Obrada prirodnih jezika

Elektrotehnički fakultet Univerziteta u Beogradu

Master akademske studije

modul Softversko inženjerstvo

2022/2023

Zadaci u obradi prirodnih jezika

Vuk Batanović, Elektrotehnički fakultet Univerziteta u Beogradu

Hijerarhija NLP zadataka

- Proučavanje jezika se može okvirno podeliti na šest grupa/podoblasti:
  - *Fonetika/fonologija* – proučavanje lingvističkih zvukova
  - *Morfologija* – proučavanje oblika reči
  - *Sintaksa* – proučavanje strukturnih odnosa između reči
  - *Semantika* – proučavanje značenja reči (*leksička semantika*) i sekvenca reči (*kompoziciona semantika*)
  - *Pragmatika* – proučavanje načina na koji se jezik koristi u komunikaciji
  - *Diskurs* – proučavanje lingvističkih jedinica širih od pojedinačnog iskaza

Prepoznavanje govora

- Zadatak – pretvoriti audio signal koji sadrži govor na prirodnom jeziku u tekstualni oblik
- Problem na raskršću NLP-a u užem smislu, koji je pre svega fokusiran na obradu tekstualnih podataka, i šire oblasti obrade signala
- Rešava se primenom modela sekvenci
  - Skriveni Markovljevi lanci (engl. *Hidden Markov Models - HMM*)
  - Uslovna nasumična polja (engl. *Conditional Random Fields – CRF*)
  - Neuralne mreže

Segmentacija teksta na rečenice

- Obično prvi deo pipeline-a u procesiranju tekstualnih podataka
- Pre segmentacije na rečenice potrebno je u nekim slučajevima uraditi filtriranje sadržaja
  - Uklanjanje XML/HTML tagova, *boilerplate* teksta, itd.
- Obično se radi heuristički, pomoću regularnih izraza – tačnost oko 95%
- Najbolji rezultati se dobijaju primenom nadgledanog mašinskog učenja nad dovoljno velikim korpusom primera (do 99% tačnosti)
- Na mnogim jezicima interpunkcija označava kraj rečenice, tako da se može koristiti za segmentaciju

Segmentacija teksta na rečenice

- Problem – interpunkcija se ponekad javlja u tekstu i kada se ne radi o kraju rečenice
  - Redni brojevi – *25. maj*
  - Decimalna tačka/grupisanje brojeva – *5.46 / 100.000.000,00*
  - Skraćenice – *dr sci. med. / npr.*
  - Inicijali – *reditelj Dž. Dž. Ejbrams*
  - URLovi i IP adrese – *www.etf.bg.ac.rs / 127.0.0.1*
  - Emotikoni – (^.^) / !>_<!
  - Idiosinkratična upotreba zbog brendiranja – *.hack//SIGN*
- Dodatan problem – ponekad se ovi izuzeci mogu javiti i kada se zaista radi o kraju rečenice
  - *Sledeći čas biće održan 11.05.*

Tokenizacija

- Podela teksta na tokene – reči i ostale smislene elemente (brojeve, URLove,…)
  - Interpunkcija se smatra posebnim tokenima (koji se nekad odbacuju)
  - Problem – interpunkcija je nekada integralni deo tokena
    - *Ejpril O'Nil, Žan-Pol Sartr, T-Mobile, M*A*S*H, B-52, 1960-ih, C#,…*
  - Kako razlikovati takve situacije od onih gde interpunkcija nije deo tokena
    - *pruga Beograd—Bar*
- I ovde se često koriste heuristički pristupi zasnovani na regularnim izrazima i ručno pisanim pravilima
  - Najjednostavnija vrsta tokenizacije – korišćenjem blanko znakova (engl. *white-space tokenization*)
    - Neprimenjiva na jezike koji ne koriste razmak za razdvajanje reči – npr. kineski, japanski, tajlandski

Tokenizacija

- Takođe je moguće obučavanje modela mašinskog učenja nad dovoljno velikim korpusom primera
- Tokenizacija je primetno teža kada polazni tekst nije „čist“ – npr. dobijen putem OCR-a
- Segmentacija teksta na rečenice i tokene je preduslov za skoro sve druge vrste obrada teksta – greške nastale u ovim koracima propagiraju na više nivoe obrade i imaju efekta na sve zadatke koji su složeniji

Jezički modeli

- Služe za probabilističko modelovanje jezika
- Pitanje – koja je verovatnoća da se u posmatranom jeziku javi rečenica s?
- Svaka rečenica se tretira kao sekvenca tokena $s = t_1\ t_2 \ldots t_m$, gde je m broj tokena u rečenici
- Sledi da je verovatnoća javljanja rečenice s:

$$P(s) = P(t_1 t_2 \ldots t_{m-1} t_m) = P(t_1)P(t_2|t_1)P(t_3|t_1 t_2) \cdots P(t_m|t_1 t_2 \ldots t_{m-1})$$

$$= \prod_{i=1}^{m} P(t_i|t_1 t_2 \ldots t_{i-1})$$

- Računanje verovatnoća – na osnovu frekvencija javljanja (sekvenca) tokena u velikom korpusu tekstova

Jezički modeli

- Pored pravih tokena iz teksta, ponekad se na početak i kraj rečenica dodaju posebni tokeni $< s >$ i $</s >$
  - Pomoću njih se izražavaju verovatnoće da rečenica počinje ili se završava nekim pravim tokenom
- Problem proređenosti podataka – dugačke sekvence tokena će vrlo verovatno biti viđene samo jednom
- Markovljeva pretpostavka – verovatnoća javljanja svakog tokena zavisi samo od k prethodnih tokena

$$P(t_i|t_1 t_2 \dots t_{i-1}) \approx P(t_i|t_{i-k} t_{i-k+1} \dots t_{i-1})$$

Jezički modeli

- Unigramski jezički modeli
  - Svaki token (tj. *unigram*) se posmatra zasebno, ignorišu se sekvencijalne zavisnosti između tokena

$$P(t_m|t_1 t_2 \dots t_{m-1}) = P(t_m)$$

  - Verovatnoće unigrama se računaju kao njihova relativna frekvencija u velikom korpusu tekstova

$$P(t_x) = \frac{Count(t_x)}{\sum_{t_i \in V} Count(t_i)}$$

Jezički modeli

- Bigramski jezički modeli
  - Sekvencijalne zavisnosti između tokena se ograničavaju na zavisnosti između dva uzastopna tokena (tj. *bigrama*)
    - Iako i ovo predstavlja simplifikaciju, bigramski modeli su realnija aproksimacija od unigramskih

$$P(t_m|t_1 t_2 \dots t_{m-1}) = P(t_m|t_{m-1})$$

  - Verovatnoće bigrama se računaju kao njihova relativna frekvencija u velikom korpusu tekstova

$$P(t_x|t_{x-1}) = \frac{Count(t_{x-1}\, t_x)}{Count(t_{x-1})}$$

Jezički modeli

- Na isti način se mogu kreirati trigramski, 4-gramski,… n-gramski jezički modeli (koriste sekvence od 3, 4,… n uzastopnih tokena)
- Problem n-gramskih jezičkih modela – verovatnoća javljanja sekvenci tokena koje nisu viđene u datom korpusu će biti nula
- Radi suzbijanja ovog efekta koriste se različiti oblici poravnanja modela (engl. *smoothing*)
  - Poravnanje dodeljuje neku malu nenultu verovatnoću i tokenima/sekvencama tokena koje nisu prisutne u korpusu tekstova
  - Laplasovo poravnanje – brojanje sekvenci tokena počinje od jedinice
  - *Good-Turing* poravnanje
  - *Kneser-Ney* poravnanje
- Pored poravnanja koriste se i druge tehnike – interpolacija, *backoff*,…

Jezički modeli

- Perpleksnost (engl. *perplexity*) – intrinzička mera kvaliteta jezičkih modela
  - Inverzna verovatnoći podataka (iz test seta), normalizovana shodno broju reči
  - $PP(t_1 t_2 \dots t_{m-1} t_m) = \sqrt[m]{\frac{1}{P(t_1 t_2 \dots t_{m-1} t_m)}} = \sqrt[m]{\frac{1}{\prod_{i=1}^{m} P(t_i | t_1 t_2 \dots t_{i-1})}}$
  - Niža perpleksnost – bolji jezički model
  - Nije zamena za ekstrinzičku evaluaciju
- Ekstrinzička mera kvaliteta jezičkih modela – performanse na nekom konkretnom NLP zadatku

Jezički modeli

- Omogućavaju predviđanje sledeće reči u sekvenci na osnovu ranijih reči
- Važno za korekciju pravopisa
  - *Uzeo je u ruke debeo put.*
  - *Uzeo je u ruke debeo prut.*
- Važno za redijakritizaciju teksta, u jezicima/pismima gde se koriste slova sa dijakritičkim oznakama (u srpskoj latinici: č, ć, š, đ, ž)
  - *Otvorena je zgrada glavne poste.*
  - *Otvorena je zgrada glavne pošte.*
- Iako su izuzetno korisni u ovim zadacima, ima situacija gde jezički modeli ne mogu da uvek pruže tačan odgovor
  - *Kupio sam najlepse zelje.*

Neuralni jezički modeli

- Pored opisanih klasičnih n-gramskih jezičkih modela, postoje i noviji, neuralni jezički modeli
  - Isprva razmatrane različite arhitekture (rekurentnih) neuralnih mreža – RNN, LSTM
- Savremeni neuralni jezički modeli su zasnovani na varijantama *Transformer* neuralne arhitekture
  - Umesto klasične tokenizacije koriste mikrotokenizaciju
    - BPE (*Byte Pair Encoding*)
    - *Wordpiece*
- Jedan od najpoznatijih predstavnika: GPT *(Generative Pre-trained Transformer)* porodica modela: GPT, GPT 2, GPT 3, ChatGPT, GPT 4
- Najpoznatija biblioteka za ovakve modele: *HuggingFace Transformers*

Neuralni jezički modeli

- Jezičko modelovanje se može koristiti ne samo za predviđanje sledeće reči u sekvenci, nego i bilo koje nedostajuće reči u sekvenci (engl. *Masked Language Models* (MLM))
  - Ovo se koristi pri obučavanju BERT (*Bidirectional Encoder Representations from Transformers*) modela
- *Obrada prirodnih jezika je MASK predmet na master studijama.*
  - Koja reč se krije iza tokena *MASK* ?
- Postojeći deo sekvence reči predstavlja kontekst, na osnovu koga se vrši predviđanje nedostajućeg dela sekvence
- Lako je konstruisati ogroman broj primera ovog problema, korišćenjem velikih neanotiranih korpusa tekstova
  - Spada u paradigmu samonadgledanog učenja

Neuralni jezički modeli

- Iako naoko rudimentarno, učenje predviđanja naredne/nedostajuće reči u sekvenci omogućava neuralnim jezičkim modelima da steknu dosta robusno opšte znanje o jeziku
  - Ako su dovoljno veliki i ako se obučavaju nad ogromnim količinama podataka
    - BERT: 340M parametara, 3,4 milijarde tokena
    - BERTić: 110M parametara, 8 milijardi tokena
    - GPT: 110M parametara, 985M tokena
    - GPT 2: 1.6B parametara, 8M dokumenata
    - GPT 3: 175B parametara, 500 milijardi tokena
    - GPT 4: ? (>1T parametara)

Neuralni jezički modeli

- Ovakvi modeli se nazivaju prethodno obučenim (engl. *pre-trained models*)
- Oni se zatim mogu fino podesiti (engl. *fine-tuning*) za rešavanje konkretnijih NLP problema, uz relativno male količine anotiranih primera
  - U zavisnosti od složenosti problema, za fino podešavanje može biti dovoljno i samo nekoliko hiljada ili čak stotina primera
  - Spada u transferno učenje (engl. *transfer learning*) – znanje iz jednog domena se prenosi na rešavanje problema u drugom domenu
- Ovo je danas dominantna paradigma razvoja najsavremenijih NLP rešenja
  - Ako su prethodno obučeni modeli dostupni za posmatrani jezik
- Najveći neuralni jezički modeli čak pokazuju obećavajuće performanse i u rešavanju nekih NLP problema za koje nisu fino prilagođeni
  - *Zero-shot learning*

Morfološka normalizacija

- U morfološki bogatim jezicima (poput srpskog) jedna ista reč (gledano semantički) se može javiti u velikom broju različitih oblika
  - Promena reči po padežima, rodovima, brojevima, licima
- Računarski sistemi nisu u stanju da razlikuju između različitih pojavnih oblika jedne iste reči i različitih reči
  - U oba slučaja postoji razlikovanje u nizu karaktera koji čine jednu reč
  - Problem – (dramatično) povećanje proređenosti podataka
    - Broj javljanja u skupu podataka za svaku reč koja ima više oblika je daleko manji, pošto se svaki pojavni oblik tretira kao posebna reč
    - Naročito važno kod manjih skupova podataka
- Rešenje – svođenje različitih oblika reči na zajedničku osnovu – morfološka normalizacija

Morfološka normalizacija

- Dve vrste morfoloških promena
  - Flektivna morfologija – odnosi se na različite oblike jedne reči
    - Npr. promena po padežima: *škola, škole, školu, školi, školom,…*
  - Derivaciona morfologija – odnosi se na izvođenje novih reči
    - Izvođenje dodavanjem sufiksa: *škola – školski, školovanje, školarac, školovati*
    - Izvođenje dodavanjem prefiksa: *neprijatelj, predalek, ulov*
    - Izvođenje slaganjem/kompozicijom više reči: *severoistok, ribolov, gasovod*
    - Kombinovano izvođenje: *iskorišćenost, uramljen, odškolovati, vinogradski*
- Prilikom morfoloških promena često dolazi i do glasovnih promena
  - *Zao – zlog, zla, zloba,…* (nepostojano „a“, prelazak „l“ u „o“)
  - *Seljak – seljački, seljaci,…* (palatalizacija, sibilarizacija)

Morfološka normalizacija

- Dva principijelna pristupa morfološkoj normalizaciji
  - Stemovanje – odsecanje krajeva reči, čime se dobija njen *stem*
    - Alati koji sprovode stemovanje se nazivaju *stemeri*
      - Najpopularniji stemer za engleski jezik – Porterov stemer
  - Lematizacija – zamenjivanje svih oblika neke reči njenom *lemom* – osnovnim rečničkim oblikom
    - Za imenice – nominativ jednine; za glagole – infinitiv; za prideve – nominativ jednine muškog roda,…
    - Alati koji sprovode lematizaciju se nazivaju *lematizatori*

Stemovanje

- Blisko ideji traženja korena reči, ali bez korišćenja lingvističkog znanja
  - Stemeri odsecaju određene sekvence karaktera sa krajeva reči, ali ne poznaju koncept sufiksa kao takvog
    - *dramski* -> *dram* (-ski jeste sufiks)
    - *Vronski* -> *Vron* (-ski nije sufiks)
- Uklanja kako flektivne, tako i neke derivacione promene (ne utiče na prefikse)
- Stemeri ne uzimaju u obzir okolni kontekst, svaka reč se obrađuje zasebno
- Stemeri se najčešće implementiraju u vidu spiska pravila (mapiranja ili regularnih izraza) za odsecanje krajeva reči
  - Do pravila se može doći i mašinskom analizom, ali se često (i) ručno sastavljaju
  - To ih čini dosta brzim alatima
- Moguće greške zbog prejakog ili preslabog stemovanja

Lematizacija

- Kompleksniji zadatak od stemovanja
- Lematizatori se najčešće implementiraju korišćenjem modela mašinskog učenja i morfoloških rečnika koji mapiraju različite oblike reči u njihove leme
- Da bi se pravilno sprovela, lematizacija zahteva uvid u kontekst u kome se posmatrana reč javlja (tj. okolne reči)
  - *Sedam* – može da bude broj, ili prvo lice jednine prezenta glagola *sedati*
- Zbog ove potrebe, kao i načina implementacije, lematizatori su uglavnom primetno sporiji od stemera

Obeležavanje vrsta reči

- Engl. *Part-of-Speech tagging* (*POS tagging*)
- Zadatak označavanja glavne kategorije kojoj posmatrana reč pripada
  - Glagoli, imenice, pridevi, itd.
- Kategorije koje se koriste pri automatskom obeležavanju (tzv. *tagset*) ne moraju da uvek korespondiraju 1-1 sa lingvističkim kategorijama
  - Najrašireniji skup oznaka za engleski – *Penn Treebank POS Tagset*
  - Najrašireniji skup oznaka za srpski – *MULTEXT-East*
- Obično se sprovodi pre lematizacije
  - Informacija o vrsti reči pomaže u pravilnom određivanju leme
- Za implementaciju se predominantno koriste modeli sekvenci

Obeležavanje morfosintaktičkih osobina reči

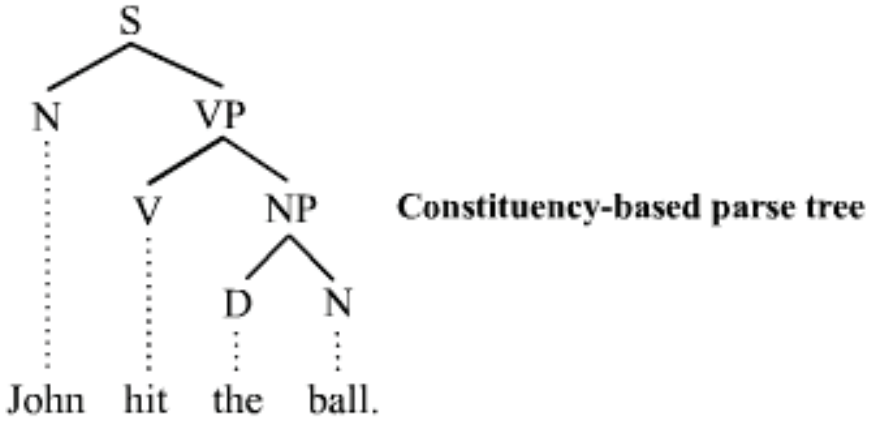
- Zadatak srodan obeležavanju vrsta reči
- Cilj – obeležiti ne samo osnovnu vrstu reči, nego i podvrstu, i opisati njene morfosintaktičke osobine
  - Padež, rod, broj, lice,…
- Skup svih kreiranih opisa se naziva morfosintaktički deskriptor (engl. *morphosyntactic descriptor – MSD*)
- Šta se tačno obeležava u ovom procesu umnogome zavisi od konkretnog jezika i konkretnog skupa oznaka koje se koriste
  - Zadatak je od veće važnosti za morfosintaktički bogatije jezike
  - Najrašireniji skup oznaka za srpski – *MULTEXT-East*
- Za implementaciju se koriste veoma slične tehnike kao i za obeležavanje vrsta reči

Parsiranje

- Zadatak analize gramatičke strukture rečenice i njenog raščlanjivanja na sintaksne delove, pri čemu se označavaju i međusobni odnosi između delova
- Parseri označavaju elemente kao što su subjekat, objekat, glagol, atributi, i povezuju ih koristeći neki gramatički formalizam
- Dva osnovna tipa parsiranja
  - Konstituentsko parsiranje – zasnovano na konstituentskim gramatikama
  - Dependencijalno parsiranje – zasnovano na dependencijalnim gramatikama
- Ranije su konstituentski parseri bili dominantni, danas je primat preuzelo dependecijalno parsiranje
  - Projekat *Universal Dependencies* – cilj: isti skup osnovnih oznaka za parsiranje svih jezika

Parsiranje

- Konstituentski parseri kreiraju sintaksno stablo rečenice u kome su reči listovi stabla
- Oznake (za engleski)
  - N – imenica (*noun*)
  - V – glagol (*verb*)
  - D – član (*determiner*)
  - NP – imenička sintagma (*noun phrase*)
  - VP – glagolska sintagma (*verb phrase*)
  - S – rečenica (*sentence*)

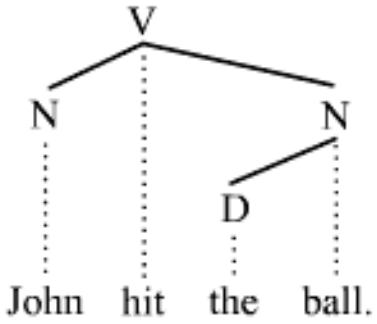


Ilustracija izgleda konstituentskog stabla parsiranja

Slika preuzeta sa:
http://en.wikipedia.org/wiki/Parse_tree

Parsiranje

- Dependencijalni parseri kreiraju sintaksno stablo rečenice u vidu sintaksnih dependencija/zavisnosti između parova reči
- Svaka sintaksna zavisnost ima upravnu reč i zavisnu reč
- Glagol je sintaktički centar jedne klauze, sve ostale reči zavise od njega, neposredno ili posredno.



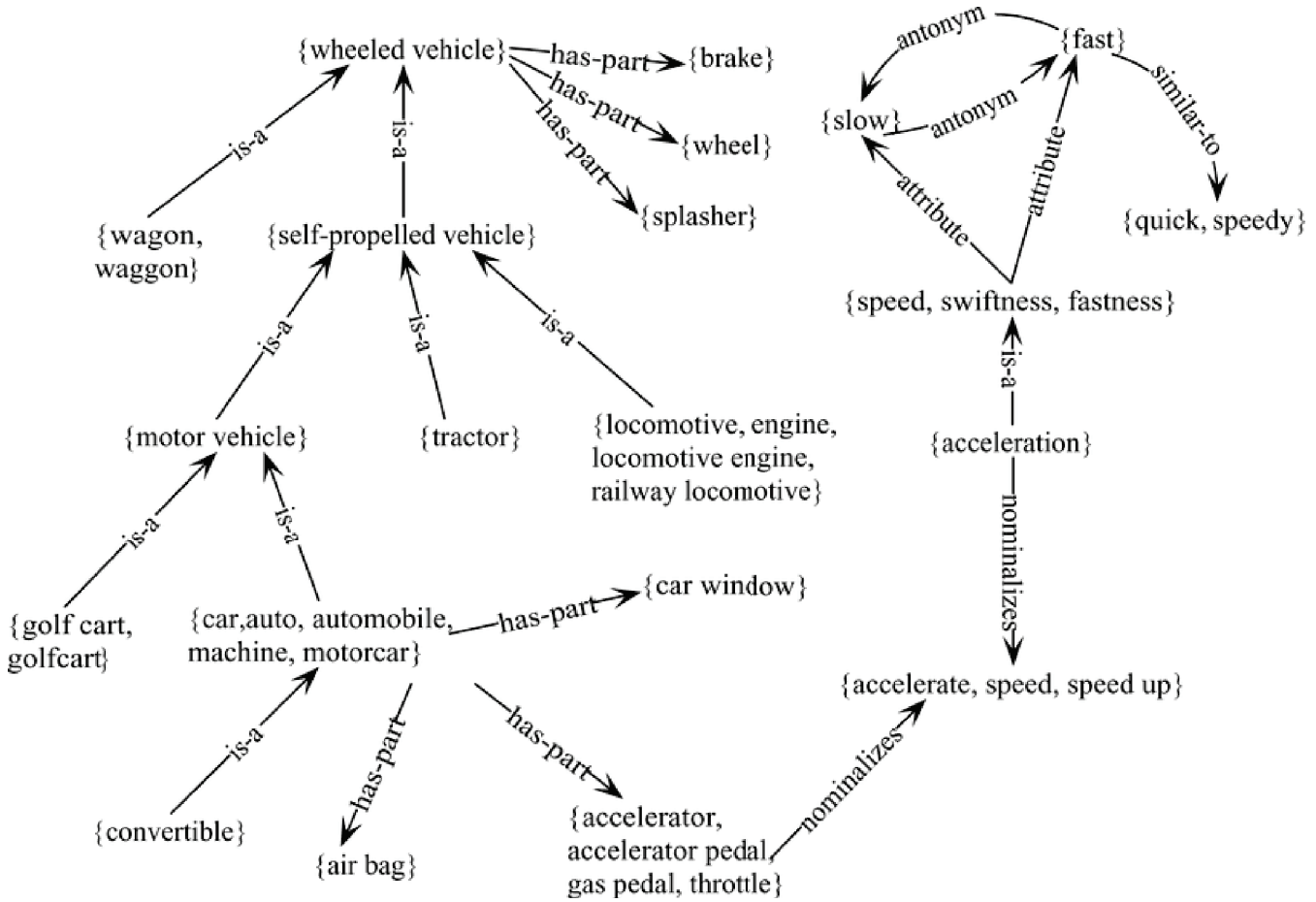
Dependency-based parse tree

Ilustracija izgleda dependencijalnog stabla parsiranja

Slika preuzeta sa:
http://en.wikipedia.org/wiki/Parse_tree

Leksička semantika

- Za mnoge NLP zadatke višeg nivoa potrebno je znati značenje reči
- Značenje reči je moguće dobiti na dva načina
  - Eksplicitno – preko baza znanja
  - Implicitno – preko praćenja upotrebe reči u velikim korpusima tekstova – distribuciona semantika
- *WordNet*
  - Najpoznatija baza znanja o značenjima reči, sadrži definicije i primere upotrebe
  - Reči sa sličnim značenjima su grupisane u skupove sinonima – *synsets*
  - Između *synset*-a postoje hijerarhijski odnosi – hipernim/hiponim (npr. *životinje* se nalazi iznad *ptice*), kao i druge semantičke relacije koje zavise od vrste reči
    - Na taj način je formiran veliki graf svih reči
    - Semantička bliskost dve reči se određuje kao njihova udaljenost u grafu



Ilustracija izgleda dela *WordNet*-a na engleskom jeziku

Slika preuzeta iz rada: R. Navigli, *Word Sense Disambiguation: A Survey*, ACM Computing Surveys 41(2), 2009

“You shall know a word by the company it keeps.”

Džon Rupert Firt, britanski lingvista

- Distribuciona hipoteza – reči sa sličnim značenjima se javljaju u sličnim kontekstima – imaju sličnu *distribuciju* tj. raspodelu u odnosu na druge reči
  - Značenje reči se može izvesti iz toga u kojim kontekstima se ta reč javlja
  - Da bi se utvrdila distribucija reči, neophodan je **veliki** (neanotiran) korpus tekstova

Distribuciona semantika

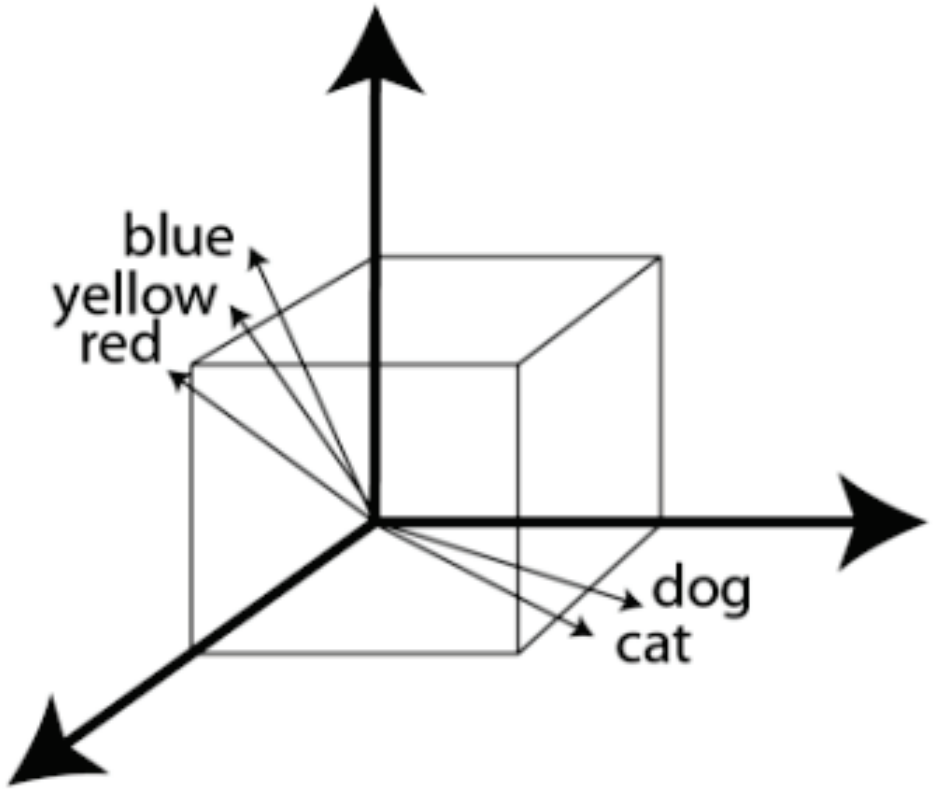
- Šta znači reč „*bardiwac*“ ?
  - *He handed her a glass of bardiwac.*
  - *Beef dishes are made to complement the bardiwacs.*
  - *Nigel staggered to his feet, face flushed from too much bardiwac.*
  - *Malbec, one of the lesser-known bardiwac grapes, responds well to Australia’s sunshine.*
  - *I dined off bread and cheese and this excellent bardiwac.*
  - *The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.*
- Čak i ako ne znamo definiciju reči, na osnovu dovoljno primera njene upotrebe možemo zaključiti da se radi o sorti vina/grožđa

Distribuciona semantika

- Primer: reči *stolica* i *fotelja* se često javljaju u sličnim kontekstima – sledi da su te dve reči semantički bliske
  - *Seo sam na stolicu/fotelju.*
  - *Presvukao sam stolicu/fotelju.*
- Broj javljanja reči u određenom kontekstu se može izraziti u vidu matrice učestalosti međusobnog javljanja posmatrane reči i okolnih, kontekstnih reči (engl. *co-occurrence matrix*)
  - Redovi matrice predstavljaju posmatrane reči, a kolone kontekstne reči
  - U kontekstne reči se obično ubraja nekoliko reči sa obe strane posmatrane reči
    - Kontekst je moguće posmatrati i u vidu celih dokumenata: *LSA – Latent Semantic Analysis* model
  - Zbog izuzetne veličine matrice obično se sprovodi minimizacija njene dimenzionalnosti

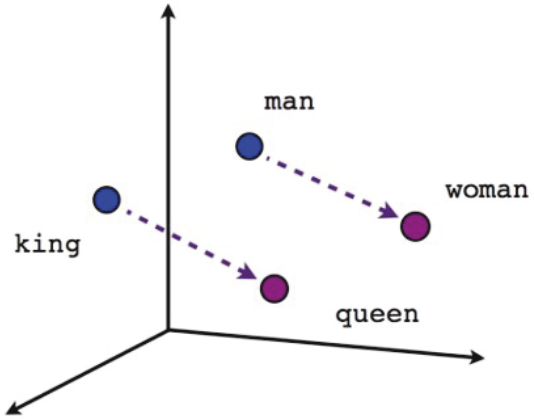
Vektori značenja reči

- Vektori značenja reči (engl. *word embeddings*) – niskodimenzionalni vektori fiksne dužine koji reprezentuju semantiku reči
  - Reči sa sličnim značenjima će imati slične vektore tj. biće blizu jedna drugoj u semantičkom prostoru reči
    - Udaljenost vektora se obično izražava putem kosinusne sličnosti
- Problem interpretacije – dimenzije/ose semantičkog prostora uglavnom nemaju jasno značenje
- Uprkos tome, položaji vektora omogućavaju elementarno rezonovanje u vidu analogija

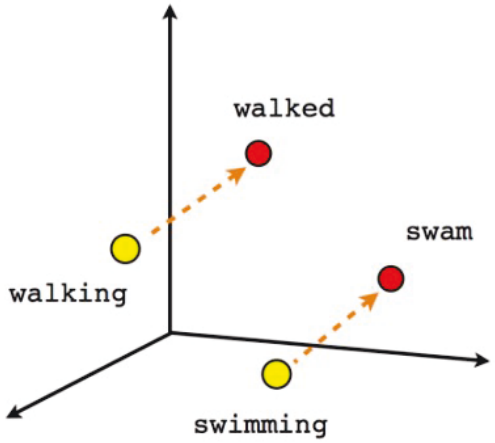


Primer položaja vektora značenja reči u semantičkom prostoru

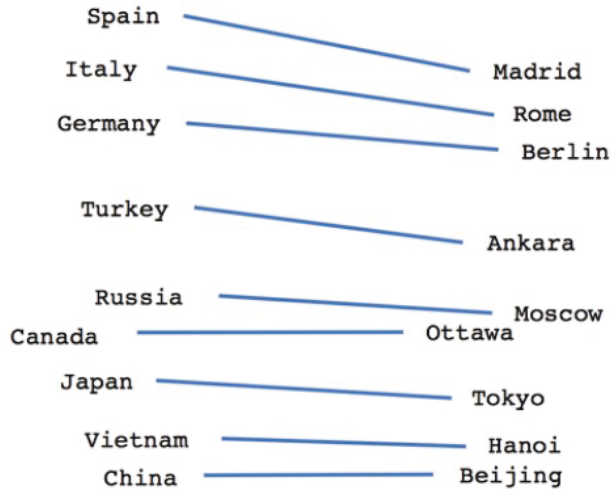
Slika preuzeta sa: http://www.yamagata-europe.com/en-gb/blog/neural-machine-translation-what-s-under-the-hood-part-2



Male-Female



Verb tense



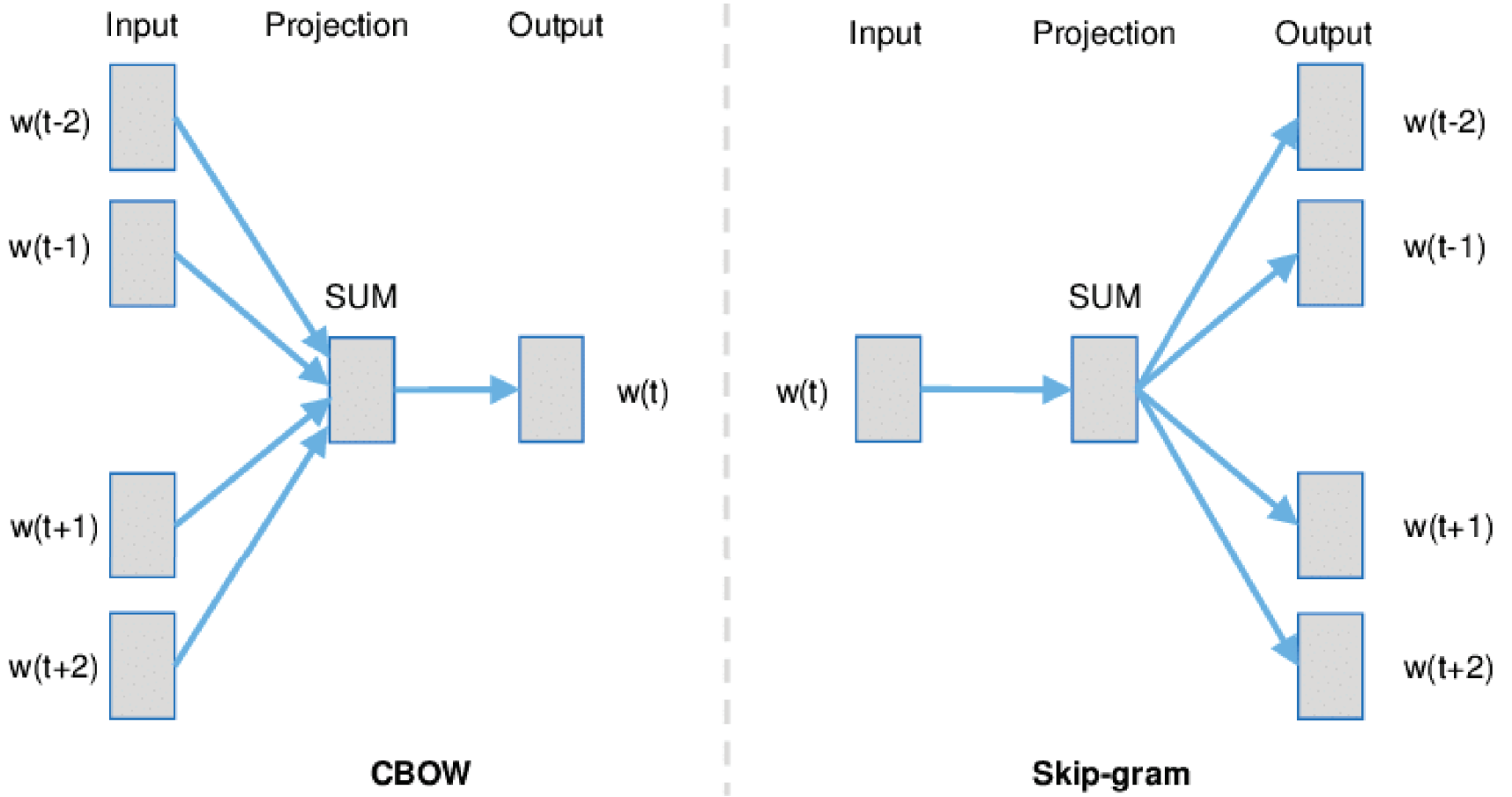
Country-Capital

Ilustracija svojstava vektora značenja reči

Slika preuzeta sa: http://www.brainlabsdigital.com/blog/what-word2vec-means-for-seo/

Vektori značenja reči

- Poznati savremeni algoritmi za generisanje vektora značenja reči:
  - *GloVe* (*Global Vectors*) – zasnovan na brojanju javljanja reči u kontekstima tj. izradi matrice zajedničkog pojavljivanja reči
  - *Word2vec* – neuralni/prediktivni model, umesto brojanja vrši predviđanje
    - CBOW (*Context Bag-of-Words*) – na osnovu datih reči iz konteksta predviđa se nedostajuća reč
    - *Skip-gram* – na osnovu date reči predviđa se njen kontekst
  - Iako naizgled imaju drugačiju logiku rada, pokazano je da i *word2vec* i *GloVe* zapravo rešavaju suštinski isti problem
  - *fastText* – unapređenje *word2vec skip-gram* modela u kome se reči posmatraju kao skupovi znakovnih n-grama
    - Obično bolji rezultati na problemima gde morfologija reči igra ulogu, a blago lošiji na semantičkim problemima



Ilustracija logike rada CBOW i *skip-gram* varijanti *word2vec* algoritma

Slika preuzeta iz rada: J. Landthaler et al, *Extending Thesauri Using Word Embeddings and the Intersection Method*, Second Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2017), London, 2017

Vektori značenja reči

- Klasični vektori značenja reči obuhvataju sva moguća značenja reči
  - Npr. u rečenici *Sto ljudi, sto ćudi* vektor značenja za reč *sto* se istovremeno odnosi i na deo nameštaja i na numeričku vrednost
  - Omogućava da vektori značenja budu statički – izračunati unapred i uvek isti za sve upotrebe određene reči
- Kontekstno-osetljivi vektori značenja (engl. *contextual embeddings*)
  - Različiti vektori značenja za istu reč, zavisno od njenog konkretnog konteksta tj. od okolnih reči
  - Vektori značenja su dinamički – moraju se iznova računati za svaku novu upotrebu tj. nov kontekst u kome se određena reč javlja
  - Dobijaju se uz pomoć neuralnih jezičkih modela: BERT, GPT,…

Klasifikacija tekstova

- Zadatak – svrstati zadati tekst u jednu od predefinisanih kategorija
- Vrste klasifikacije
  - Po temama
  - Po autorstvu
  - Detekcija spama
  - Analiza sentimenta se često koncipira kao zadatak klasifikacije
- Tipično se za potrebe klasifikacije tekstova koriste modeli mašinskog učenja – naivni Bajesov klasifikator, logistička regresija, metoda potpornih vektora,…
  - Da bi se modeli obučili, neophodno je najpre (ručno) anotirati ispravnu kategoriju za podatke iz skupa za obučavanje

Klasifikacija tekstova

- Pristup vreće reči (engl. *bag-of-words* – *BOW*)
  - Tretirati tekstualni dokument kao neuređen skup reči koje sadrži – ignoriše se njihov redosled
  - Veoma jednostavan pristup za potrebe klasifikacije, ali prilično dobar za duže tekstove/dokumente
  - Svaka reč se tretira kao jedna od odlika u klasifikacionom modelu
  - Pored pojedinačnih reči, moguće je kao odlike tretirati i sekvence od više uzastopnih reči (vreća n-grama)
- Moguće su i druge vrste odlika
  - Npr. odlike dobijene na osnovu vektora značenja reči u tekstu

Klasifikacija tekstova

- Česti koraci u pretprocesiranju tekstova pre klasifikacije
  - Tokenizacija tekstova
  - Morfološka normalizacija (opciono)
  - Uklanjanje stop reči (opciono)
- Stop reči – reči koje se vrlo često javljaju u jeziku a ne nose semantičku informaciju (ili je ona zanemarljiva)
  - Funkcionalne reči – predlozi, veznici, uzvici, rečce
  - Moguće je definisati i neke domenske stop-reči

Analiza sentimenta

- Zadatak – odrediti kakav sentiment se izražava u zadatom tekstu
- Najčešće i najjednostavnije se koncipira kao problem klasifikacije tekstova na pozitivne/negativne (/neutralne)
- Moguće i korišćenje numeričkih oznaka za izražavanje jačine sentimenta
  - Na primer, skala od -5 do +5, gde je -5 najnegativniji, a +5 najpozitivniji sentiment
- Analizu sentimenta je moguće raditi na različitim nivoima granularnosti teksta
  - Nivo dokumenata
  - Nivo rečenica/paragrafa
  - Nivo aspekata – pod aspektom se podrazumeva bilo koji element prema kome se izražava sentiment (ličnost, pojava, događaj, tema, svojstvo, itd.)
    - ***Hrana*** *je bila odlična, ali je* ***usluga*** *spora!*

Aspektna analiza sentimenta

- Osnovni elementi u aspektnoj analizi sentimenta
  - Aspektni izraz (engl. *aspect term*) – konkretna reč ili izraz u tekstu prema kojoj/kome se ispoljava sentiment
  - Izraz ispoljavanja sentimenta (engl. *opinion term*) – konkretna reč ili izraz u tekstu pomoću koje/koga se ispoljava sentiment
  - Aspektna kategorija (engl. *aspect category*) – odnosi se na određen aspekt domena koji se razmatra, pokriva više/puno aspektnih termina
    - Npr. kategorija *hrana* u domenu restorana može pokrivati puno termina za različita jela
  - Polarnost sentimenta – najčešće se klasifikuje u pozitivnu/negativnu/neutralnu
  - Primer: ***Roštilj*** *im je izvrstan!* – aspektni izraz: *roštilj*; aspektna kategorija: *hrana*; izraz ispoljavanja sentimenta: *izvrstan*; polarnost sentimenta: *pozitivna*
- Određivanje ovih elemenata (samostalno ili u parovima) predstavlja posebne podzadatke u aspektnoj analizi sentimenta

Analiza sentimenta

- Neka konceptualna pitanja u obeležavanju sentimenta
  - Šta tačno spada u neutralnu klasu (ako postoji) – odsustvo sentimenta (tj. objektivni tekstovi) i/ili mešavina sentimenata (i pozitivni i negativni)?
  - Kako obeležiti dvosmislene tekstove?
    - *„U pravu je, ovako nešto definitivno niste videli ranije.“*
  - Kako pouzdano prepoznati sarkazam u pisanim tekstovima?
    - U govoru sarkazam se često ispoljava preko tona govornika
    - Poseban problem detekcije sarkazma
- Granularnije posmatranje sentimenta često olakšava konzistentnu anotaciju
  - Pogotovo kada se različit sentiment ispoljava u različitim rečenicama / prema različitim aspektima

Analiza sentimenta

- Dominantan pristup – modeli nadgledanog mašinskog učenja
- Alternativan/dopunski pristup – sentiment leksikoni
  - Rečnici reči (i izraza) koji nose određen sentiment
  - Svakom unosu u leksikon pridodata polarnost sentimenta koji izražava, a obično i ocena jačine sentimenta
  - *SentiWordNet* – opšti sentiment leksikon izveden iz *WordNet*-a
- Analiza sentimenta je domenski osetljiv problem
  - Reč *nepredvidiv* ima pozitivnu konotaciju ako se govori o radnji filma, a negativnu ako se govori o sistemu upravljanja u automobilu
  - Izraz *pročitajte knjigu* ima pozitivnu konotaciju ako se nalazi u recenziji knjige, a negativnu ako se nalazi u recenziji filma snimanog po knjizi

Analiza sentimenta

- Komercijalno izuzetno atraktivan zadatak
  - Ogroman broj preduzeća želi da zna šta njihovi korisnici misle o njihovim proizvodima i o proizvodima konkurencije
  - Od interesa i za sve ostale društvene aktere koji žele da prate sentiment javnosti prema nekom entitetu
- (Kompleksniji) srodni problemi
  - Klasifikacija tekstova po subjektivnosti (subjektivni/objektivni)
  - Prepoznavanje emocija – izlazak izvan okvira pozitivno/negativno i određivanje konkretnih emocija kao što su ljutnja, radost, žalost, itd.
  - Detekcija stavova (prema tvrdnji sagovornika – slaganje/neslaganje)
  - Detekcija sarkazma

Prepoznavanje imenovanih entiteta

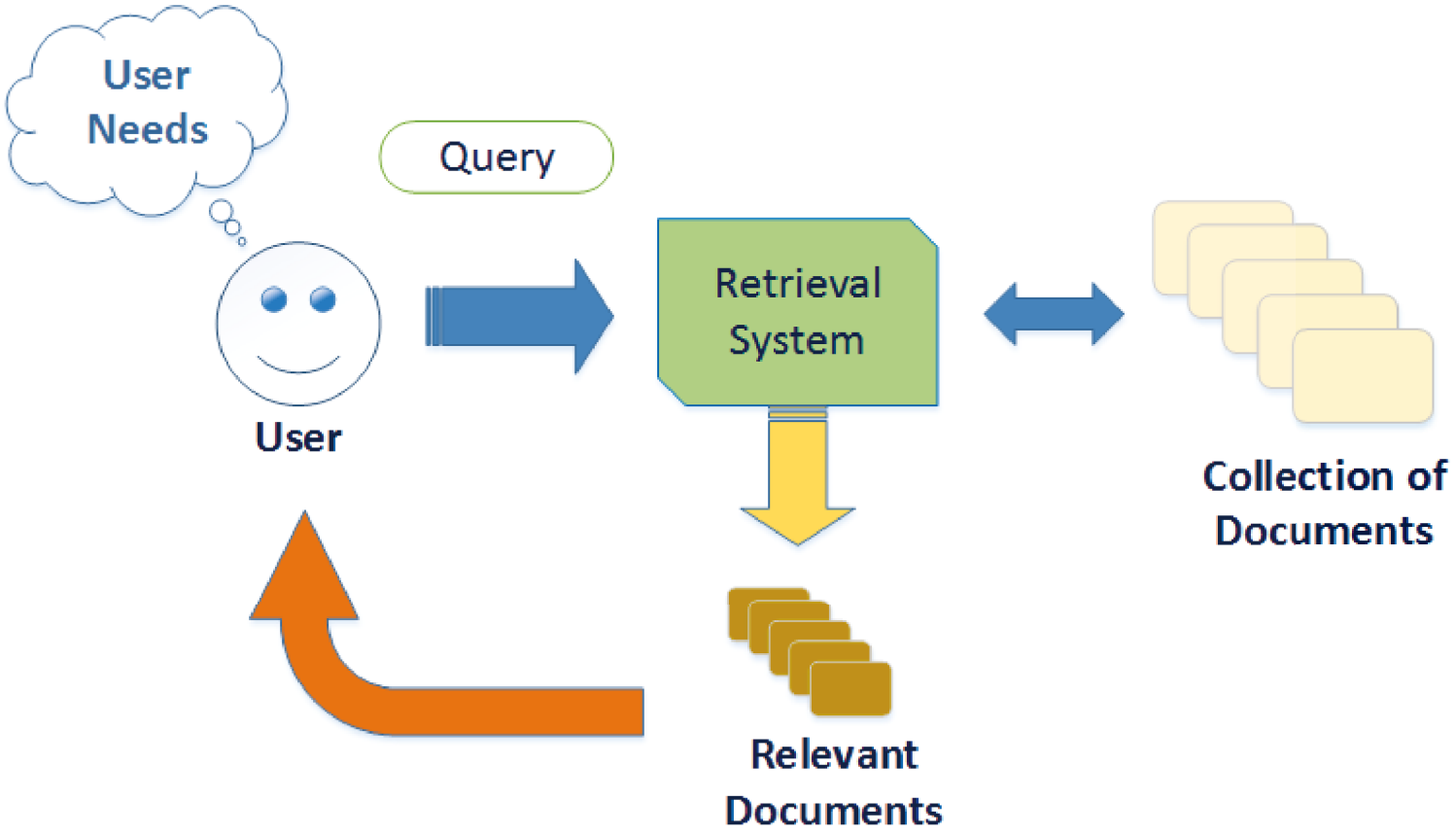
- Zadatak prepoznavanja imena različitih entiteta i označavanja njihovog tipa
- Glavne kategorije imenovanih entiteta
  - Osobe – *PERson* – *Majkl Džordan, Dž. Edgar Huver*
  - Lokacije – *LOCation* – *Srbija, Beč, Trg republike*
  - Organizacije – *ORGanization* – *Banca Intesa, Imlek, Univerzitet u Beogradu*
- Moguće je definisati i dodatne/drugačije kategorije entiteta – vremenski izrazi, prisvojni oblici ličnih imena, razno,…
- Komplikovanije od prostog detektovanja vlastitih imenica
  - *Trg republike* jeste imenovani entitet, ali su obe imenice koje ga čine zajedničke

Prepoznavanje imenovanih entiteta

- B-I-O sistem anotiranja tokena
  - B – *begins* – token kojim započinje neki imenovani entitet
  - I – *inside* – token koji pripada već započetom imenovanom entitetu
  - O – *outside* – token koji ne pripada nijednom imenovanom entitetu
- Moguće tretirati kao problem klasifikacije (označavanje jednog po jednog tokena) ili kao problem označavanja sekvence
- Kao odlike u modelima mašinskog učenja se za ovaj problem često koriste sadržaj tokena, njegova lema, vrsta reči i morfosintaktički deskriptor za posmatrani token, prisustvo/odsustvo kapitalizacije, itd.

Dohvatanje informacija

- Zadatak pronalaženja izvora informacija (obično dokumenata) koji su nestrukturirane prirode (najčešće tekstovi) i koji zadovoljavaju potrebe korisnika
- Obično je pri dohvatanju informacija potrebno obraditi/proći kroz veliku kolekciju postojećih dokumenata
- Korisnik formuliše svoju potrebu za informacijama kroz zadavanje upita
- Sistem poredi upit sa indeksom dokumenata u kolekciji i vraća one koji odgovaraju upitu
- Cilj – da sistem vrati sve dokumente koji su relevantni za zadati upit, kao i da ne vrati nijedan irelevantan dokument



Ilustracija zadatka dohvatanja informacija

Slika preuzeta sa: http://ir.cs.ui.ac.id/new/

Dohvatanje informacija

- Osnovni pristup – model vektorskog prostora
  - Tretirati i dokumente i upite kao neuređene skupove reči (vreće reči)
  - Indeksirati koje reči se javljaju u kojim dokumentima
  - Računati sličnost svakog dokumenta sa upitom kao sličnost njihovih BOW vektora
    - Mera sličnosti – euklidska, kosinusna, Menhetn,…
    - Moguće je isfiltrirati sadržaj dokumenata tako da neke reči ne ulaze u sastav vektora dokumenata (npr. stop reči), ili dati veću težinu nekim rečima
- Česte upotrebe
  - Dohvatanje informacija iz ličnih baza dokumenata, npr. pretraga e-mailova
  - Dohvatanje informacija sa interneta
  - Dohvatanje informacija iz specijalizovanih ili domenskih baza dokumenata

Dohvatanje informacija

- Često se koriste dve vrste težinskog ponderisanja
  - TF (engl. *Term Frequency*) ponderisanje
  - IDF (engl. *Inverse Document Frequency*) ponderisanje
  - Često se kombinuju – TF-IDF ponderisanje
  - Često se primenjuju i u drugim NLP problemima, npr. klasifikaciji tekstova
- TF (engl. *Term Frequency*) ponderisanje
  - Relevantnost dokumenta d za određeni upit raste sa porastom frekvencije javljanja reči t iz upita u dokumentu, ali ne linearno, već logaritamski
    - Ako se reč iz upita javi u dokumentu A 10x, a u dokumentu B 1x, A jeste relevantnije od B, ali ne 10x

$$TF = \begin{cases} 1 + \log_{10} Count(t)_d & , Count(t)_d > 0 \\ 0 & , Count(t)_d = 0 \end{cases}$$

Dohvatanje informacija

- IDF (engl. *Inverse Document Frequency*) ponderisanje
  - Reči koje se javljaju u svim dokumentima su manje važne od reči koje se javljaju u malom broju dokumenata – retkim rečima treba dati veću težinu
  - df_t je broj dokumenata u kojima se reč t javlja
  - N je ukupan broj dokumenata u sistemu

$$IDF = \log_{10} \frac{N}{df_t}$$

  - Logaritam se koristi da bi ublažio efekat inverzne frekvencije po dokumentima
- TF-IDF ponderisanje – težina reči raste sa njenom učestalošću i retkošću

$$TFIDF = (1 + \log_{10} Count(t)_d) \times \log_{10} \frac{N}{df_t}$$

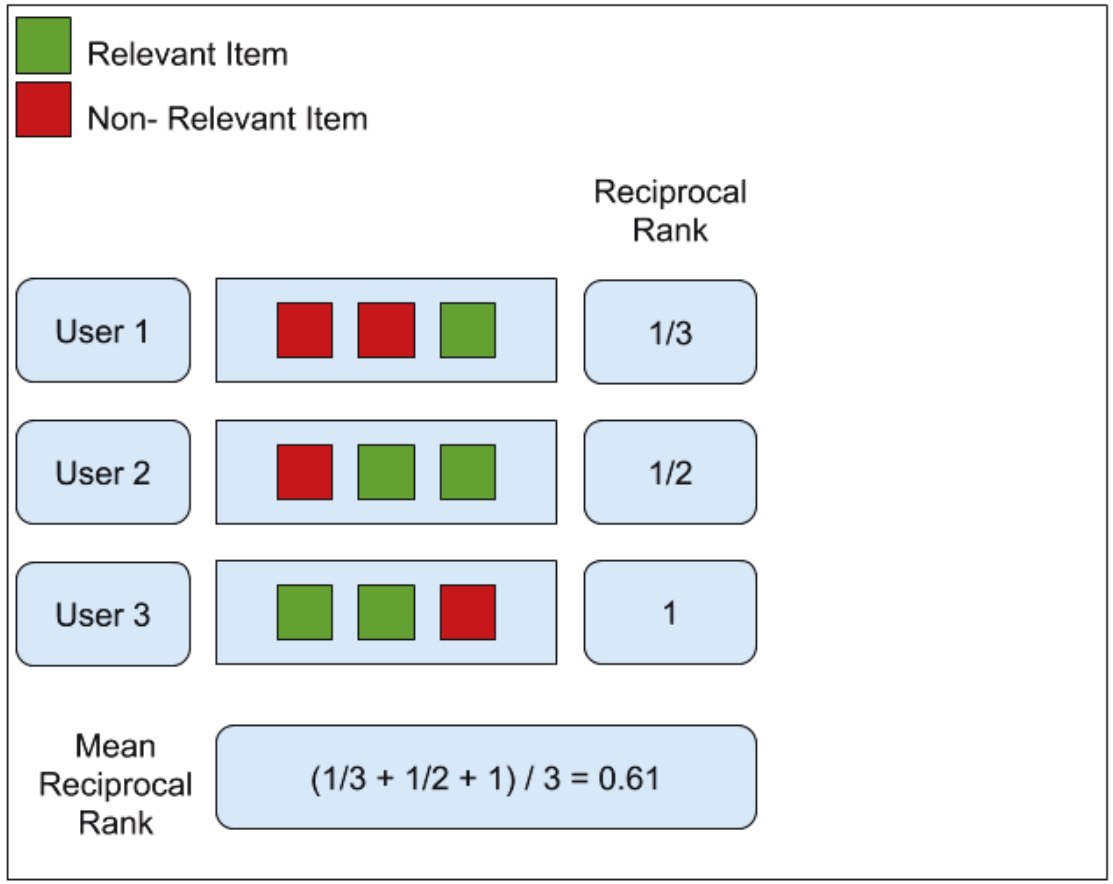
Dohvatanje informacija

- Jedna od osnovnih metrika za merenje uspešnosti sistema za dohvatanje informacija je srednji recipročni rang (engl. *Mean Reciprocal Rank – MRR*):

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i}$$

  gde je Q broj upita, a $rank_i$ je rang (prvog) tačnog odgovora za upit i

- Relevantnost odgovora se posmatra na binaran način (odgovor jeste/nije relevantan za zadati upit)
- Srednji recipročni rang ignoriše sve druge relevantne odgovore iza prvorangiranog



Ilustracija računanja srednjeg recipročnog ranga

Slika preuzeta sa: http://medium.com/swlh/rank-aware-recsys-evaluation-metrics-5191bba16832

Četbotovi

- Termin četbot (engl. *chatbot*) se koristi za razne dijaloške sisteme koji se često dramatično međusobno razlikuju po složenosti
  - Najjednostavniji četbotovi – prepoznavanje ključnih reči
  - Najsavremeniji četbotovi – konverzacioni veliki jezički modeli – OpenAI ChatGPT, Google Bard
    - Pored velikih jezičkih modela uključuju tehnike poboljšanja kvaliteta sistema korišćenjem učenja sa podrškom (engl. *reinforcement learning*)
- Tipični elementi/podzadaci savremenih četbotova
  - Prepoznavanje namere (engl. *intent recognition*) – kategorizacija cilja koji korisnik želi da ostvari porukom
  - Prepoznavanje imenovanih entiteta

Četbotovi

- Tipični elementi/podzadaci savremenih četbotova
  - Upravljanje dijalogom – često se formalizuje kao kretanje kroz predefinisana stanja u konverzacionom grafu
    - Omogućava sistemu da prati prethodne interakcije i da vraća smislene odgovore u posmatranom stanju
  - Odgovaranje na pitanja (engl. *question answering*)
    - Pronalaženje i (re)formulisanje odgovora na postavljeno pitanje na osnovu baze znanja ili skupa dokumenata
    - Složeniji oblik problema dohvatanja informacija
  - Generisanje odgovora
    - Zasnovano na pravilima/šablonima
    - Zasnovano na naprednijim tehnikama generisanja prirodnog jezika (engl. *natural language generation*)

Određivanje semantičke sličnosti tekstova

- Zadatak – za zadata dva teksta odrediti stepen njihove semantičke sličnosti/različitosti na nekoj skali (obično 0–5)
- Cilj – da sistem prepozna kada su dva teksta semantički bliska, iako su možda veoma različiti u pogledu leksike i sintakse
  - *Ubrzo će doći do poskupljenja struje.*
  - *Cena električne energije će uskoro porasti.*
- Netrivijalan problem, naročito za kratke tekstove
- Većina pristupa se zasniva na formiranju vektorske reprezentacije značenja svakog od tekstova i poređenju tih vektora
- Igra veliku ulogu u sistemima za pretraživanje i dohvatanje informacija, sumarizaciji teksta, odgovaranju na pitanja,…

Mašinsko prevođenje

- Zadatak – prevesti iskaz sa jednog jezika na drugi tako da se sačuva i semantička informacija i prirodnost izražavanja
- Problemi
  - Kompleksnost, nejasnost, višesmislenost jezika
    - *The computer outputs the data; it is fast.* – *Računar proizvodi podatke; on je brz.*
    - *The computer outputs the data; it is stored on a disk.* – *Računar proizvodi podatke; oni se smeštaju na disk.*
  - Razlike u gramatičkim svojstvima izvorišnog i ciljnog jezika
    - Redosled reči
      - U engleskom: subjekat-predikat-objekat (SVO)
      - U japanskom: subjekat-objekat-predikat (SOV)
      - U srpskom: u principu slobodan, ali je SVO najčešći

Mašinsko prevođenje

- Statističko mašinsko prevođenje
  - Paralelni korpusi tekstova se koriste kao primeri ispravnih prevoda
    - U početku se kao par jezika najčešće javljao engleski i francuski – parlament Kanade zvanično koristi oba jezika, te su paralelni transkripti skupštinskih sednica najčešće korišćeni pri izradi modela
    - Često se koriste transkripti sednica evropskog parlamenta
  - *State-of-the-art* do pre nekoliko godina – modeli zasnovani na frazama (engl. *phrase-based machine translation*)
    - Ne prevode se pojedinačne reči, već grupe reči – uparuju se izrazi u jednom jeziku sa ekvivalentnim izrazima u drugom
  - *State-of-the-art* danas – neuralno mašinsko prevođenje
    - *Sequence-to-sequence* modeli – tekst na polaznom jeziku se enkoduje u vektorski format, a zatim se takav vektor dekodira u izlazni jezik
      - Dužina ulaznih i izlaznih sekvenci može da se razlikuje

Sumarizacija teksta

- Zadatak – proizvesti skraćenu verziju početnog teksta koja sadrži informacije bitne/relevantne za korisnika
- Primeri sumarizovanih tekstova koje ljudi pišu – novinski naslovi, apstrakti naučnih radova, opisi knjiga,…
- Automatska sumarizacija se može raditi na osnovu jednog polaznog dokumenta ili više njih
- Po tipu razlikuju se
  - Ekstraktivna i apstraktivna sumarizacija
    - Ekstraktivna – sadrži neizmenjene delove polaznog teksta
    - Apstraktivna – generiše (bar delimično) nov tekstualni sadržaj koji komprimuje informacije sadržane u polaznom tekstu
      - Obuhvata zadatak generisanja prirodnog jezika (engl. *Natural Language Generation – NLG*)

Sumarizacija teksta

- Po tipu razlikuju se
  - Generička i upitno-orijentisana sumarizacija
    - Generička – sumarizuje celokupan sadržaj dokumen(a)ta
    - Upitno-orijentisana – sumarizacija se fokusira na one delove dokumen(a)ta koji su relevantni za upit korisnika
- Osnovni algoritam ekstraktivne sumarizacije
  - Odrediti stepen relevantnosti za svaku rečenicu
    - Moguće korišćenjem nadgledanih ili nenadgledanih metoda
  - Odabrati najrelevantne rečenice i eliminisati redundantnosti
  - Odrediti redosled rečenica/informacija

Ostali NLP zadaci

- Postoje i drugi definisani zadaci/problemi u NLP-u
  - Razrešavanje koreferenci/anafora
  - Simplifikacija teksta
  - Ekstrakcija odnosa između entiteta
  - Ekstrakcija ključnih reči
  - Modelovanje tema
  - …

Ostali NLP zadaci

- Relativno često se formulišu novi zadaci ili nove varijante starih zadataka
  - Npr. obeležavanje imenovanih entiteta u domenu biomedicine
- Novi zadaci se često popularizuju putem tzv. deljenih zadataka (engl. *shared tasks*) – javnih međunarodnih takmičenja timova koji prilažu svoja rešenja zadatog problema
  - Na takav način se istraživačkoj zajednici daje pristup i odgovarajućim (anotiranim) podacima za posmatrani problem