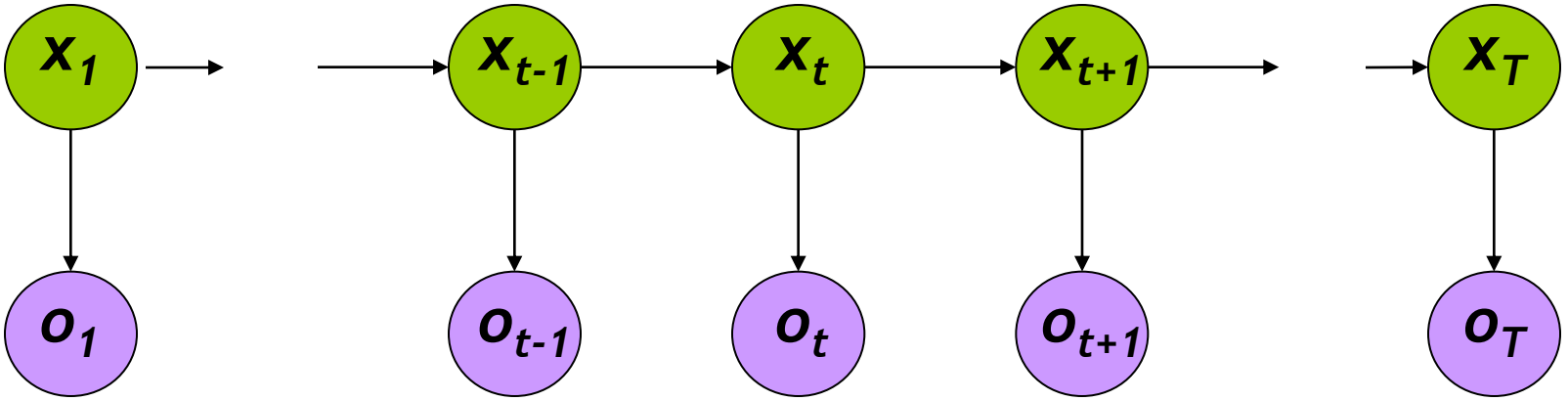


Given an observation sequence and a model, compute the probability of the observation sequence

$$O = (o_1 \ldots o_T),\ \mu = (A, B, \Pi)$$

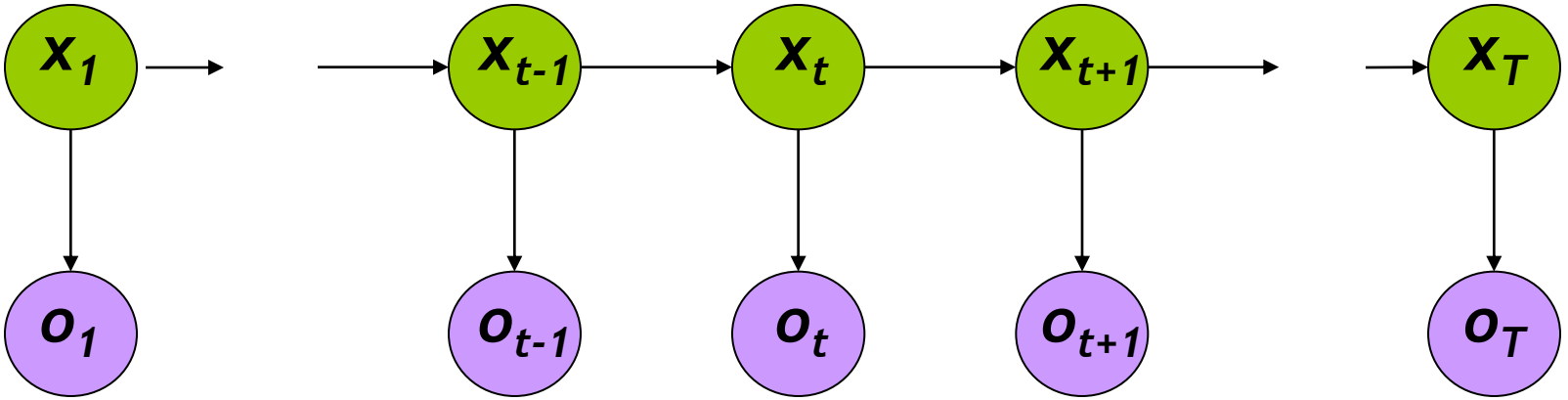
$$\text{Compute}\, P(O \mid \mu)$$

Decoding



$$P(O \mid X, \mu) = b_{x_1 o_1} b_{x_2 o_2} \ldots b_{x_T o_T}$$

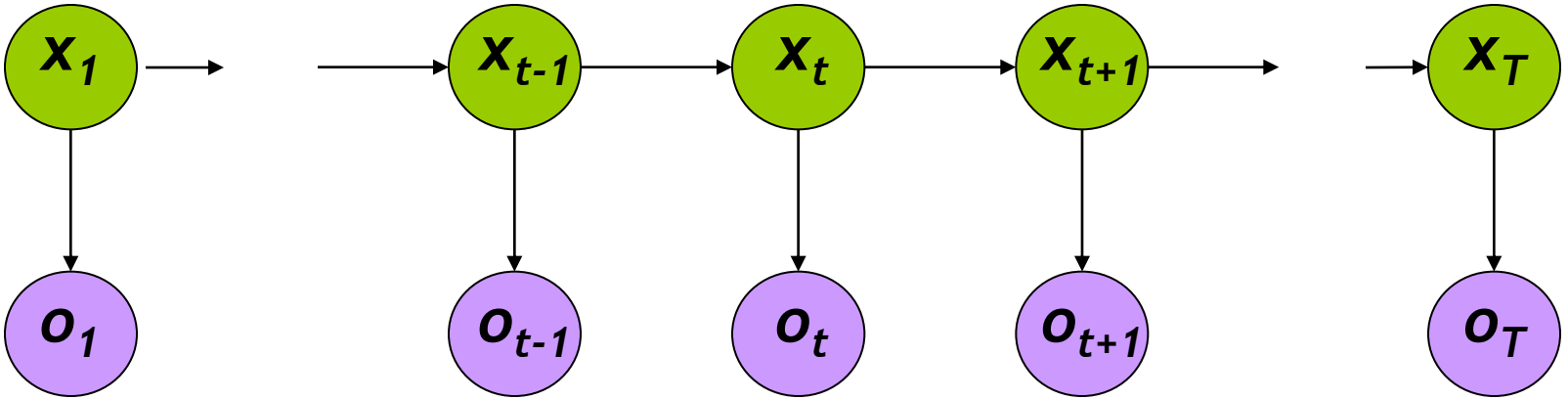
Decoding



$$P(O \mid X, \mu) = b_{x_1 o_1} b_{x_2 o_2} \ldots b_{x_T o_T}$$

$$P(X \mid \mu) = \pi_{x_1} a_{x_1 x_2} a_{x_2 x_3} \ldots a_{x_{T-1} x_T}$$

Decoding

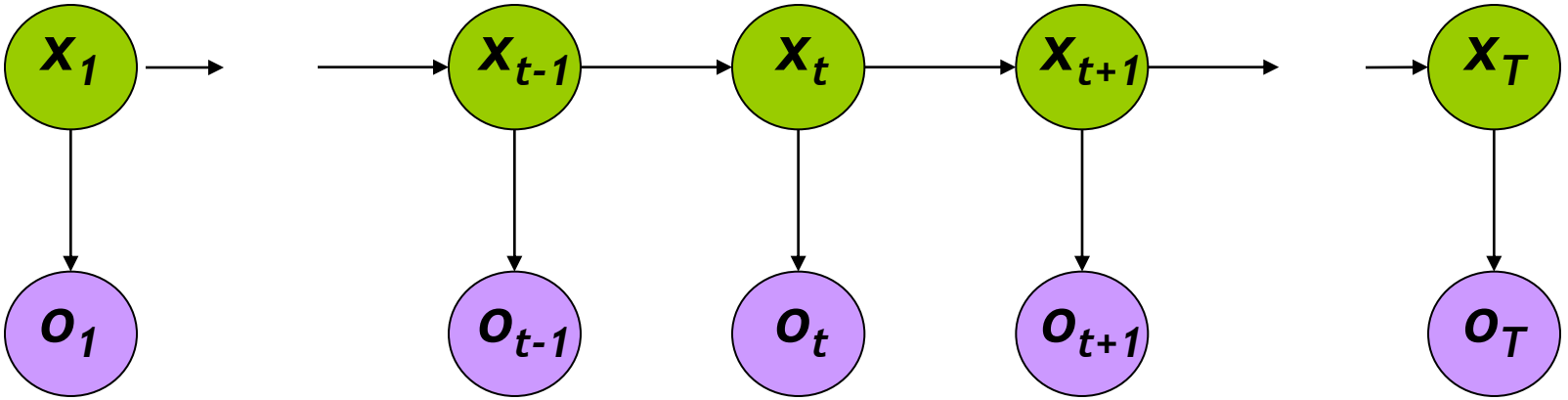


$$P(O \mid X, \mu) = b_{x_1 o_1} b_{x_2 o_2} \ldots b_{x_T o_T}$$

$$P(X \mid \mu) = \pi_{x_1} a_{x_1 x_2} a_{x_2 x_3} \ldots a_{x_{T-1} x_T}$$

$$P(O, X \mid \mu) = P(O \mid X, \mu) P(X \mid \mu)$$

Decoding



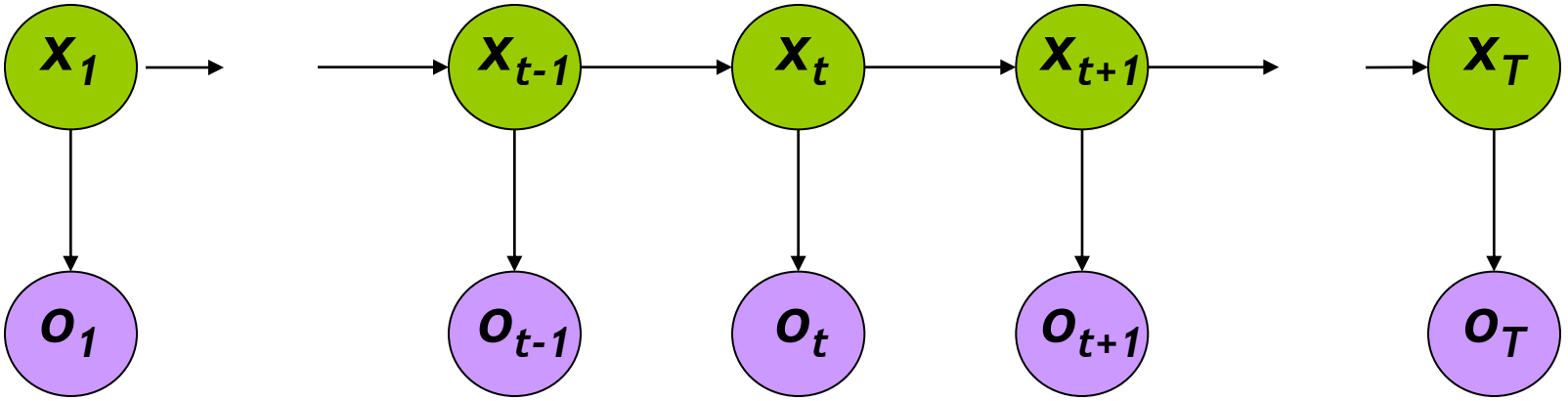
$$P(O \mid X, \mu) = b_{x_1 o_1} b_{x_2 o_2} \ldots b_{x_T o_T}$$

$$P(X \mid \mu) = \pi_{x_1} a_{x_1 x_2} a_{x_2 x_3} \ldots a_{x_{T-1} x_T}$$

$$P(O, X \mid \mu) = P(O \mid X, \mu) P(X \mid \mu)$$

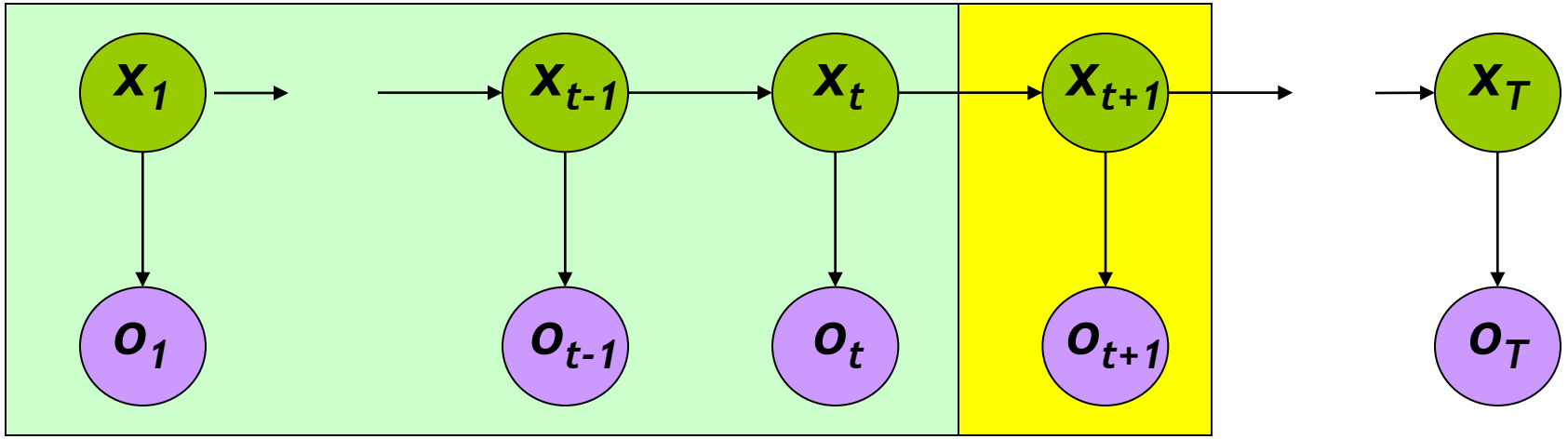
$$P(O \mid \mu) = \sum_X P(O \mid X, \mu) P(X \mid \mu)$$

Decoding



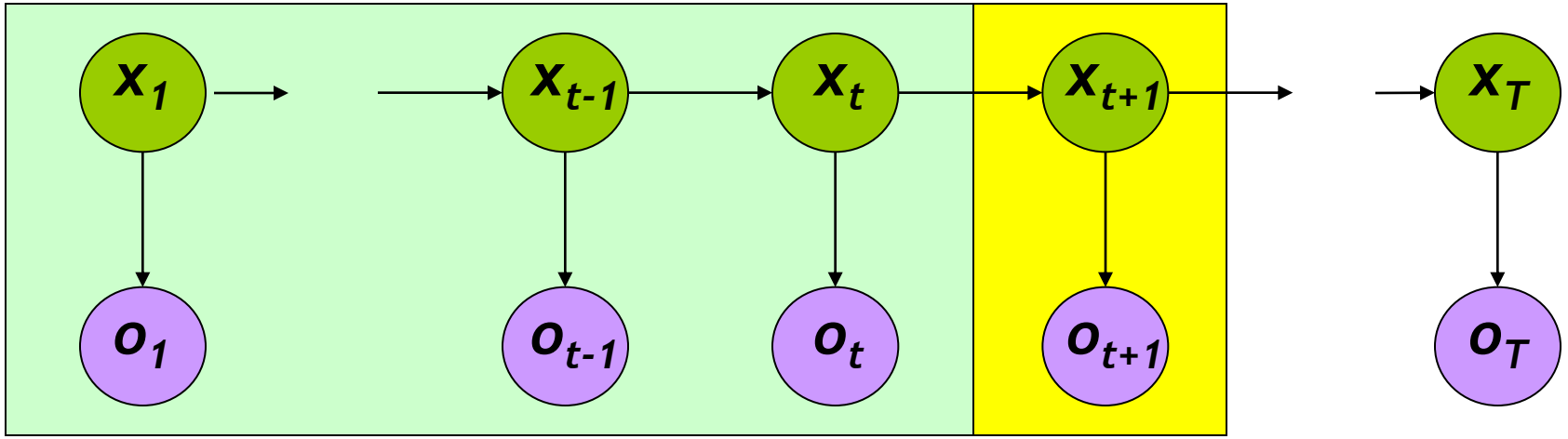
$$P(O \mid \mu) = \sum_{\{x_1 \ldots x_T\}} \pi_{x_1} b_{x_1 o_1} \prod_{t=1}^{T-1} a_{x_t x_{t+1}} b_{x_{t+1} o_{t+1}}$$

Forward Procedure



- Special structure gives us an efficient solution using *dynamic programming*.
- Intuition: Probability of the first *t* observations is the same for all possible *t*+1 length state sequences.
- Define: $\alpha_i(t) = P(o_1 \ldots o_t, x_t = i \mid \mu)$

Forward Procedure



$$\alpha_j(t+1)$$

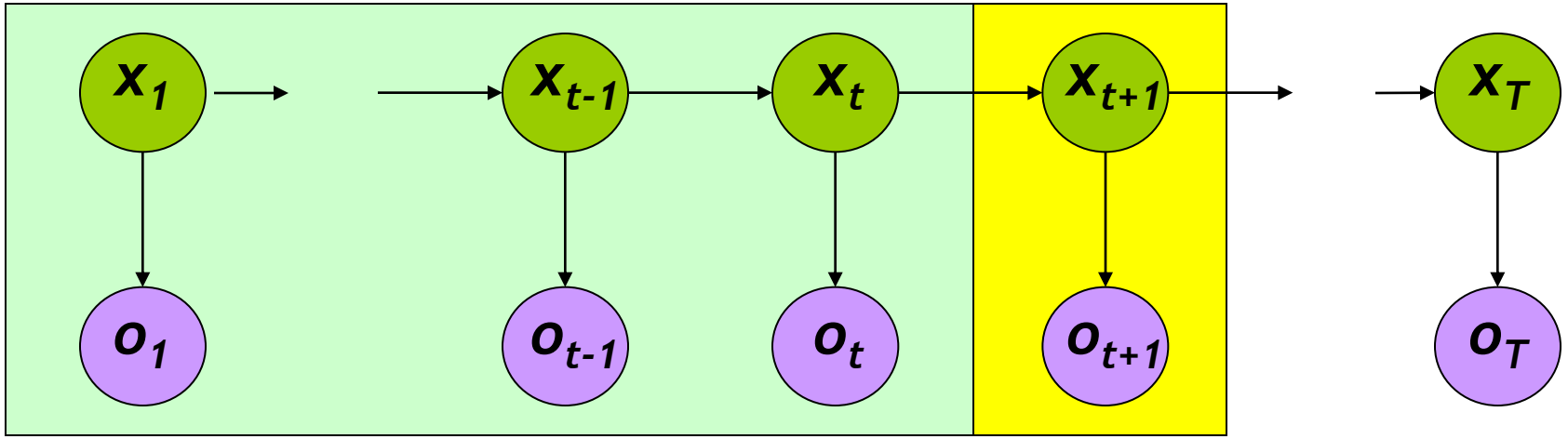
$$= P(o_1 \ldots o_{t+1}, x_{t+1} = j)$$

$$= P(o_1 \ldots o_{t+1} \mid x_{t+1} = j) P(x_{t+1} = j)$$

$$= P(o_1 \ldots o_t \mid x_{t+1} = j) P(o_{t+1} \mid x_{t+1} = j) P(x_{t+1} = j)$$

$$= P(o_1 \ldots o_t, x_{t+1} = j) P(o_{t+1} \mid x_{t+1} = j)$$

Forward Procedure



$$\alpha_j(t+1)$$

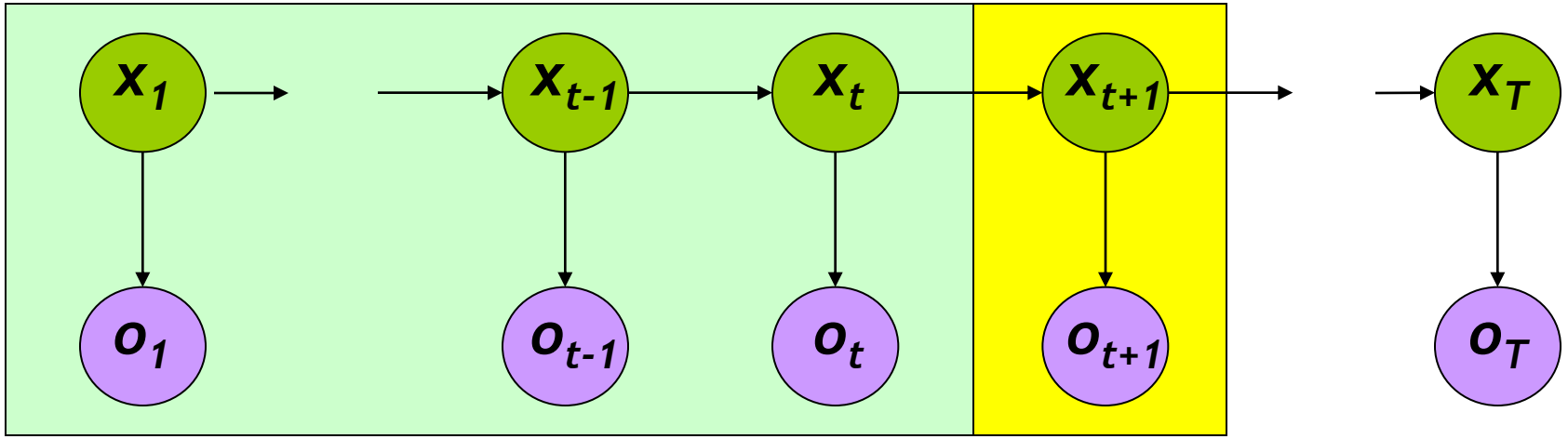
$$= P(o_1 \ldots o_{t+1}, x_{t+1} = j)$$

$$= P(o_1 \ldots o_{t+1} \mid x_{t+1} = j) P(x_{t+1} = j)$$

$$= P(o_1 \ldots o_t \mid x_{t+1} = j) P(o_{t+1} \mid x_{t+1} = j) P(x_{t+1} = j)$$

$$= P(o_1 \ldots o_t, x_{t+1} = j) P(o_{t+1} \mid x_{t+1} = j)$$

Forward Procedure



$$\alpha_j(t+1)$$

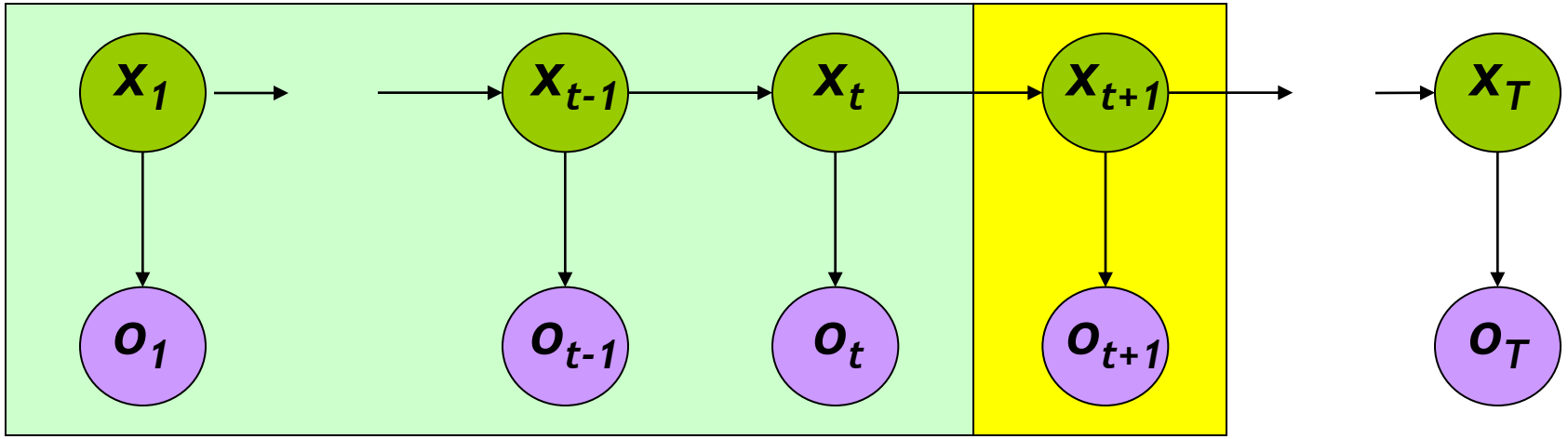
$$= P(o_1 \ldots o_{t+1}, x_{t+1} = j)$$

$$= P(o_1 \ldots o_{t+1} \mid x_{t+1} = j) P(x_{t+1} = j)$$

$$= P(o_1 \ldots o_t \mid x_{t+1} = j) P(o_{t+1} \mid x_{t+1} = j) P(x_{t+1} = j)$$

$$= P(o_1 \ldots o_t, x_{t+1} = j) P(o_{t+1} \mid x_{t+1} = j)$$

Forward Procedure



$$\alpha_j(t+1)$$

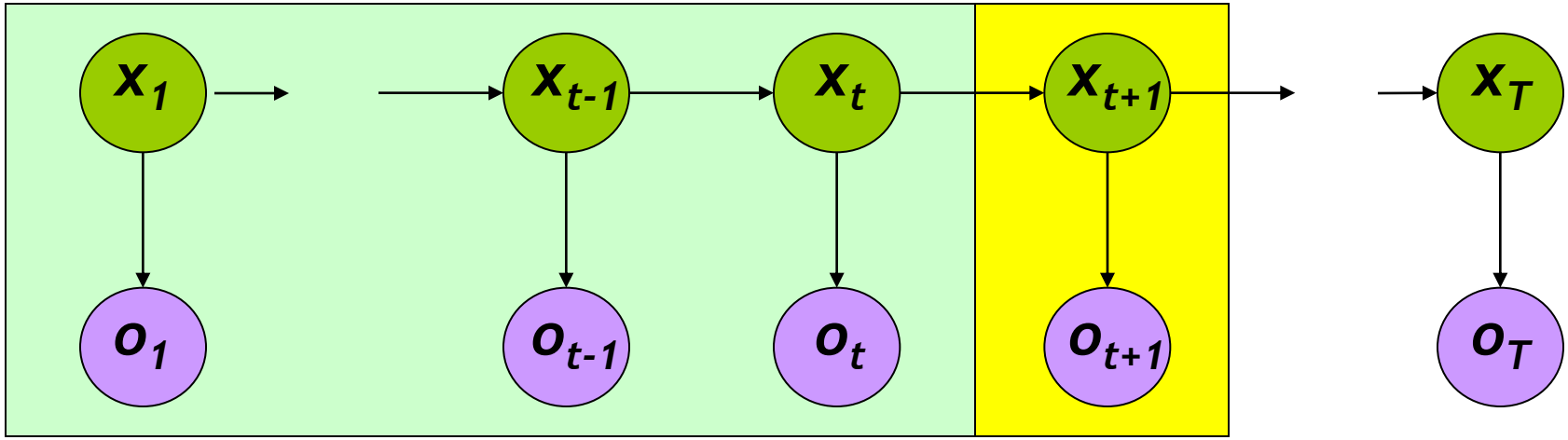
$$= P(o_1 \ldots o_{t+1}, x_{t+1} = j)$$

$$= P(o_1 \ldots o_{t+1} \mid x_{t+1} = j) P(x_{t+1} = j)$$

$$= P(o_1 \ldots o_t \mid x_{t+1} = j) P(o_{t+1} \mid x_{t+1} = j) P(x_{t+1} = j)$$

$$= P(o_1 \ldots o_t, x_{t+1} = j) P(o_{t+1} \mid x_{t+1} = j)$$

Forward Procedure



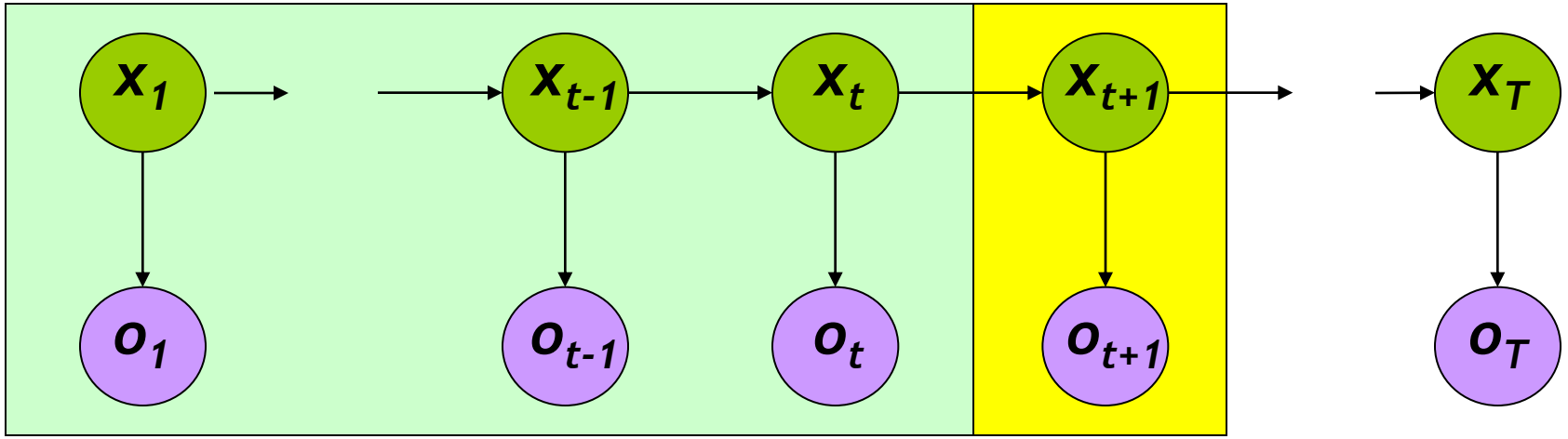
$$= \sum_{i=1\ldots N} P(o_1 \ldots o_t, x_t = i, x_{t+1} = j) P(o_{t+1} \mid x_{t+1} = j)$$

$$= \sum_{i=1\ldots N} P(o_1 \ldots o_t, x_{t+1} = j \mid x_t = i) P(x_t = i) P(o_{t+1} \mid x_{t+1} = j)$$

$$= \sum_{i=1\ldots N} P(o_1 \ldots o_t, x_t = i) P(x_{t+1} = j \mid x_t = i) P(o_{t+1} \mid x_{t+1} = j)$$

$$= \sum_{i=1\ldots N} \alpha_i(t) a_{ij} b_{jo_{t+1}}$$

Forward Procedure



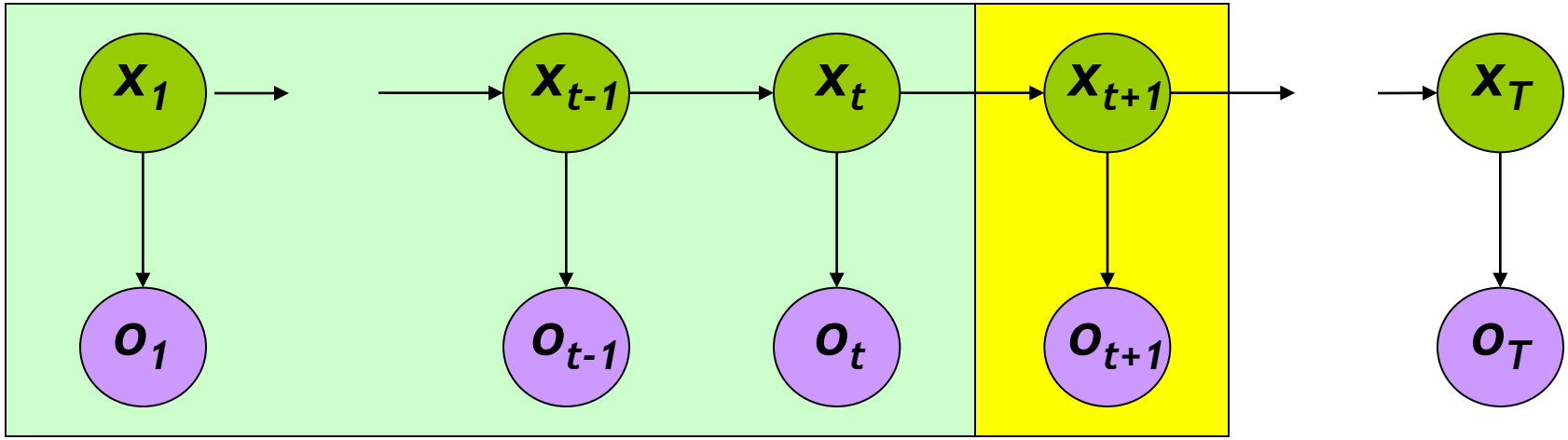
$$= \sum_{i=1\ldots N} P(o_1 \ldots o_t, x_t = i, x_{t+1} = j) P(o_{t+1} \mid x_{t+1} = j)$$

$$= \sum_{i=1\ldots N} P(o_1 \ldots o_t, x_{t+1} = j \mid x_t = i) P(x_t = i) P(o_{t+1} \mid x_{t+1} = j)$$

$$= \sum_{i=1\ldots N} P(o_1 \ldots o_t, x_t = i) P(x_{t+1} = j \mid x_t = i) P(o_{t+1} \mid x_{t+1} = j)$$

$$= \sum_{i=1\ldots N} \alpha_i(t) a_{ij} b_{jo_{t+1}}$$

Forward Procedure



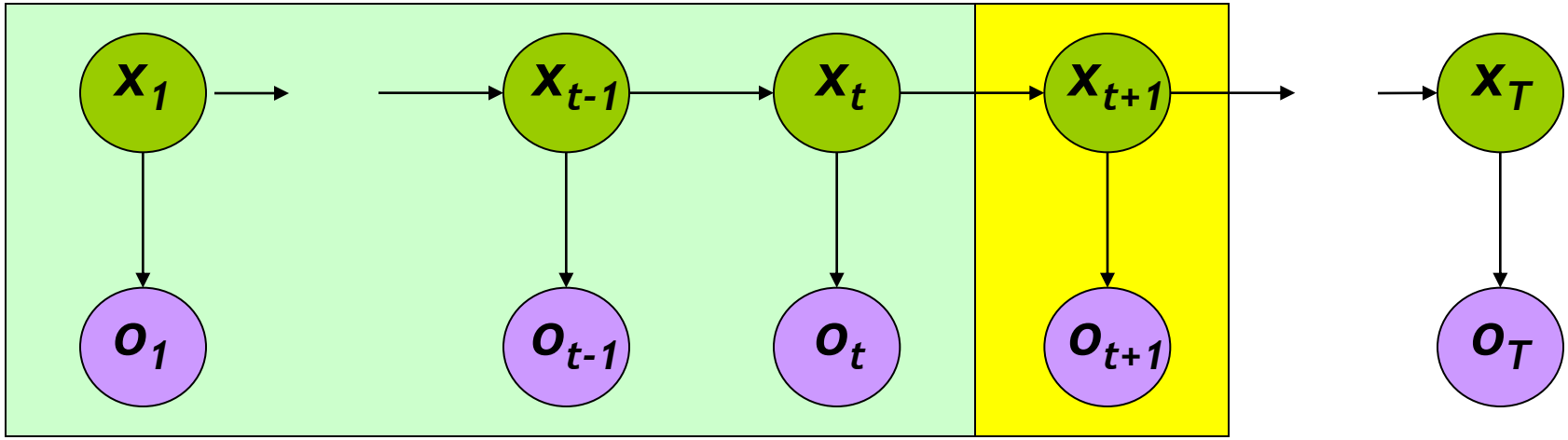
$$= \sum_{i=1\ldots N} P(o_1 \ldots o_t, x_t = i, x_{t+1} = j) P(o_{t+1} \mid x_{t+1} = j)$$

$$= \sum_{i=1\ldots N} P(o_1 \ldots o_t, x_{t+1} = j \mid x_t = i) P(x_t = i) P(o_{t+1} \mid x_{t+1} = j)$$

$$= \sum_{i=1\ldots N} P(o_1 \ldots o_t, x_t = i) P(x_{t+1} = j \mid x_t = i) P(o_{t+1} \mid x_{t+1} = j)$$

$$= \sum_{i=1\ldots N} \alpha_i(t) a_{ij} b_{jo_{t+1}}$$

Forward Procedure



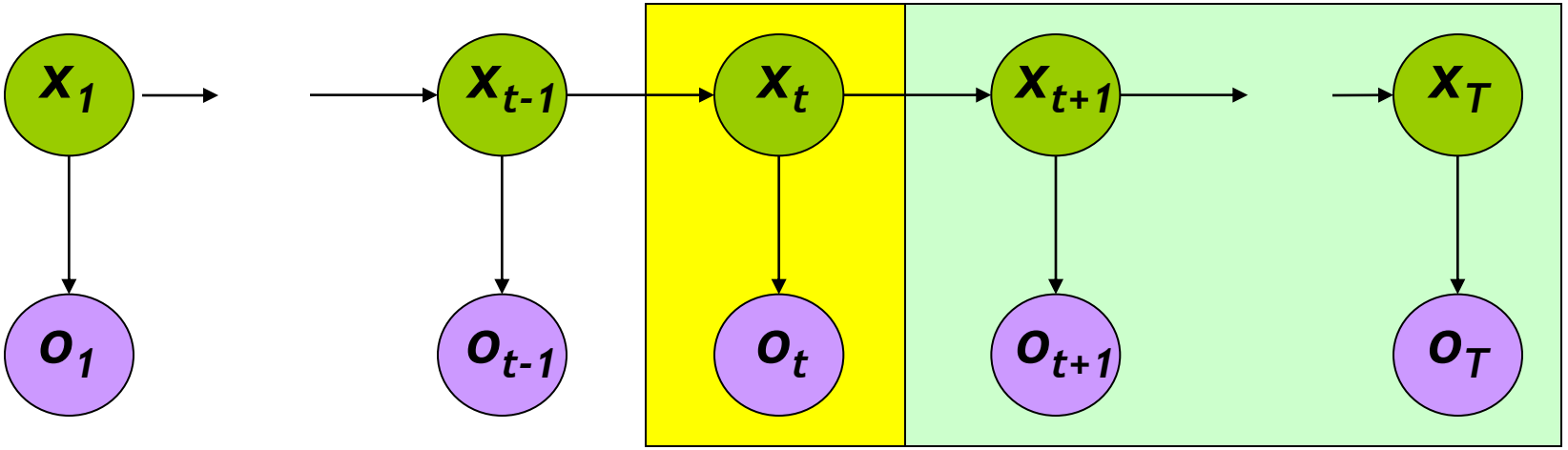
$$= \sum_{i=1\ldots N} P(o_1 \ldots o_t, x_t = i, x_{t+1} = j) P(o_{t+1} \mid x_{t+1} = j)$$

$$= \sum_{i=1\ldots N} P(o_1 \ldots o_t, x_{t+1} = j \mid x_t = i) P(x_t = i) P(o_{t+1} \mid x_{t+1} = j)$$

$$= \sum_{i=1\ldots N} P(o_1 \ldots o_t, x_t = i) P(x_{t+1} = j \mid x_t = i) P(o_{t+1} \mid x_{t+1} = j)$$

$$= \sum_{i=1\ldots N} \alpha_i(t) a_{ij} b_{jo_{t+1}}$$

Backward Procedure



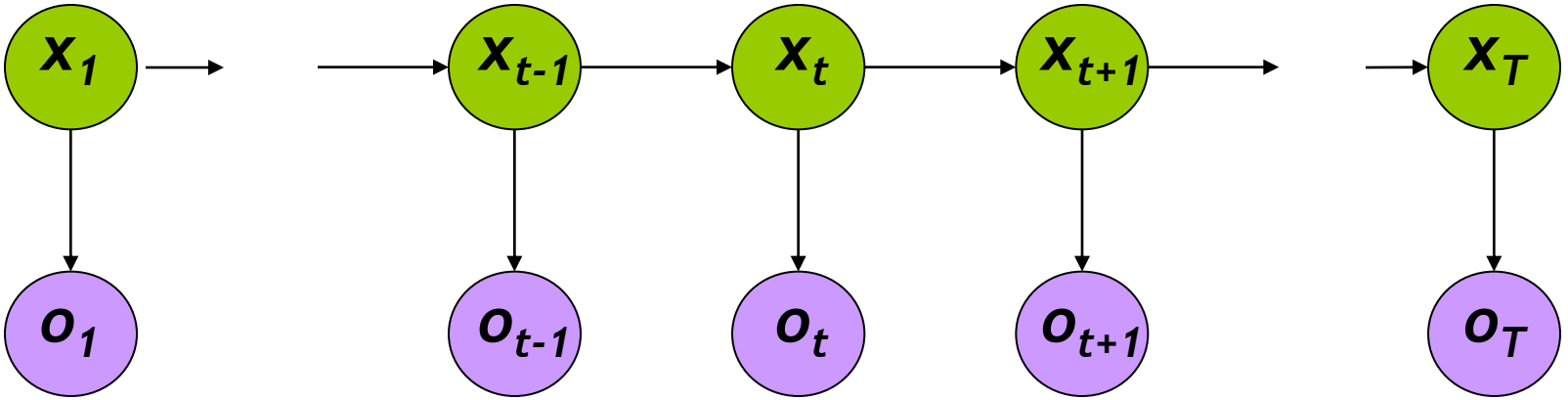
$$\beta_i(T+1) = 1$$

$$\beta_i(t) = P(o_t \ldots o_T \mid x_t = i)$$

$$\beta_i(t) = \sum_{j=1\ldots N} a_{ij} b_{io_t} \beta_j(t+1)$$

Probability of the rest of the states given the first state

Decoding Solution

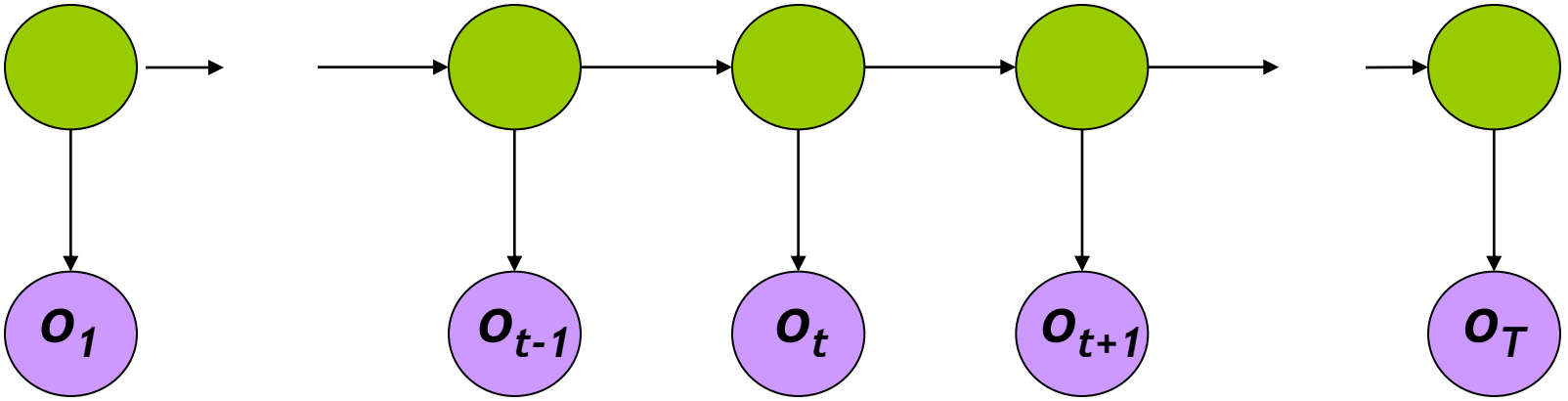


$$P(O \mid \mu) = \sum_{i=1}^{N} \alpha_i(T)$$ **Forward Procedure**

$$P(O \mid \mu) = \sum_{i=1}^{N} \pi_i \beta_i(1)$$ **Backward Procedure**

$$P(O \mid \mu) = \sum_{i=1}^{N} \alpha_i(t)\beta_i(t)$$ **Combination**

Best State Sequence

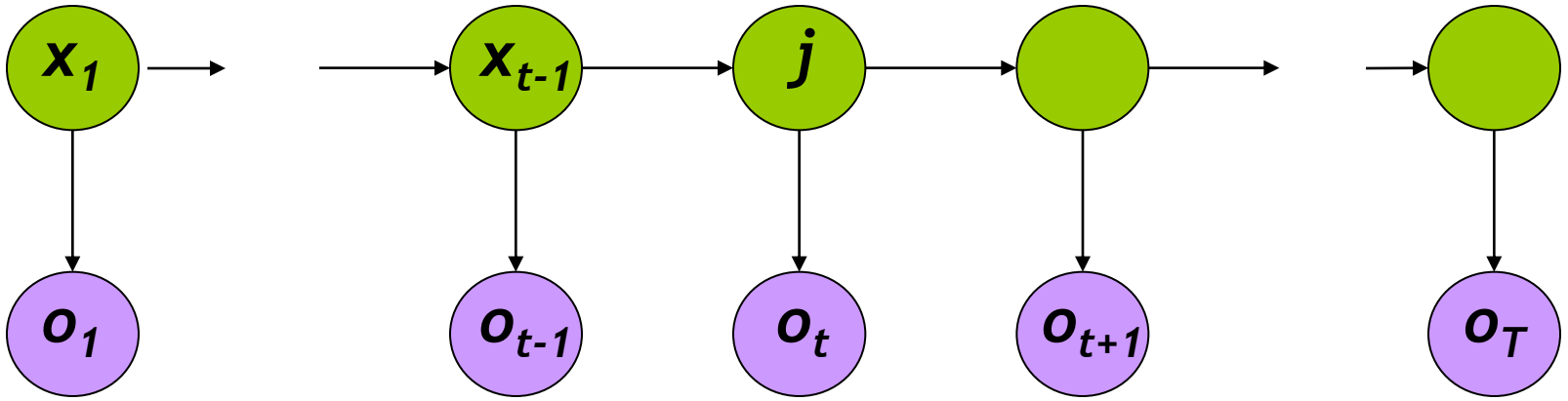


- Find the state sequence that best explains the observations

- **Viterbi** algorithm

$$\arg\max_X P(X \mid O)$$

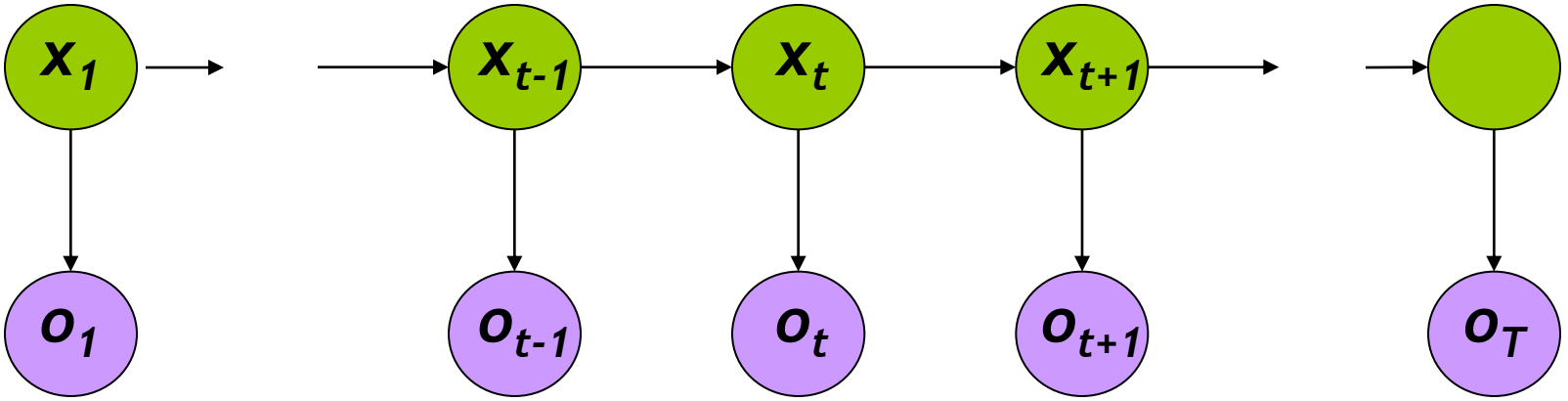
Viterbi Algorithm



$$\delta_j(t) = \max_{x_1 \ldots x_{t-1}} P(x_1 \ldots x_{t-1}, o_1 \ldots o_{t-1}, x_t = j, o_t)$$

The state sequence which maximizes the probability of seeing the observations to time t-1, landing in state j, and seeing the observation at time t

Viterbi Algorithm



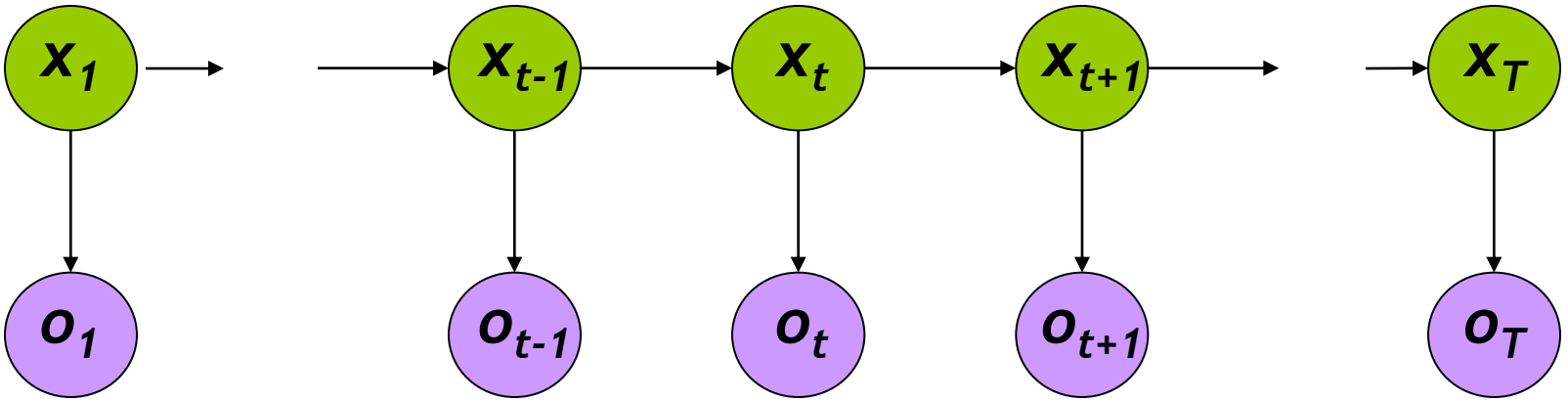
$$\delta_j(t) = \max_{x_1 \ldots x_{t-1}} P(x_1 \ldots x_{t-1}, o_1 \ldots o_{t-1}, x_t = j, o_t)$$

$$\delta_j(t+1) = \max_i \delta_i(t) a_{ij} b_{jo_{t+1}}$$

$$\psi_j(t+1) = \arg\max_i \delta_i(t) a_{ij} b_{jo_{t+1}}$$

Recursive Computation

Viterbi Algorithm



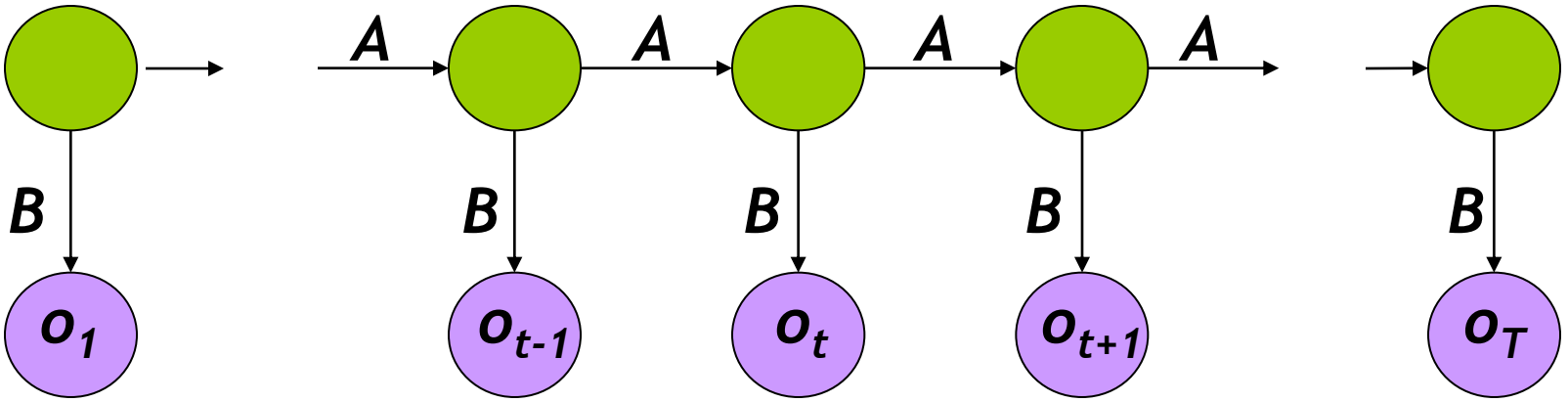
$$\hat{X}_T = \arg\max_i \delta_i(T)$$

$$\hat{X}_t = \psi_{\hat{X}_{t+1}}(t+1)$$

$$P(\hat{X}) = \arg\max_i \delta_i(T)$$

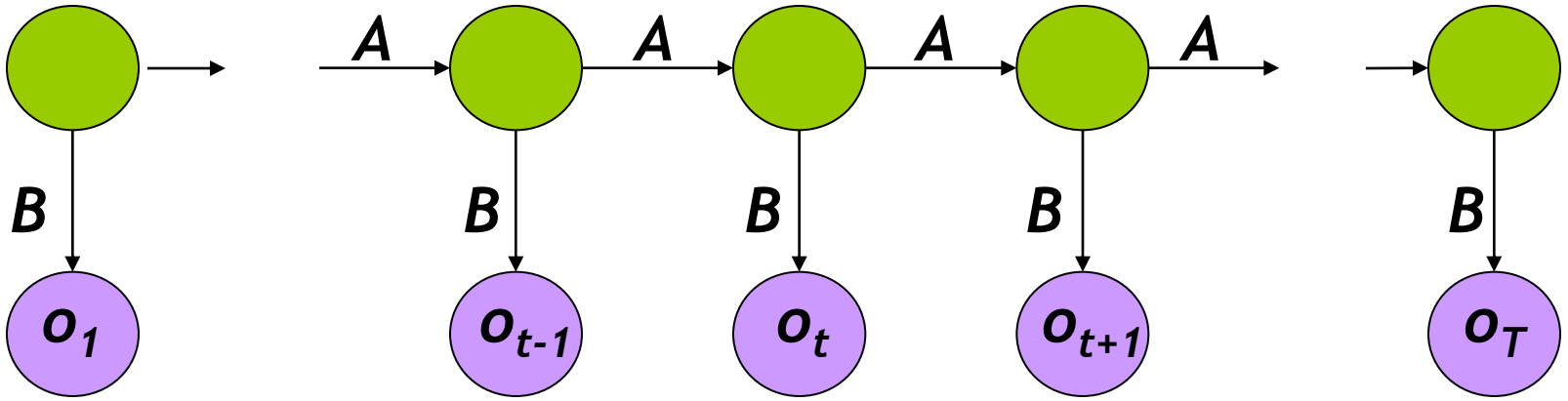
Compute the most likely state sequence by working backwards

Parameter Estimation



- Given an observation sequence, find the model that is most likely to produce that sequence.
- No analytic method
- Given a model and observation sequence, update the model parameters to better fit the observations.

Parameter Estimation



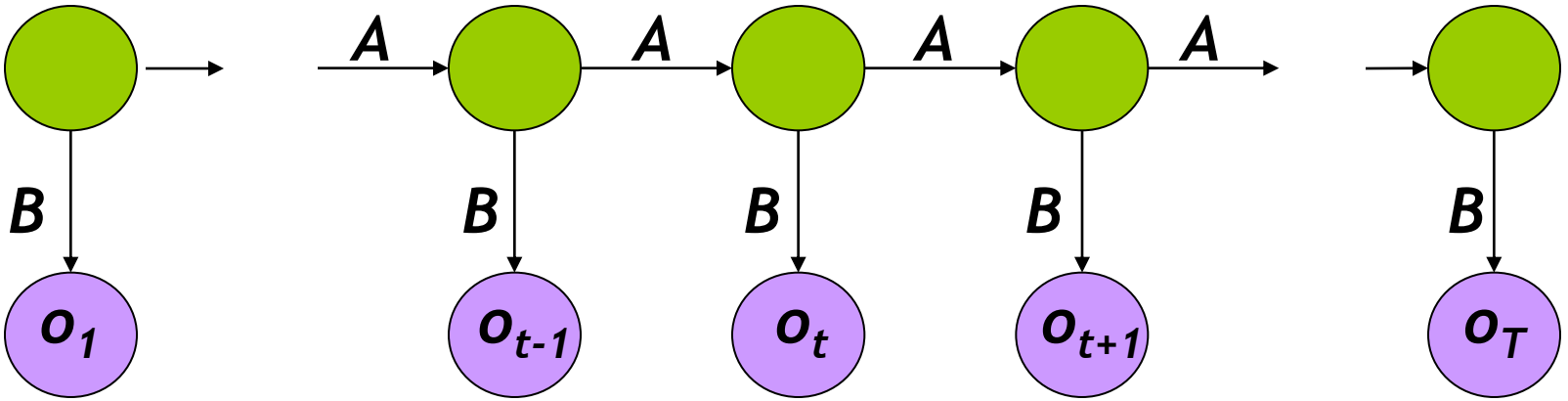
$$p_t(i,j) = \frac{\alpha_i(t) a_{ij} b_{jo_{t+1}} \beta_j(t+1)}{\sum_{m=1\ldots N} \alpha_m(t)\beta_m(t)}$$

Probability of traversing an arc

$$\gamma_i(t) = \sum_{j=1\ldots N} p_t(i,j)$$

Probability of being in state *i*

Parameter Estimation



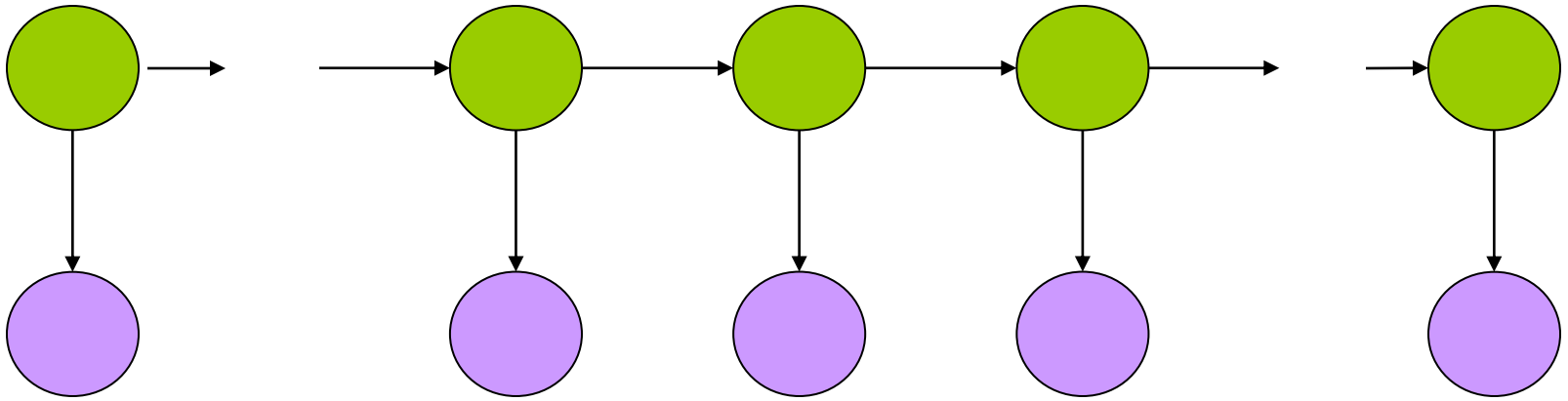
$$\hat{\pi}_i = \gamma_i(1)$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T} p_t(i,j)}{\sum_{t=1}^{T} \gamma_i(t)}$$

$$\hat{b}_{ik} = \frac{\sum_{\{t:o_t=k\}} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_i(t)}$$

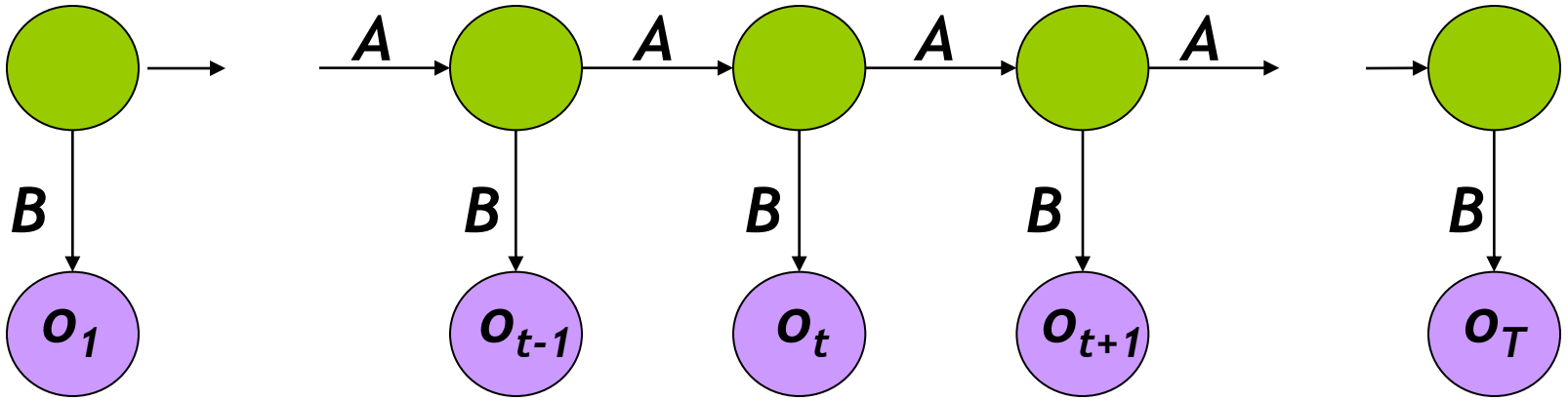
Now we can compute the new estimates of the model parameters.

HMM Applications



- Generating parameters for n-gram models
- Tagging speech
- Speech recognition

The Most Important Thing



We can use the special structure of this model to do a lot of neat math and solve problems that are otherwise not solvable.

HMM Tools (1)

- http://www-a2k.is.tokushima-u.ac.jp/member/kita/NLP/nlp_tools.html
- i unutar toga
- http://www.kanungo.com/software/software.html#umdhmm
- http://people.cs.ubc.ca/~murphyk/Software/HMM/hmm.html
- http://metameme.sdsc.edu/

HMM Tools (2)

- http://ccg.cc.gt.atl.ga.us/~anjiro/gt2k/html_book/node18.html
- http://www.cs.ualberta.ca/~sergey/MVNHMM/index.html
- http://cslu.cse.ogi.edu/HLTsurvey/ch1node7.html
- http://htk.eng.cam.ac.uk/

Zaključak

- Potrebno je imati osnovnu ideju što HMM mogu, kakve podatke traže, koja su ograničenja o kojima treba voditi računa
- Kada znamo što nam treba i zašto nam treba, tada biramo primjereni alat (toolkit) za svoj posao
- Dublji uvid u HMM dobit koristeći očigledne pogreške u našem radu s alatom

Obrada prirodnog jezika

Dr.sc. Mladen Sokele

Modeliranje indikatora obrade prirodnog jezika

Ak.g. 2012./2013.

Modeli rasta – određivanje optimalnih parametara

Modeliranje vremenskih nizova:

$$y(t) = f(t; a_1, a_2, ..., a_k)$$

***k* slobodnih parametara – minimalno *k* poznatih točaka: (t_i, N(t_i))**

1. Poznato točno *k* eksperimentalnih točaka

Sustav jednadžbi: $$N(t_i) - f(t_i; a_1, a_2, ..., a_k) = 0, \quad i = 1, ..., k$$

Za složene modele rasta – u pravilu nelinearan i eksplicitno nerješiv

Potrebno koristiti numeričke metode (npr. Newtonova iterativna metoda) ili *solver* aplikacije.

Modeli rasta

2. Poznato *n*, *n* > *k* eksperimentalnih točaka

Metoda najmanjih kvadrata

Vrijednosti parametara modela određuju se tako da je suma kvadrata razlika između mjerenih vrijednosti i izračunatih vrijednosti (s pomoću modela) minimalna:

$$S = \sum_{i=1}^{n} w_i \cdot \left[f(t_i; a_1, a_2, \ldots, a_k) - N(t_i) \right]^2$$

gdje su w_i težine. Za w_i = 1 => obična metoda najmanjih kvadrata (OLS).

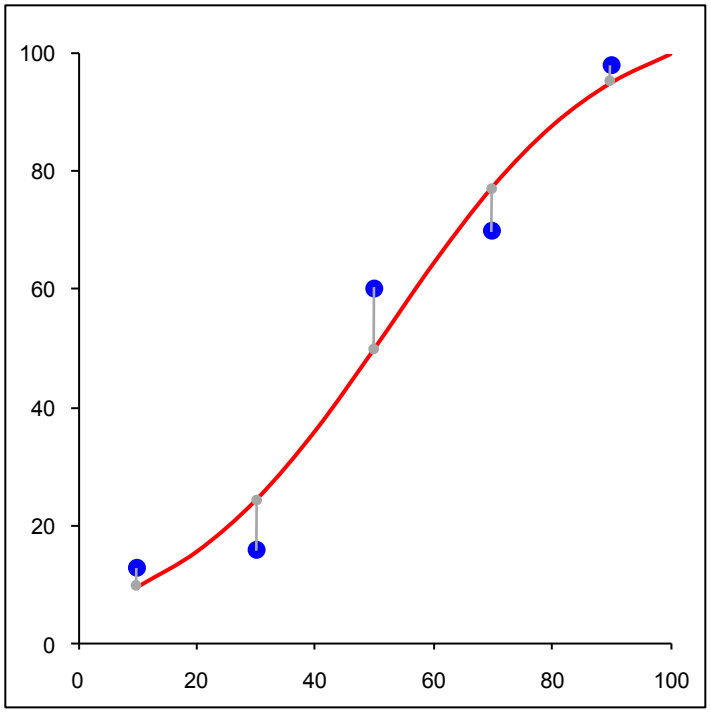
Za minimizaciju sume *S* i time određivanje optimalnih vrijednosti parametara a_i mogu se koristiti *solver* aplikacije.

Analitički izrazi za optimalne parametre dobivaju se iz sustava jednadžbi:

$$\frac{\partial S}{\partial a_j} = 0, \quad j = 1, \ldots, k$$

Modeli rasta – određivanje optimalnih parametara

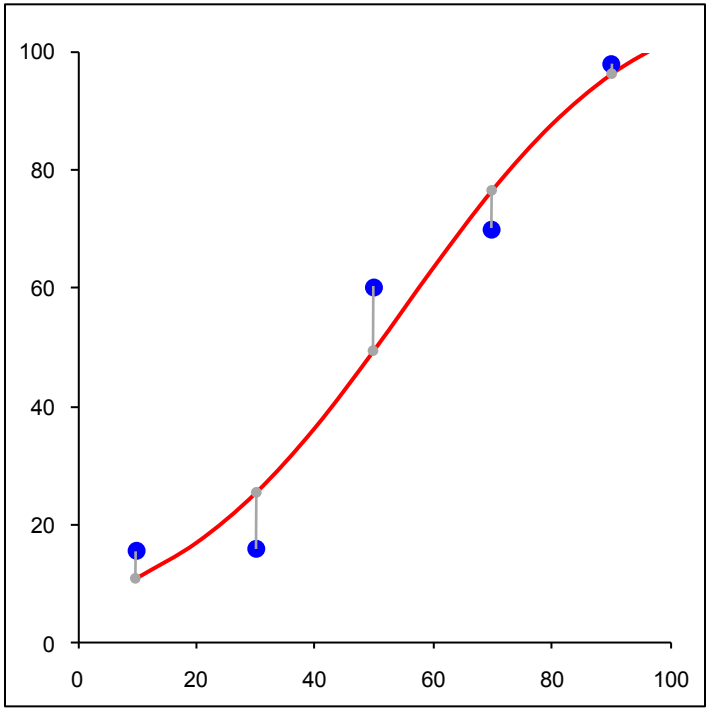
Obična metoda najmanjih kvadrata (MNK)
Ordinary least squares method (OLS)



$$\min_{\{a_1 \ldots a_k\}} \sum_{i=1}^{n} \left[f(t_i; a_1, \ldots, a_k) - N(t_i) \right]^2$$

Određivanje parametara a_j je provedeno statističkim izglađivanje pogreške sadržane u (izmjerenoj) zavisnoj varijabli $N(t)$

Metoda najmanjih kvadrata s težinama
Weighted least squares method

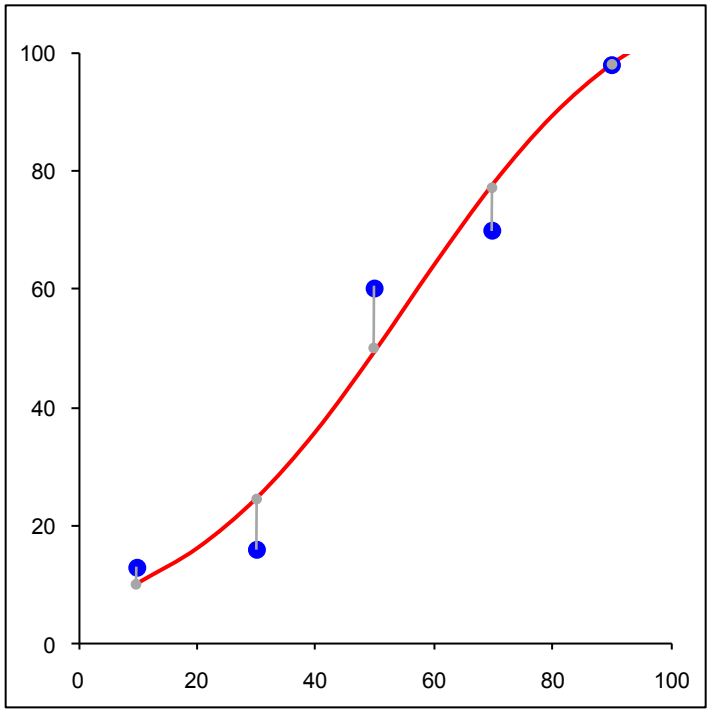


$$\min_{\{a_1 \ldots a_k\}} \sum_{i=1}^{n} w_i \cdot \left[f(t_i; a_1, \ldots, a_k) - N(t_i) \right]^2$$

Ovisno o težinama w_i model ima manje odstupanje za odabrane točke (npr. za točke u bližoj prošlosti)

Modeli rasta – određivanje optimalnih parametara

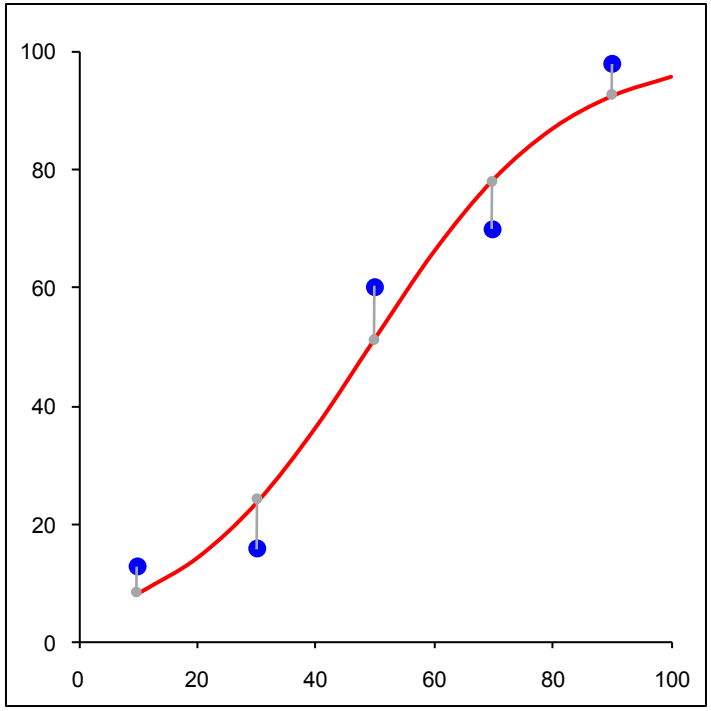
Metoda najmanjih kvadrata s modelom kroz fiksnu točku (t_f, $N(t_f)$)



$$\min_{\{a_1 \ldots a_{k-1}\}} \sum_{i=1, i \neq f}^{n} \left[f(t_i; a_1, \ldots, a_{k-1}; t_f, N(t_f)) - N(t_i) \right]^2$$

Model prolazi kroz fiksnu točku (t_f, $N(t_f)$), a ostale točke se koriste za statističko izglađivanje pogreške sadržane u zavisnoj varijabli $N(t)$

Metoda najmanjih kvadrata s fiksiranom vrijednošću parametra a_k

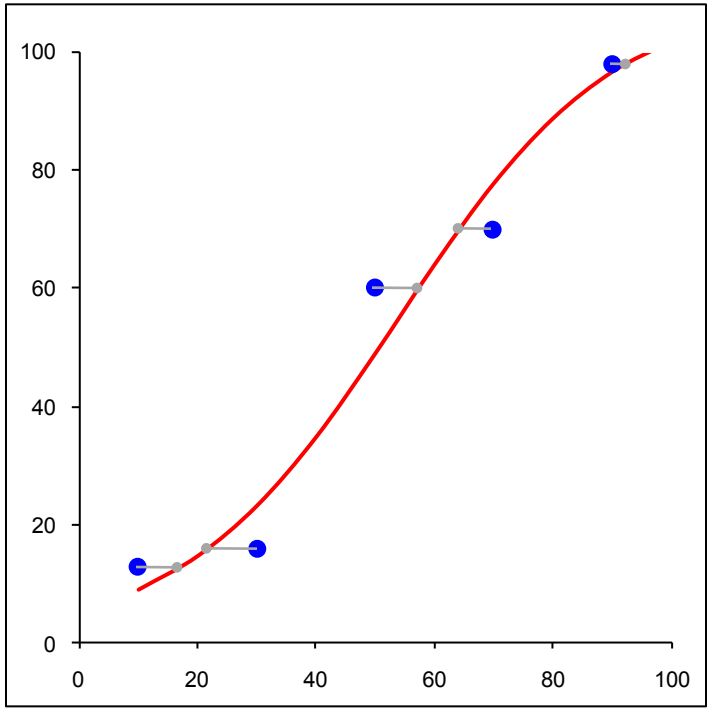


$$\min_{\{a_1 \ldots a_{k-1}\}} \sum_{i=1}^{n} \left[f(t_i; a_1, \ldots, a_k) - N(t_i) \right]^2$$

Vrijednost parametra a_k ulazi izravno u model. Ostali parametri se određuju s ciljem statističkog izglađivanja pogreške sadržane u (izmjerenoj) zavisnoj varijabli $N(t)$

Modeli rasta – određivanje optimalnih parametara

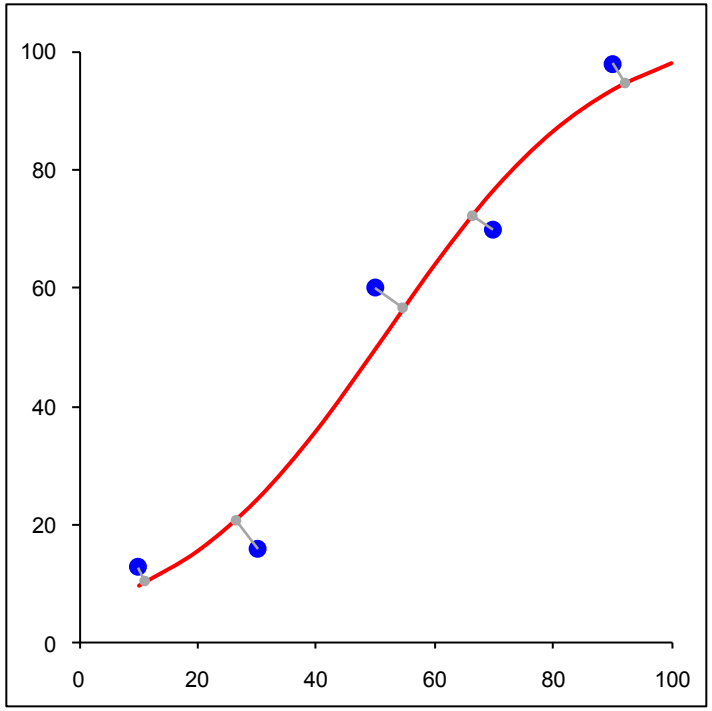
Metoda najmanjih kvadrata s inverznim modelom



$$\min_{\{a_1 \ldots a_k\}} \sum_{i=1}^{n} \left[f^{-1}(N(t_i); a_1, \ldots, a_k) - t_i \right]^2$$

Statističko izglađivanje pogreške sadržane u (izmjerenoj) nezavisnoj varijabli *t*

Metoda potpunih najmanjih kvadrata
Total least squares method (TLS)



Statističko izglađivanje pogreške sadržane u nezavisnoj *t* i zavisnoj varijabli *N*(*t*)

Vidjeti npr:

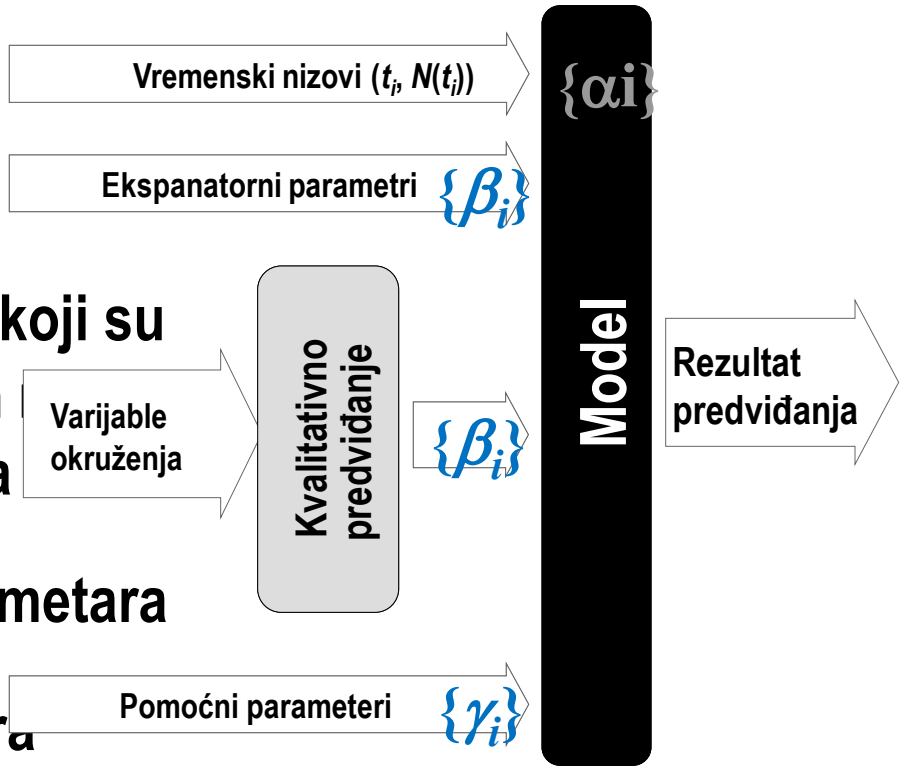
http://www.mathos.hr/rp2/seminar6.pdf
http://en.wikipedia.org/wiki/Total_least_squares

Modeli rasta – prilagodba za predviđanje

Mogućnost prihvata rezultata predviđanja temeljem procjena i izravno dobivenih eksplanatornih parametara

Pomoćni parametri – prilagodba specifičnim praktičnim potrebama

$$Y(t) = f(t; \{\alpha_i\}, \{\beta_i\}, \{\gamma_i\})$$



$\{\alpha_i\}$ – skup parametara modela koji su rezultat modeliranja vremenskih
npr. optimalne vrijednosti prema

$\{\beta_i\}$ – skup eksplanatornih parametara

$\{\gamma_i\}$ – skup pomoćnih parametara

Prilagodba modela rasta – fiksne točke

Poznato n, $n > k$ eksperimentalnih točaka, model prolazi kroz jednu zadanu (fiksnu) točku $(t_f, N(t_f))$

Vrijednost jednog parametra modela npr. a_k može se izračunati iz jednadžbe:

$$N(t_f) - f(t_f; a_1, a_2, \ldots, a_k) = 0$$

Modificirani model ima jedan parametar manje za odrediti metodom najmanjih kvadrata (MNK), te poprima oblik:

$$f(t_i; a_1, \ldots, a_k) \Rightarrow f(t_i; a_1, \ldots, a_{k-1}; t_f, N(t_f))$$

Primjer uporabe:

- **Model u sebi "uključuje" vrijednost zadnje izmjerene točke $(t_e, N(t_e))$**

Modeli rasta u praktičnoj primjeni imaju do dvije fiksne točke – npr. početnu i završnu $(t_s, N(t_s))$ i $(t_e, N(t_e))$ ⇨ do dva slobodna parametra manje za MNK.

Prilagodba modela rasta – težine

Pretpostavka:

Točke u bliskoj prošlosti su važnije za predviđanje

- Metoda najmanjih kvadrata s težinama:

$$S = \sum_{i=1}^{n} w_i \cdot \left[f(t_i; a_1, a_2, \ldots, a_k) - N(t_i) \right]^2$$

gdje su težine w_i veće za točke iz bliže prošlosti (t.j. za zadnje poznate točke).

Na primjer: $w_i = \frac{1}{q^{n-i}}, \quad q > 1$

Npr. za $q = 2$, težine su:

$w_n = 1, \quad w_{n-1} = 0.5, \quad w_{n-2} = 0.25, \quad w_{n-3} = 0.125, \ldots$

Modeli rasta – mjera uspješnosti modeliranja

Koeficijent korelacije *r* – mjera statističke povezanosti (jakost veze) izmjerenih i modeliranih vrijednosti:

$$r = \sqrt{1 - \frac{\sum_{i=1}^{n} [f(t_i) - N(t_i)]^2}{\sum_{i=1}^{n} [\overline{N} - N(t_i)]^2}}; \quad 0 \le r \le 1$$

MNK
Suma S

Gdje su:

- $N(t_i)$ **eksperimentalne (izmjerene) vrijednosti za broj korisnika u vremenskom trenutku** t_i
- $f(t_i)$ **modelirane vrijednosti za broj korisnika u vremenskom trenutku** t_i
- $\overline{N}$ **prosjek izmjerene vrijednosti za broj korisnika:** $\overline{N} = \frac{1}{n} \sum_{i=1}^{n} N(t_i)$

Što je vrijednost *r* bliža 1, to je veza jača t.j. modeliranje uspješnije

Mjere uspješnosti predviđanja

Usporedba broja korisnika dobivenim iz modela $f(t_i)$ s izmjerenim vrijednostima $N(t_i)$ s ciljem ocjene efikasnosti odabranog modela

1. **Predviđanje koje je provedeno u prošlosti - nakon što su postale poznate sve vrijednosti koje su bile predmet predviđanja (t.j. nakon što je "budućnost postala sadašnjost"):**

   → **Prosječno apsolutno odstupanje (*Mean Absolute Error*) *MAE:***

   $$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|f(t_i) - N(t_i)\right|$$

   → **Prosječno apsolutno postotno odstupanje (*Mean Absolute Percentage Error*) *MAPE* [%]:**

   $$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|f(t_i) - N(t_i)\right|}{N(t_i)}$$

   → **Korijen prosječnih kvadrata odstupanja (*Root Mean Squared Error*) *RMSE:***

   $$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[f(t_i) - N(t_i)\right]^2}$$

   MNK Suma S

Mjere uspješnosti predviđanja

2. **Predviđanja temeljem ograničenog skupa poznatih točaka (t_i, $N(t_i)$), i=1,2,...,m; $m < n$ i usporedba sa svim poznatim točkama (t_i, $N(t_i)$), i=1,2,...,n :**

   → **<u>Na primjer:</u> Odrediti parametre metodom najmanjih kvadrata iz m=n-1 poznate točke**

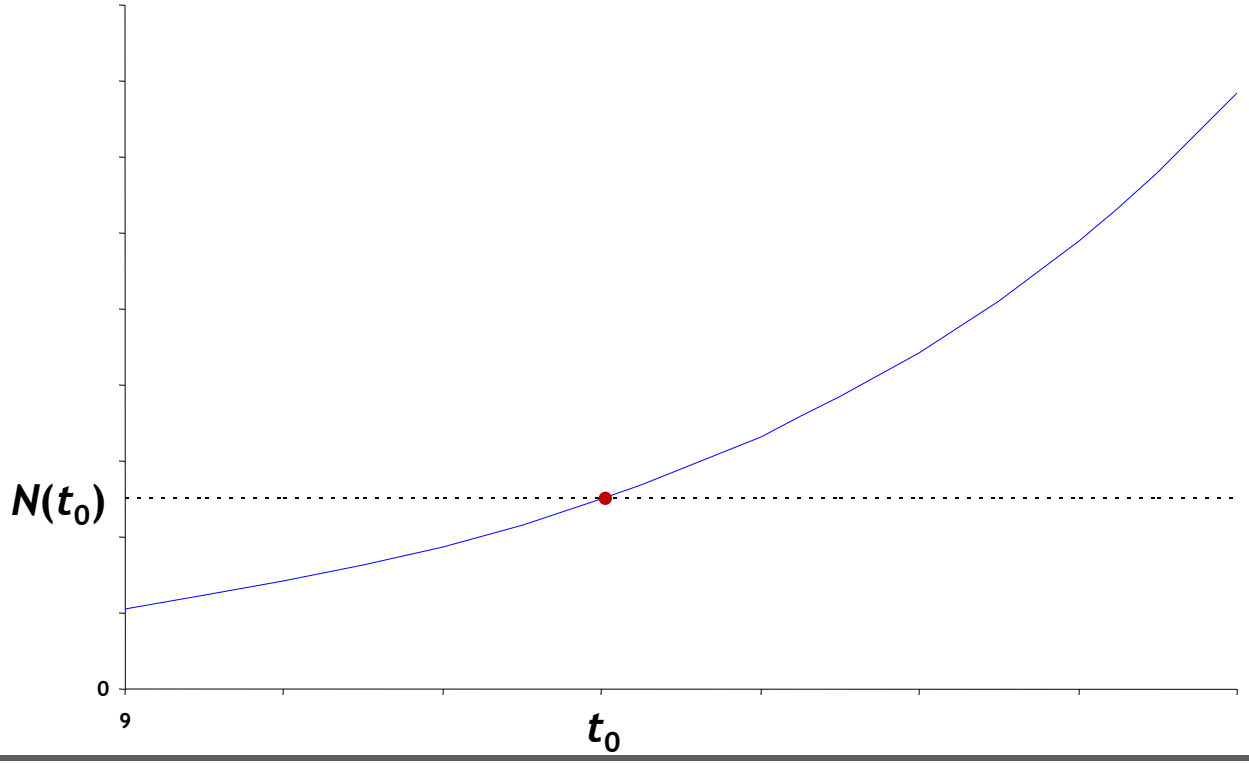
   Postupak:

   - **Odrediti parametre $\{\alpha^*_i\}$ metodom najmanjih kvadrata iz n-1 poznate točke (t_i, $N(t_i)$), i=1,2,...,n-1 za sve modele koji se ocjenjuju;**
   - **Na temelju $\{\alpha^*_i\}$, $\{\beta_i\}$ i $\{\gamma_i\}$ izračunati $f(t_n)$ putem svakog modela, te dobivene vrijednosti za $f(t_n)$ određivanjem *MAE* / *MAPE* / *RMSE* usporediti s poznatom vrijednošću za $N(t_n)$;**
   - **Odabrati model koji daje najmanji *MAE* / *MAPE* / *RMSE*;**
   - **Za odabrani model odrediti parametre $\{\alpha_i\}$ putem MNK iz svih n poznatih točaka, te uz iste $\{\beta_i\}$ i $\{\gamma_i\}$ izvršiti predviđanje za traženi vremenski interval u (stvarnoj) budućnosti.**

Modeli

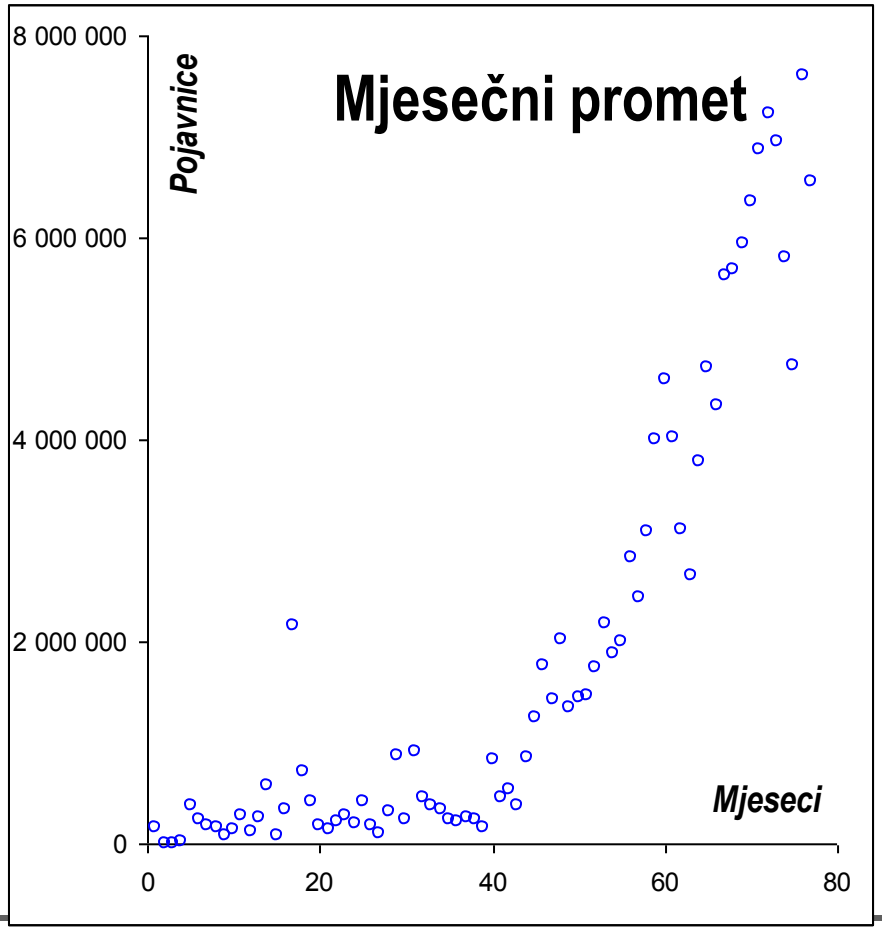
Eksponencijalni: $f(x) = e^{\frac{x-\Delta t}{\tau}} \Rightarrow f(\Delta t) = 1; \quad \frac{f(x+\tau)}{f(x)} = e$

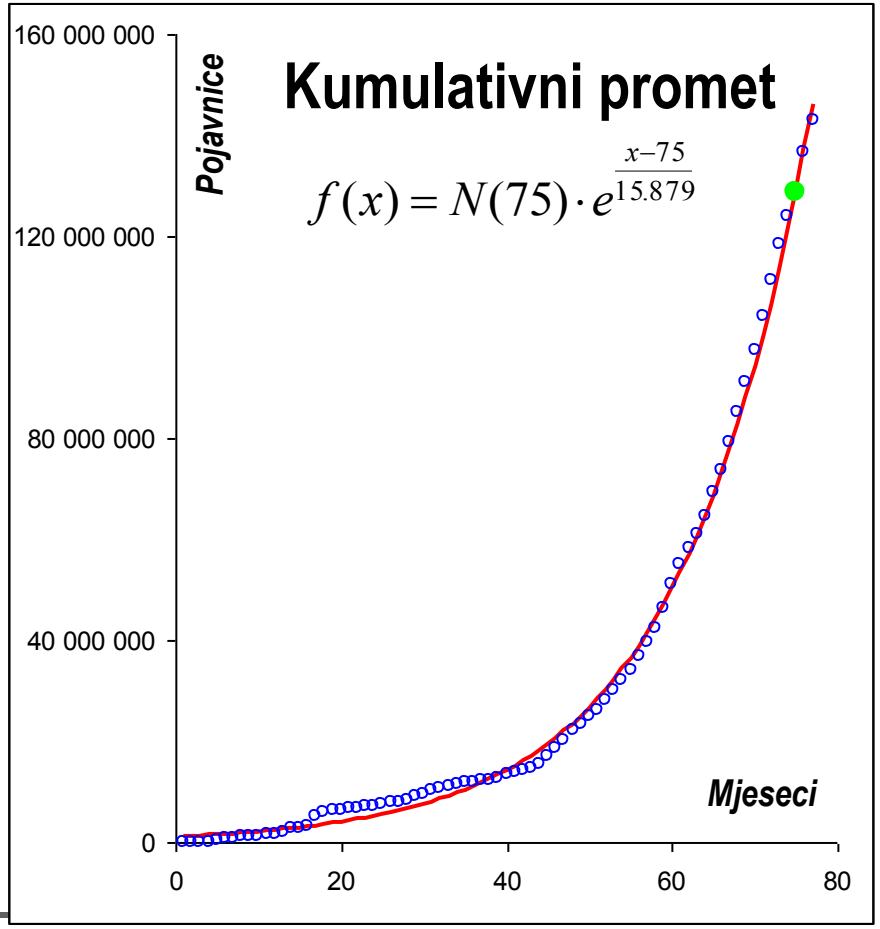
Eksponencijalni s fiksnom točkom: $f(x) = N(t_0) \cdot e^{\frac{x-t_0}{\tau}} \Rightarrow f(t_0) = N(t_0); \quad \frac{f(x+\tau)}{f(x)} = e$



Modeli

Primjer eksponencijalnog modela: početak vremenskog intervala rasta ukupnog obrađenog korpusa

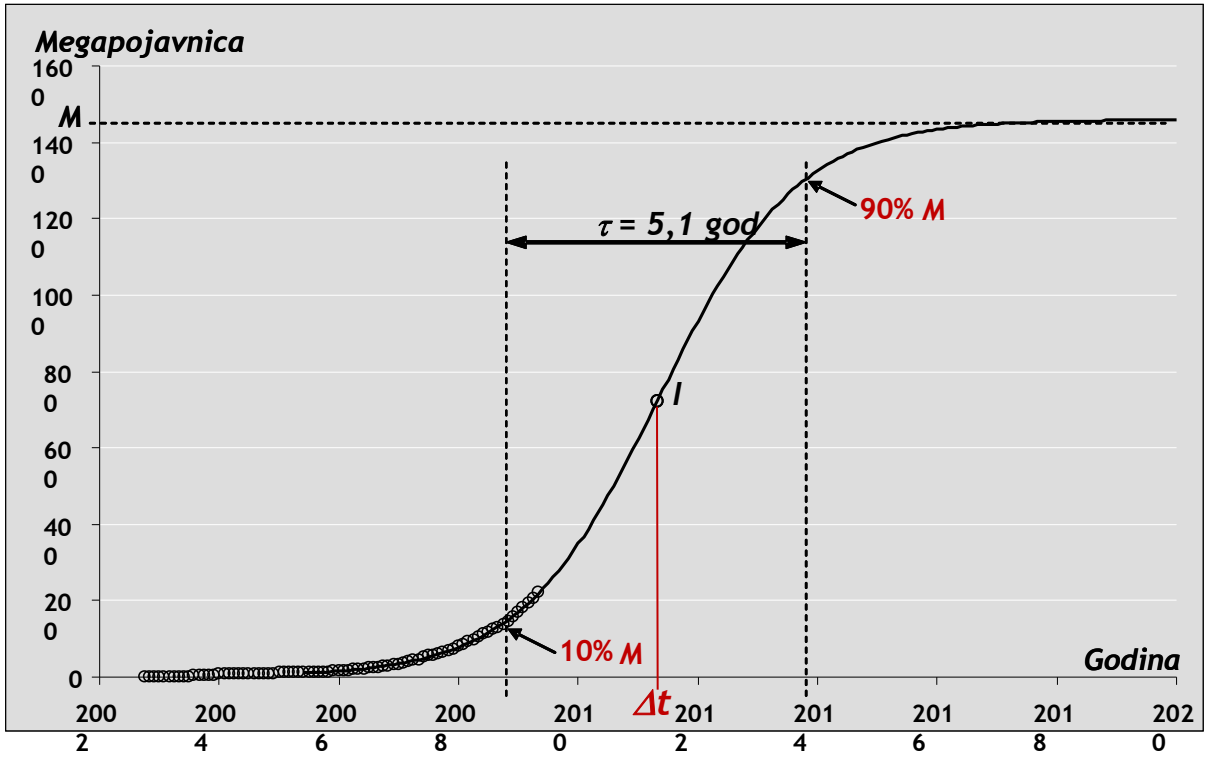




Modeli

Logistički: $f(x) = \dfrac{M}{1 + e^{-k\frac{x-\Delta t}{\tau}}} \Rightarrow \lim_{x\to-\infty} f(x) = 0; \quad \lim_{x\to+\infty} f(x) = M; \quad f(\Delta t) = \dfrac{M}{2}$

$$k = 2\ln\left(\frac{1}{u} - 1\right); \quad f(\Delta t - \frac{\tau}{2}) = uM; \quad f(\Delta t + \frac{\tau}{2}) = (1-u)M$$



u = 10% ⇨ k = 4.3944
τ = 5,1
Δt = 2011.3 (u travnju 2012.)

Modeli

Bassov model:

Ispravlja nedostatak logističkog modela - omogućuje modeliranje početka životnog vijeka usluga

Diferencijalni oblik: $$\frac{dB(t)}{dt} = qB(t)\left(1-\frac{B(t)}{M}\right) + p\left(M - B(t)\right)$$

Effect of imitators (Logistic growth) — Effect of innovators

Analitički oblik: $$B(t; M, p, q, t_s) = B(t) = M\,\frac{1-e^{-(p+q)(t-t_s)}}{1+\frac{q}{p}e^{-(p+q)(t-t_s)}}$$

M - asimptota (ukupni kapacitet tržišta)

p - koeficijent inovacije, $p > 0$

q - koeficijent imitacije, $q \geq 0$

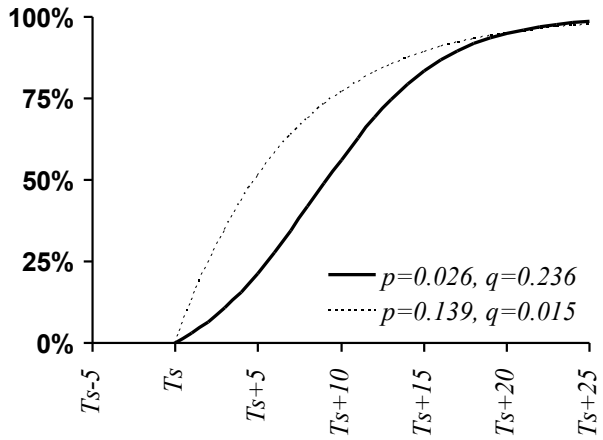
t_s - vrijeme kada je usluga lansirana na tržište, $B(t_s)=0$

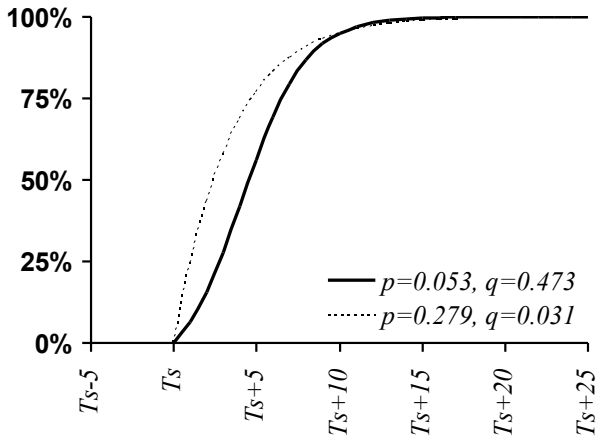
⇨ Ima 4 slobodna parametra

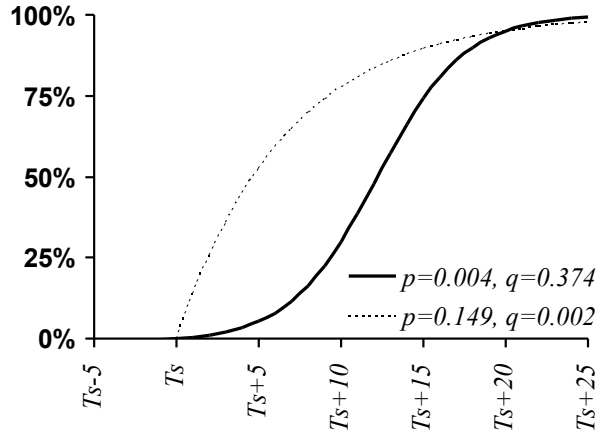
⇨ Krivulja je identična logističkoj ali je pomaknuta dolje po ordinati

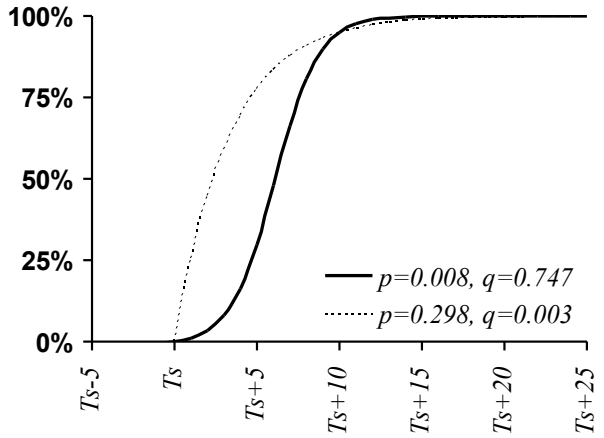
Modeli

Ovisnost oblika S-krivulje o parametrima Bassovog modela:









Modeli

Eksponencijalni saturacijski:

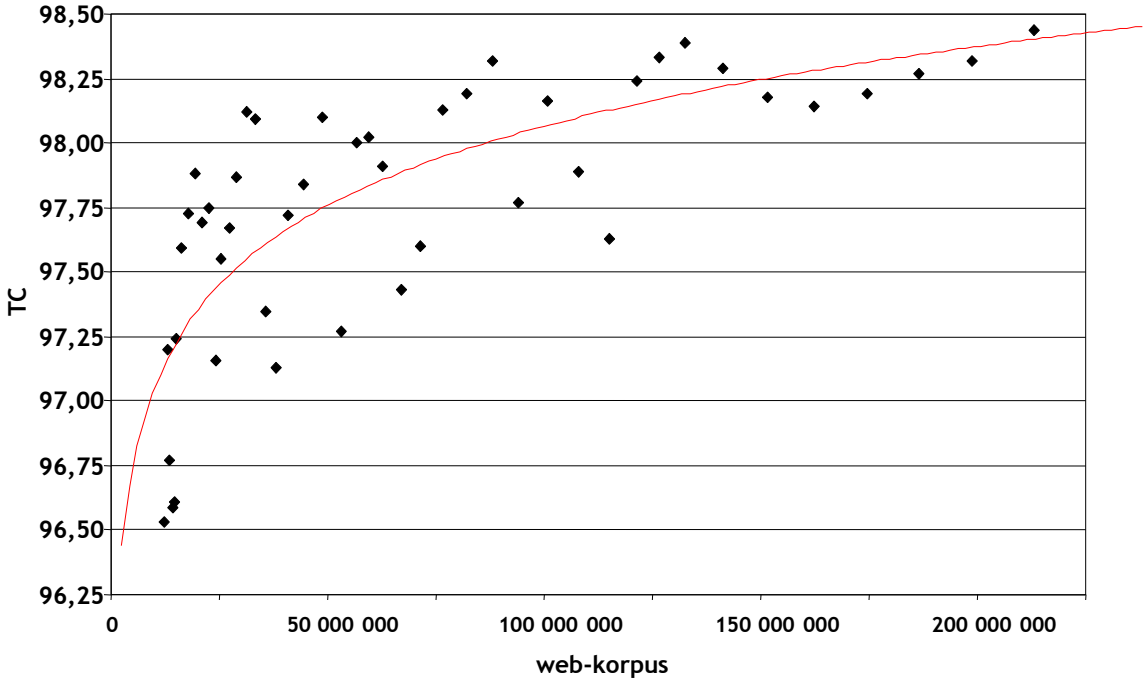
$$f(x) = M\left[1 - e^{-\frac{x-\Delta t}{\tau}}\right] \quad \Rightarrow \quad f(\Delta t) = 0; \quad \lim_{x\to\infty} f(x) = M$$

Primjer: Modeliranje indeksa prekrivanja teksta *TC*

Neovisna varijabla nije vrijeme već obim korpusa

$$TC = 100 \cdot \left[1 - \frac{BrojNeprepoznatihPojavnica}{OpsegTeksta}\right]$$

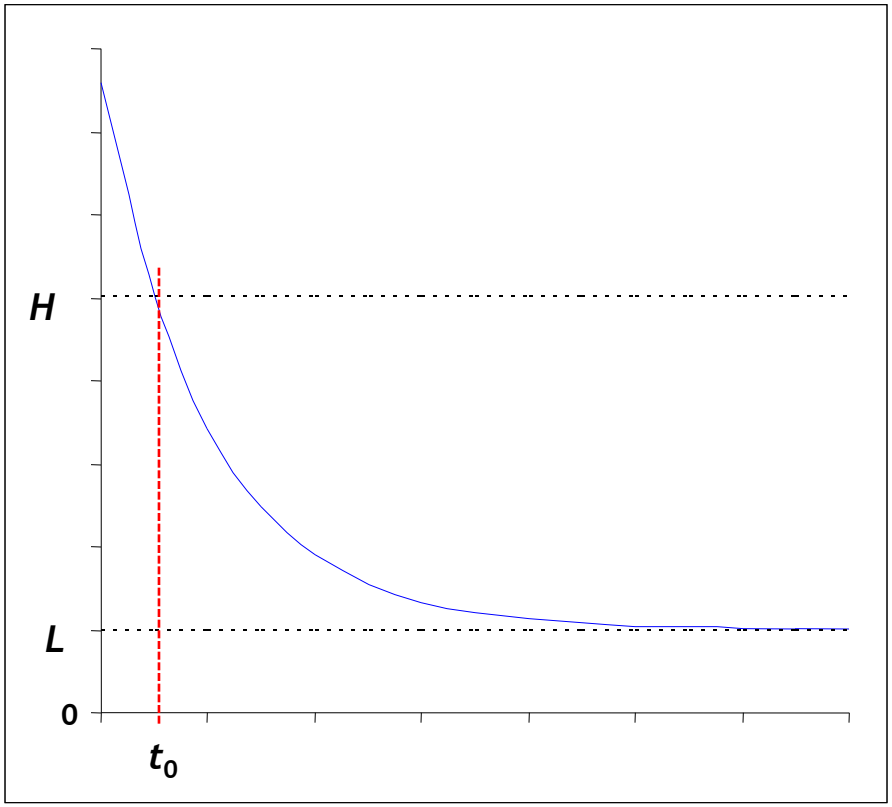
***M* = 100**



Modeli

Eksponencijalna asocijacija*:

$$f(x) = L + (H - L) \cdot e^{-\frac{x-\Delta t}{\tau}} \quad \Rightarrow \quad f(\Delta t) = H; \quad \lim_{x \to \infty} f(x) = L$$



*CurveExpert Software http://www.curveexpert.net/

Modeli

Primjer: Modeliranje indeksa učenja *LI*

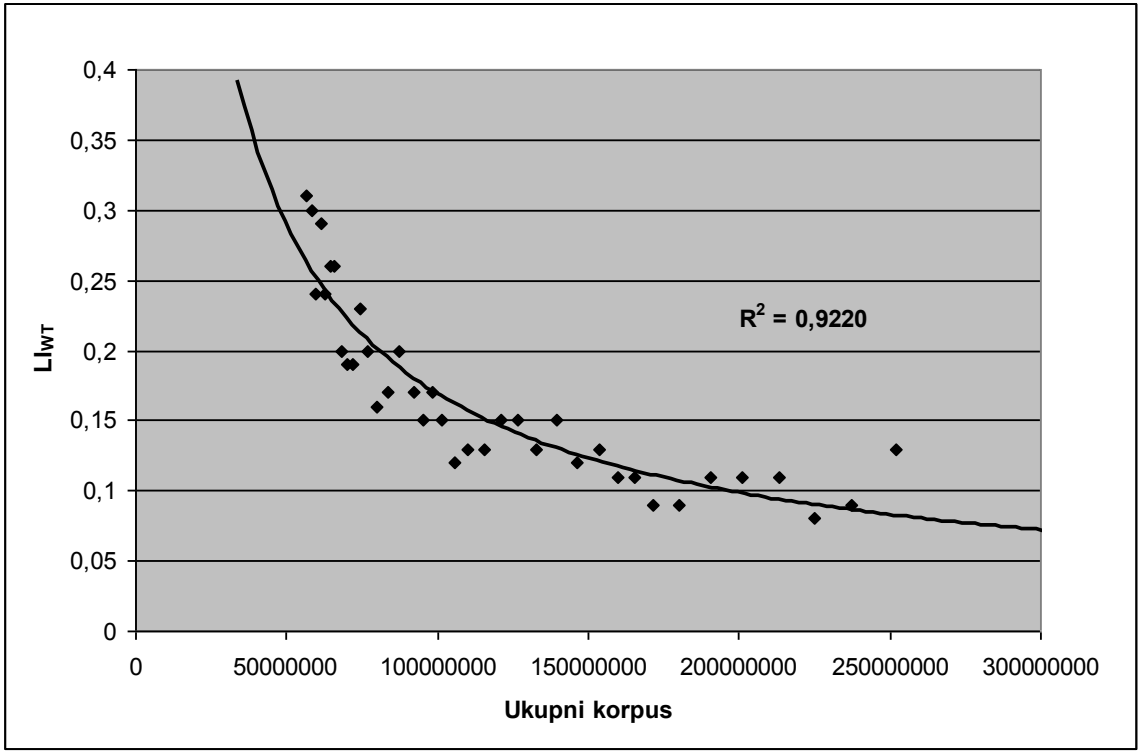
$$LI = 100 \cdot \frac{BrojNaučroihRazlicnica}{OpsegTeksta}$$

Neovisna varijabla nije vrijeme već obim korpusa

***H* = 1**

***L* = *a*, *a* > 0**

$$li(t) = a + (1-a) \cdot e^{-\frac{t-T}{\tau}}$$

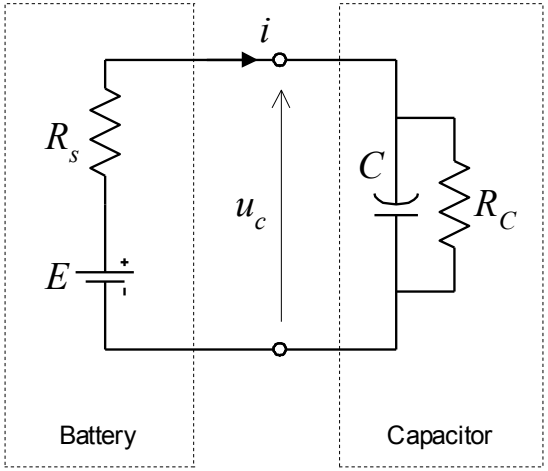


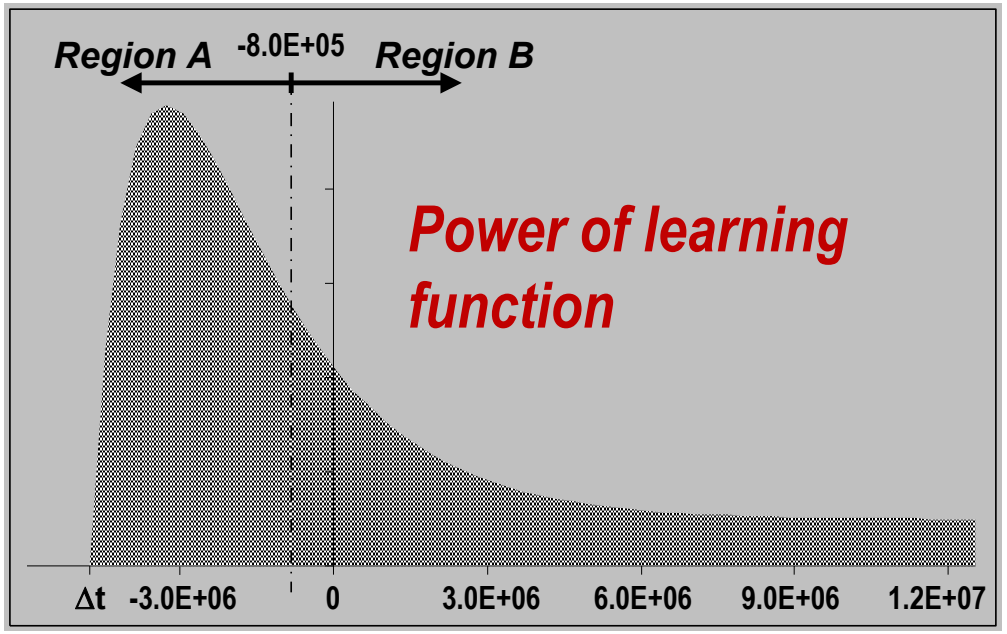
Kognoelektrička analogija

$$i(t) = k \cdot \left[a' + (1 - a') \cdot e^{-\frac{t}{\tau}} \right]$$ ~ indeks učenja *LI*

$$u_C(t) = k \cdot b' \cdot \left(1 - e^{-\frac{t}{\tau}} \right)$$ ~ indeks prekrivanja teksta *TC*

$$PoL = \left[a + (100 - a) \cdot e^{-\frac{PPC - \Delta t}{\tau}} \right] \cdot b \cdot \left(1 - e^{-\frac{PPC - \Delta t}{\tau}} \right)$$

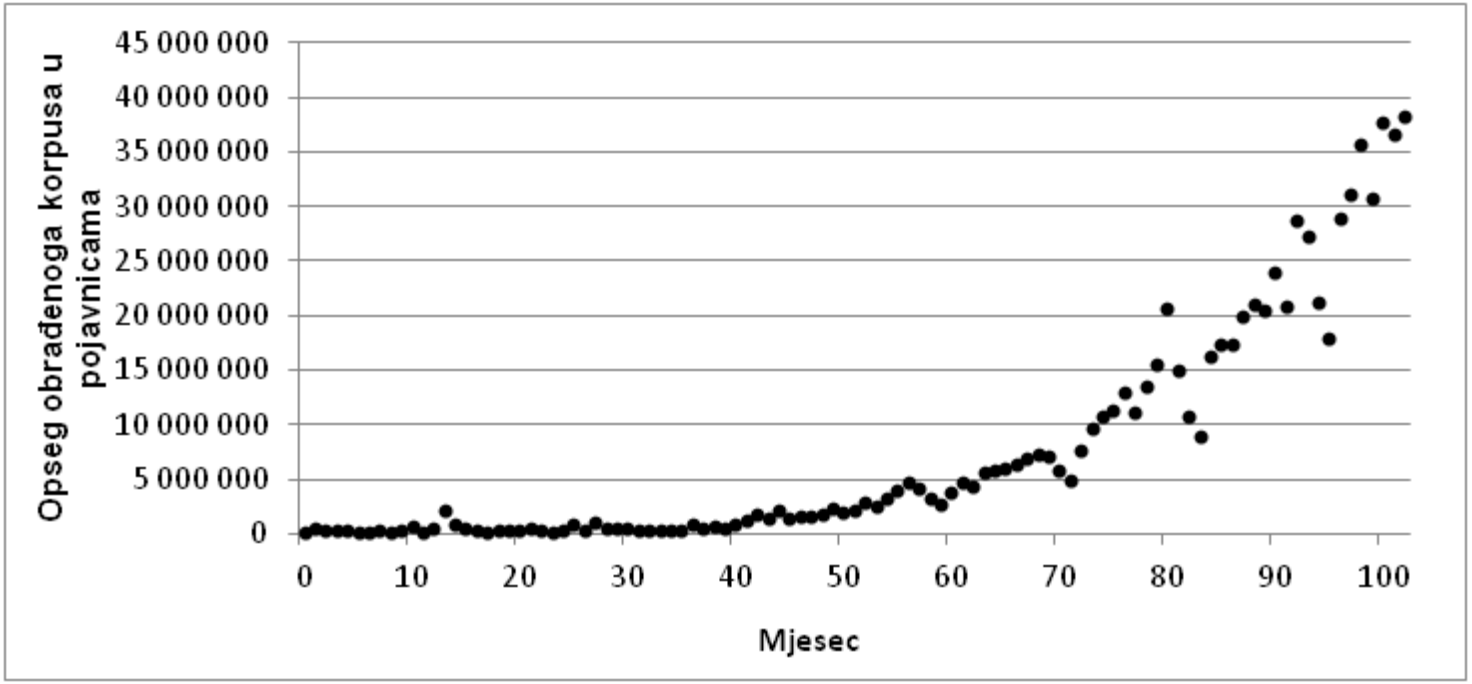




Electrical Equivalence of Word Learning

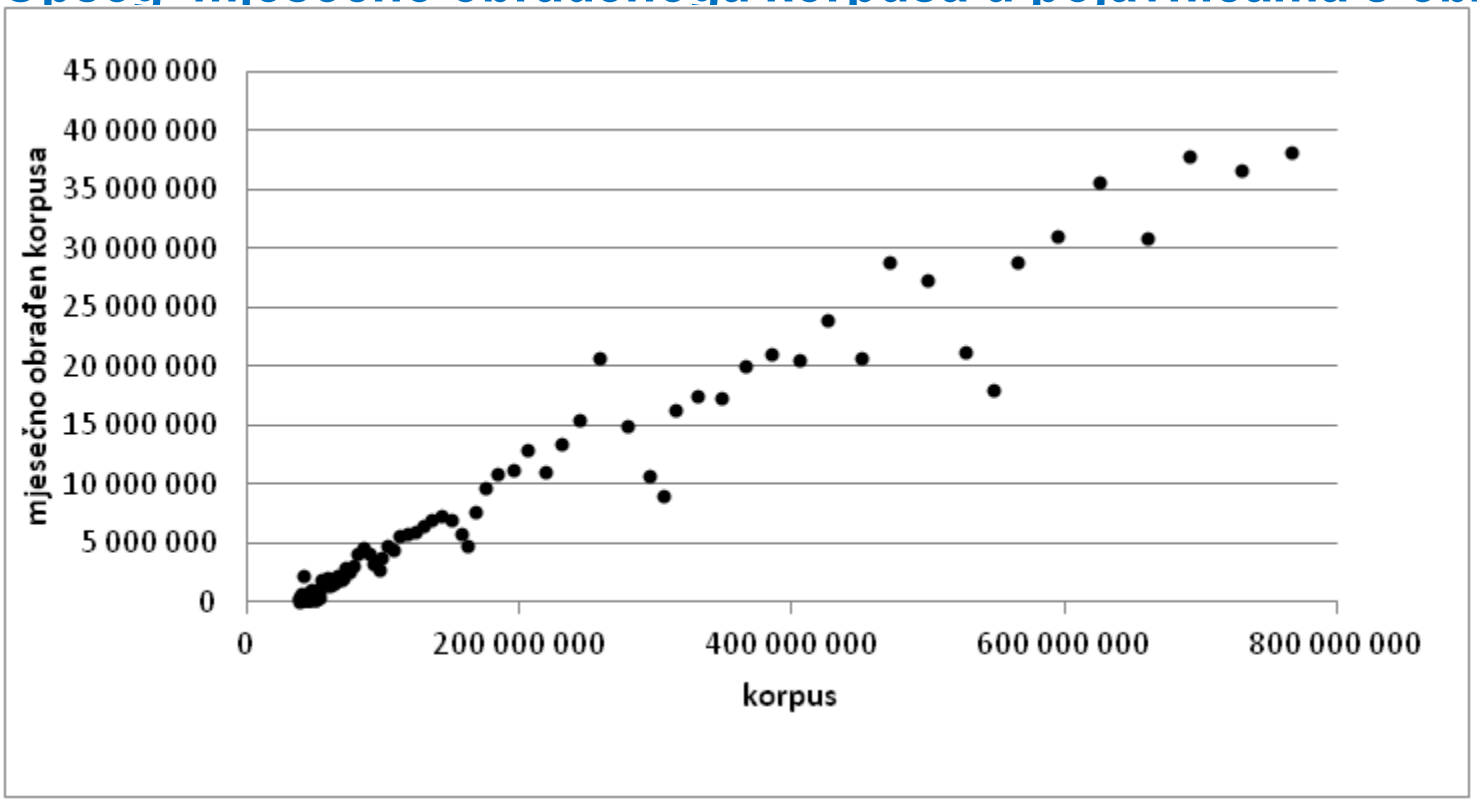
*Modeliranje prometa usluge strojne provjere pravopisa**

Mjesečno obrađen korpus:

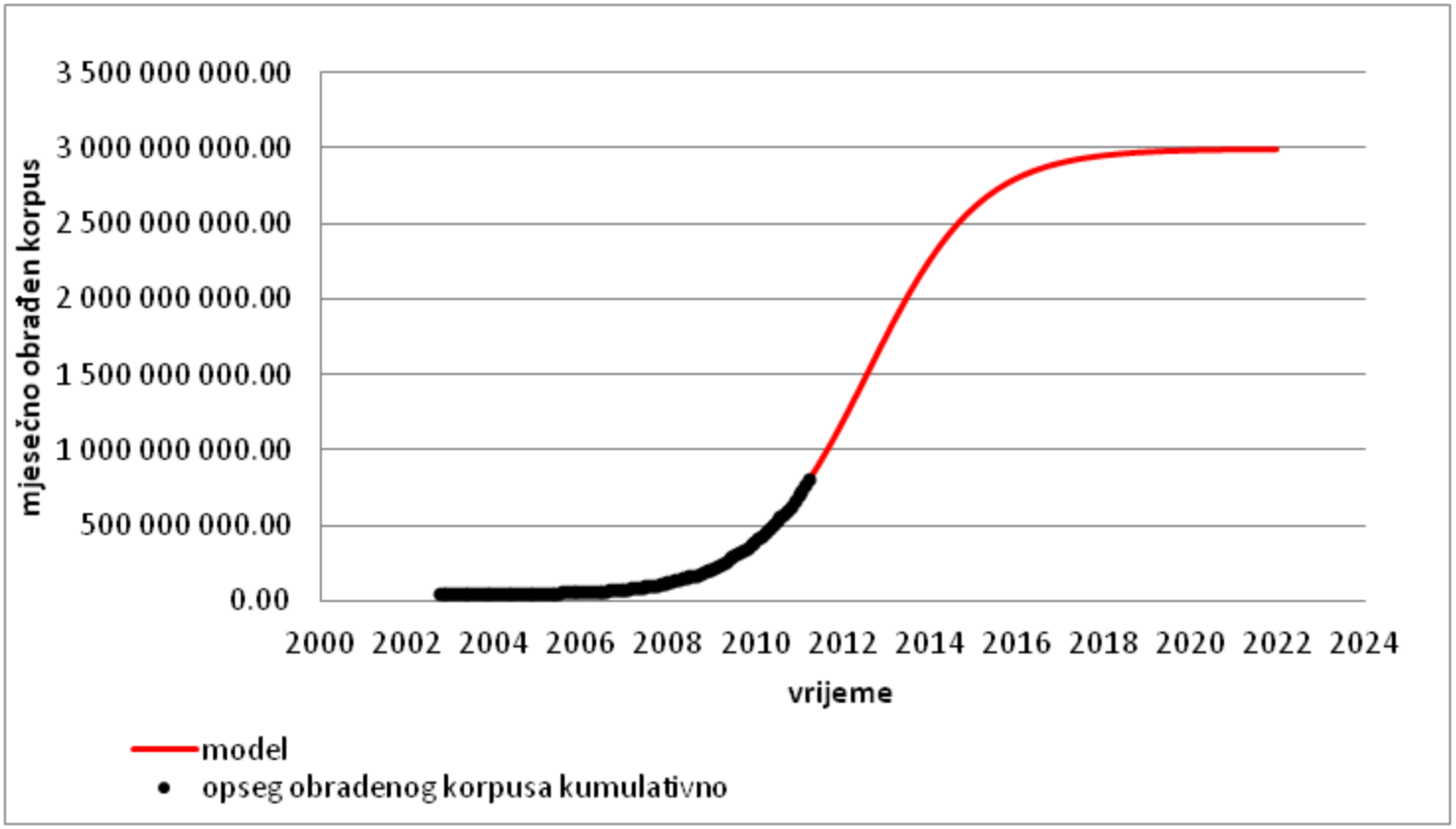


** Kristina Mrvelj, diplomski rad, 2012.*

Opseg mjesečno obrađenoga korpusa u pojavnicama s obzirom na ve



Modeliranje veličine korpusa (kumulativ mjesečno obrađenog korpusa)



Bassov model s fiksnom vrijednošću zadnje točke

Broj novonaučenih različnica