Obrada prirodnih jezika

Elektrotehnički fakultet Univerziteta u Beogradu

Master akademske studije

modul Softversko inženjerstvo

2022/2023

Uvod u obradu prirodnih jezika

Vuk Batanović, Elektrotehnički fakultet Univerziteta u Beogradu

Obrada prirodnih jezika

- Engl. *Natural Language Processing – NLP*
- Oblast na raskršću računarskih nauka (veštačke inteligencije/mašinskog učenja), inženjerstva, lingvistike i (kognitivne) psihologije
- Često se za NLP koristi i alternativan naziv – računarska lingvistika (engl. *computational linguistics*)
- Pod *prirodnim* se podrazumevaju jezici koje ljudi prirodno koriste za međusobnu komunikaciju
  - Engleski, srpski, japanski,…
  - Nasuprot veštačkim jezicima
    - Jezici specijalne namene – npr. Morzeovi znaci
    - Programski jezici

Čemu računarska obrada prirodnih jezika?

- Podaci na prirodnim jezicima u tekstualnom obliku su prisutni svuda i sve ih je više
  - Knjige, časopisi, novine, e-mail, društvene mreže,…
- Najveći deo ljudskog znanja je dostupan samo u nestrukturiranom tekstualnom obliku (nasuprot strukturiranim izvorima znanja poput relacionih baza podataka)
- Tekstualni podaci gotovo uvek imaju svoju internu strukturu, ali ona se obično ne može lako i direktno prevesti u mašinski čitljiv oblik
  - Lingvistička struktura – sintaktička i diskursna struktura
  - Struktura formatiranja – podela na pasuse, sekcije/poglavlja/glave

Zašto je računarska obrada prirodnih jezika teška?

- Prirodni jezici imaju veći broj osobina koji ih čine teškim za mašinsku obradu
  - Veliki stepen kompleksnosti
  - Česte nejasnoće u izražavanju
  - Česte dvosmislenosti/višesmislenosti u izražavanju
  - Razumevanje iskaza često zavisi od posedovanja šireg znanja o svetu (engl. *world knowledge*)
- Potpuno razumevanje prirodnih jezika je AI-potpun problem
  - Da bi se rešio, potrebno je da računari dostignu/prestignu ljudski nivo inteligencije
  - Nije moguće rešiti ovaj problem korišćenjem nekog uže specijalizovanog, jednostavnijeg algoritma

Kompleksnost jezika

- Postoji više različitih aspekata kompleksnosti jezika
- Morfološka kompleksnost – više različitih oblika jedne iste reči
- Sintaktička kompleksnost – više različitih sintaktičkih struktura koje su validne za jedan isti iskaz
- Semantička kompleksnost – više različitih značenja za jednu istu reč/iskaz
  - Frazeologizmi/idiomi – ustaljene jezičke jedinice od bar dve reči, imaju jedinstveno značenje koje se ne može izvesti iz značenja reči u njihovom sastavu
    - *Carski rez*
    - *Kupiti mačku u džaku*
    - *Izvoditi besne gliste*
    - *Novinarska patka*
    - *Obrati (zelen) bostan*
    - *Slepi putnik*

Kompleksnost jezika

- Semantička kompleksnost
  - Za razumevanje značenja frazeologizama često je neophodan kulturološki kontekst/znanje o svetu
    - *Lupati kao Maksim po diviziji*
    - *Razbiti kao bugarsku skupštinu*
    - *Ahilova peta*
  - Ovo deluje intuitivno jasno kada razmatramo kulturu u kojoj smo odrasli, ali postaje izuzetno teško kada se razmatra kultura koja nam je nepoznata
    - U kineskom jeziku postoji 5000 – 20000 idioma, izvedenih iz klasičnih kineskih tekstova
      - *Pritajeni tigar, skriveni zmaj*
    - Ekstreman primer iz naučne fantastike – *Star Trek: The Next Generation, Season 5, Episode 2: Darmok*

Nejasnoće u izražavanju

- Nejasnoće u izražavanju na prirodnim jezicima su ponekad slučajne, ali su vrlo često i namerne
  - Naročito kada se radi o političkim ili reklamnim porukama
- Primeri nejasnih reklamnih poruka (preuzeti iz rada *Vidaković, Trninić-Janjić (2017): Verbalni mamci u reklamnim oglasima na engleskom i srpskom jeziku*)
  - *Traje 3x duže nego što ste očekivali.* (Fairy)
  - *Muzika nikada nije bolje zvučala.* (zvučnici)
  - *Vozite dalje. Živite bolje.* (Renault)
  - *Izuzetna nega sa dva patenta pomaže u borbi protiv znakova starenja kože.* (krema za kožu)
  - *Verujte snazi Q10 plus bisernih perli u borbi protiv bora.* (krema za kožu)

Nejasnoće u izražavanju

- Primeri političkih fraza
  - *Nastavak strukturnih reformi/restrukturiranje*
  - *Treba ostaviti prošlost iza sebe i okrenuti se budućnosti*
  - *Reformska agenda je neophodna za ekonomski oporavak zemlje*
  - *Po prvi put u novijoj istoriji*
  - *Neophodan je pragmatičan pristup*
  - *Rešili smo da preuzmemo odgovornost*

Nejasnoće u izražavanju

- Povezivanjem i kombinovanjem političkih fraza dobijaju se celi iskazi za koje je nejasno šta (i da li išta) znače
  - *Mnogo se s pravom očekuje od mjera u sklopu projekta za Slavoniju, Baranju i Srijem, posebice kada je riječ o otvaranju prostora novim proizvodnim i drugim razvojnim investicijama s krajnjim ciljem pune zaposlenosti, rasta plaća i standarda. To je od ključne važnosti kako bismo zadržali mlade i očuvali ljudske resurse za buduće generacije, poručila je Grabar Kitarović.*
  - *Takođe smo od strane jedne od najaktivnijih međunarodnih nevladinih organizacija, koja se bavi Opštim periodičnim pregledom, preporučeni kao primjer pozitivne prakse u vođenju preglednog mehanizma, odnosno načina uključivanja UN agencija i nacionalnih nevladinih organizacija u proces – kazao je Đukanović.*

Dvosmislenosti/višesmislenosti u izražavanju

- Veći broj različitih izvora dvosmislenosti/višesmislenosti u jeziku
  - Kategorijska
  - Leksička
  - Sintaktička
  - Referencijalna
  - Pragmatička
- Većina iskaza sadrži bar neki oblik dvosmislenosti
  - Ljudi najčešće to ni ne primećuju
    - Ljudi dvosmislenosti uglavnom razrešavaju korišćenjem šireg tekstualnog konteksta, znanja o svetu i zdravorazumskih pretpostavki
  - Ovo svojstvo jezika se vrlo često koristi u šalama i vicevima

Kategorijska dvosmislenost/višesmislenost

- Proističe iz toga što jedan isti oblik reči može da predstavlja više vrsta reči
- *Ona je bila mlada.*
  - Ona je imala malo godina.
  - Ona je bila nevesta.
- *Gore gore gore gore.*
  - Jedan isti oblik reči može da bude i imenica, i glagol, i prilog, i pridev
  - Lošija brda gore iznad. / Brda iznad lošije gore. (pored kategorijske, postoji i sintaktička dvosmislenost)

Leksička dvosmislenost/višesmislenost

- Polisemija – više srodnih značenja jedne reči, nastaje zbog uočavanja sličnosti
  - *Crkva – građevina/institucija*
  - *Glava – deo tela/strana novčića/deo knjige/starešina, rukovodilac/…*
  - *Miš – životinja/uređaj za interakciju sa računarom*
  - *Kucati – tekst/na vrata*
- Homonimija – više potpuno nepovezanih značenja jedne reči
  - *Sto – nameštaj/broj*
  - *Pop – sveštenik/muzički žanr*
  - *Zavijati – umotavati/otegnuto urlati*
- Može da iskomplikuje tumačenje značenja rečenice
  - *Siledžijama će da odgovaraju nadležni organi.*

Sintaktička dvosmislenost/višesmislenost

- Prostiče iz toga što neka sekvenca reči može imati više strukturnih tumačenja
  - Često proizlazi iz dvosmislenosti vezivanja predložno-padežnih konstrukcija
- Nekad je rešiva korišćenjem šireg znanja o svetu
  - Naslov: *Sve složenije operacije u novopazarskoj Opštoj bolnici*
  - Primeri iz: Uvod u opštu lingvistiku, Ranko Bugarski (2003)
    - *Dve devojčice ujele meduze.*
    - *Veliki broj studenata ugrožava nastavu.*
    - *Razgovor doktora Petra Petrovića o braku sa pitomcima JNA*

Sintaktička dvosmislenost/višesmislenost

- Nekad ni šire znanje o svetu (osim ako nije vrlo specifično) ne pomaže
  - *Održao je predavanje na skupu o tehnološkim kretanjima u Londonu.*
  - *Kupio sam košulju i džemper na pruge.*
  - *Vidim čoveka bez naočara.*
  - *Doneo sam ti voće iz Španije.*
  - *Pozvao je prijatelje sa Malte.*
  - Primeri iz: Uvod u opštu lingvistiku, Ranko Bugarski (2003)
    - *Jovanu je lako verovati.*
    - *Stvaranje umetnika*
    - *Kritika članstva*

Referencijalna dvosmislenost/višesmislenost

- Proizlazi iz više mogućih tumačenja na koji entitet ili deo rečenice se neki iskaz odnosi
- *Petar je javio Marku da je kupio auto.*
  - Petar je kupio auto.
- *Petar je javio Marku da je dobio posao.*
  - Ko je dobio posao?
- – *„Profesorka, da li treba da popunjavamo neme karte ili ne?“ – „Naravno!”*
  - Referencijalna dvosmislenost zbog izostavljanja određenih reči (elipse)
- *Ako je bankarska kartica ispravna, klijent treba da unese svoj PIN broj u ATM. U suprotnom, biće ispisana poruka o neuspeloj transakciji.*
  - Da li se poruka ispisuje kada je kartica neispravna, ili kada klijent ne unese PIN?

Pragmatička dvosmislenost/višesmislenost

- Proizlazi iz više mogućih tumačenja cilja koji se želi postići nekim iskazom, iako je osnovno značenje iskaza jasno
- Pragmatičke dvosmislenosti/višesmislenosti se često ne mogu razrešiti ni uvidom u širi tekstualni kontekst
- *Znate li koliko je sati?*
  - Može da predstavlja upit za tačno vreme ili prekor povodom kašnjenja
- *Na putu sam.*
  - Može da predstavlja napomenu o tome da je govornik na odsustvu, ili uveravanje da ubrzo stiže.

Istorijat obrade prirodnih jezika

- 1950 – Alan Tjuring predložio test inteligencije za mašine koji se zasniva na pisanoj komunikaciji u realnom vremenu između mašine i čoveka
  - Postavka – jedan čovek (sudija) komunicira pisanim putem sa drugim čovekom i sa mašinom; oba sagovornika pokušavaju da se predstave kao ljudi
  - Cilj – da mašina uspešno prevari sudiju, tako da on nije siguran koji sagovornik je mašina a koji čovek
  - Tjuring predvideo da će do kraja 20. veka mašina sa 10 GB memorije biti u stanju da zavara 30% ljudi tokom petominutnog razgovora
  - Od 1991. – Lebnerova nagrada – godišnje takmičenje programa u formatu Tjuringovog testa
  - Glavna ideja – korišćenje jezika na nivou ljudi je dovoljno kao operativni test inteligencije
    - Kasnije dovedeno u pitanje – Kineska soba, Džon Serl

Istorijat obrade prirodnih jezika

- 1957 – Noam Čomski objavio *Sintaksičke strukture*
  - Opisuje jezik kroz ideju generativne gramatike – skupa logičkih pravila pomoću kojih je moguće generisati beskonačan broj mogućih ispravnih rečenica u zadatom jeziku i pridodati im ispravan strukturni opis
  - Jedno od najvažnijih dela u modernoj lingvistici
- Krajem 1950-tih i početkom 1960-tih – podela istraživanja na dve paradigme – simboličku i stohastičku
- Simbolička paradigma
  - Rad Čomskog i drugih na formalnoj teoriji jezika i generativnoj sintaksi
  - Razvoj algoritama za parsiranje
  - Istraživanja u domenu rezonovanja i logike – prvi sistemi za razumevanje prirodnih jezika zasnovani na prepoznavanju ključnih reči i jednostavnim heuristikama za prepoznavanje obrazaca i odgovaranje na pitanja

Istorijat obrade prirodnih jezika

- 1966 – Džozef Vajzenbaum napravio program ELIZA – prvi NLP sistem sposoban za (vrlo ograničenu) konverzaciju sa korisnikom
  - Zasnovan na jednostavnom detektovanju jezičkih obrazaca pomoću ključnih reči i transliranju upita korisnika u izlaz mašine
  - Program imitirao odgovore psihoterapeuta
    - User_1: You are like my father in some ways.
    - ELIZA_1: WHAT RESEMBLANCE DO YOU SEE
    - User_2: You are not very aggressive but I think you don’t want me to notice that.
    - ELIZA_2: WHAT MAKES YOU THINK I AM NOT AGGRESSIVE
    - User_3: You don’t argue with me.
    - ELIZA_3: WHY DO YOU THINK I DON’T ARGUE WITH YOU
    - User_4: You are afraid of me.
    - ELIZA_4: DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU

Istorijat obrade prirodnih jezika

- Stohastička paradigma
  - 1964 – Bajesovske metode primenjene na zadatak određivanja autorstva Federalističkih spisa
  - 1963/64 – Brown Corpus of American English – korpus od milion reči raspodeljenih po 500 tekstova izvučenih iz različitih žanrova (vesti, romani, akademski tekstovi,…)
  - Veliki uticaj na razvoj sistema za prepoznavanje govora
    - *Noisy channel* modeli
    - Skriveni Markovljevi lanci (engl. *Hidden Markov Models*)

Istorijat obrade prirodnih jezika

- 1950-tih i 60-tih godina puno pažnje poklanjano problemu mašinskog/automatskog prevođenja
  - Veliko interesovanje državnih i vojnih aktera u SAD-u zbog Hladnog rata
  - 1954 – Georgetown–IBM eksperiment – demonstracija automatskog prevođenja preko 60 rečenica sa ruskog na engleski
    - Autori tvrdili da će mašinsko prevođenje biti rešen problem za tri do pet godina
- Istraživači izuzetno potcenili kompleksnost problema
  - 1966 - zbog velikih obećanja, a slabih rezultata, američki savetodavni komitet za automatsku obradu jezika (ALPAC) dao negativno mišljenje o rezultatima istraživanja mašinskog prevođenja
    - Dovelo do dramatičnog pada finansiranja i zanimanja za ovu problematiku, sve do 1980-tih i pojave statističkih sistema za prevođenje
    - Posredno dovelo do usporenog razvoja u celoj oblasti NLP-a

Istorijat obrade prirodnih jezika

- 1970-tih godina – Modeli zasnovani na konačnim automatima (engl. *finite-state models*) i ručno pisanim pravilima postaju prominentni u fonološkim, morfološkim i sintaktičkim problemima
- 1968-70 – Teri Vinograd napravio SHRDLU – prvi sistem za razumevanje prirodnog jezika (engl. *natural language understanding*)
  - Simulacija robota koji se kreće u svetu blokova koje treba pomeriti
  - Prvi pokušao da u NLU sistem ugradi sveobuhvatnu gramatiku engleskog jezika
  - Sistem bio u stanju da razume do tada neviđen nivo kompleksnosti komandi
    - *Move the red block on top of the smaller green one*
  - Uspeh sistema demonstrirao da je problem parsiranja prirodnog jezika rešen u meri da se može preći na probleme semantike i analize diskursa

Istorijat obrade prirodnih jezika

- Kasnih 1980-tih godina – Probabilistički modeli i modeli zasnovani na upotrebi podataka se razvijaju za probleme iz sintakse i semantike
- Od 1990-tih godina na dalje – *„Statistička revolucija“*
  - Velike količine tekstualnih podataka postaju dostupne preko interneta
  - Porast brzine i memorijskih kapaciteta računara dovodi do komercijalne primene većeg broja NLP zadataka (prepoznavanje govora, provera pravopisa, dohvatanje informacija, itd.)
  - Modeli zasnovani na upotrebi podataka/mašinskom učenju, kao i metodologija evaluacije razvijena za takva rešenja, nameću se kao standard u većini NLP zadataka
  - Izuzetno intenzivan razvoj novih modela, konceptualizacija novih zadataka, i razvoj NLP rešenja i za jezike osim engleskog

Obrada prirodnih jezika danas

- 2002 – prvi radovi na temu analize sentimenta
- 2011 – *Apple Siri*
- 2011 – *IBM Watson* pobedio šampione u kvizu *Jeopardy!*
- Od oko 2013 na dalje – sve šira upotreba neuralnih mreža u NLP-u
  - 2013 – *word2vec*
  - 2016/2017 - *WaveNet*
  - 2017 – *Google Neural Machine Translation*
  - 2018/2019 – *BERT*
- Istraživanja se sve više kreću u pravcu multilingvalnih, kros-lingvalnih, i multimodalnih rešenja (kombinovanje analiza teksta i slike/videa)

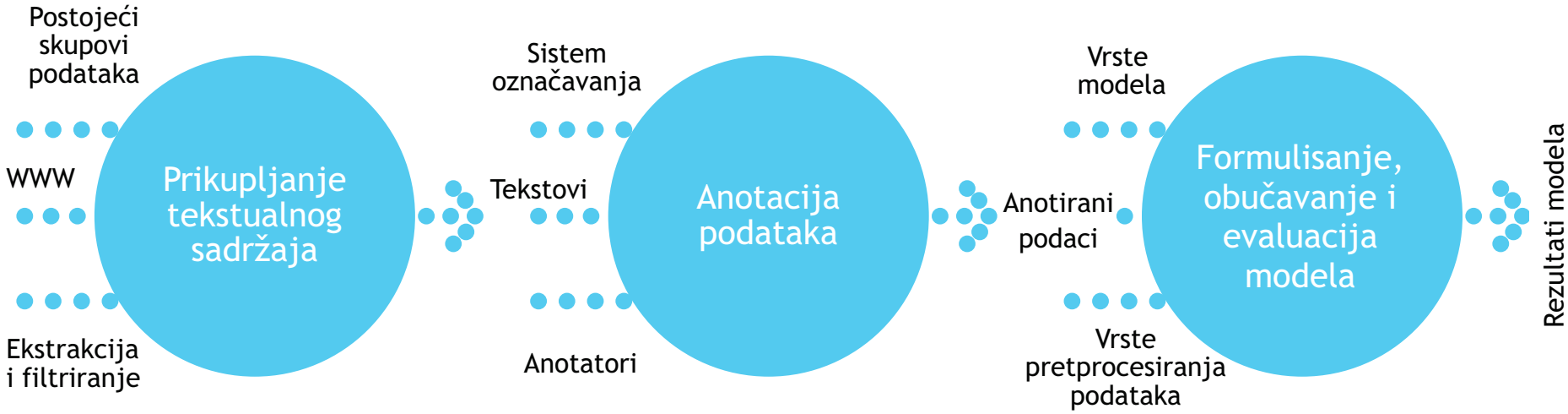
Konceptualni pristupi obradi prirodnih jezika

- Metode zasnovane na ručno sastavljenim pravilima i bazama znanja (engl. *rule-based NLP*)
  - Simboličko procesiranje, konačni automati, logičko rezonovanje
  - Generalno govoreći prilično tačne metode, ali veoma krhke
    - Loše performanse na primerima koji nisu pokriveni postojećim pravilima/bazama znanja
  - Izrada sveobuhvatnih skupova pravila/baza znanja za određenu problematiku je izuzetno naporna i često zahteva ekspertsko znanje
    - Jezik se vremenom menja – neophodno je stalno ručno ažurirati sistem
- Metode zasnovane na statistici i podacima (engl. *statistical NLP*)
  - Daleko robusnije i fleksibilnije, ali zahtevaju odgovarajuće podatke, obično anotirane, koji se koriste za uočavanje obrazaca
  - (Nadgledano) mašinsko učenje, teorija verovatnoće

> “Every time I fire a linguist, the performance of our speech recognition system goes up.”

Frederik Jelinek, istraživač automatskog prepoznavanja govora

- Ispravan lingvistički tretman nekog jezičkog fenomena ne mora nužno da dovede do željenih poboljšanja u smislu praktičnih efekata
- Sa druge strane, odsustvo lingvističkog znanja pri izradi rešenja ograničava obradu jezika na relativno rudimentarne/inženjerske zadatke



Dijagram faza u procesu razvoja statističkih rešenja za obradu prirodnih jezika

Slika preuzeta iz: V. Batanović, *Metodologija rešavanja semantičkih problema u obradi kratkih tekstova napisanih na prirodnim jezicima sa ograničenim resursima*, Elektrotehnički fakultet, Univerzitet u Beogradu, 2020.

Prikupljanje tekstualnog sadržaja

- Problem pronalaženja adekvatnih izvora tekstualnih podataka
  - Nekada je moguće iskoristiti digitalizovane kolekcije tekstova ili postojeće skupove podataka koji su razvijeni za rešavanje nekih srodnih problema
    - Češći slučaj za velike svetske jezike
  - Internet se često koristi kao izvor podataka
    - Treba voditi računa o tome da podaci nad kojima će se model kasnije primenjivati budu iste/slične prirode kao oni nad kojima se obučava
    - Tipično se koriste biblioteke za parsiranje i obradu HTML i *JavaScript* koda, kao što su: *JSoup*, *BeautifulSoup*, *Scrapy*, *HTMLUnit*,…
- Za manje jezike (poput srpskog) može biti teško pronaći dovoljne količine podataka
- Prikupljeni tekstovi se obično filtriraju na neki način pre dalje obrade

Anotacija podataka

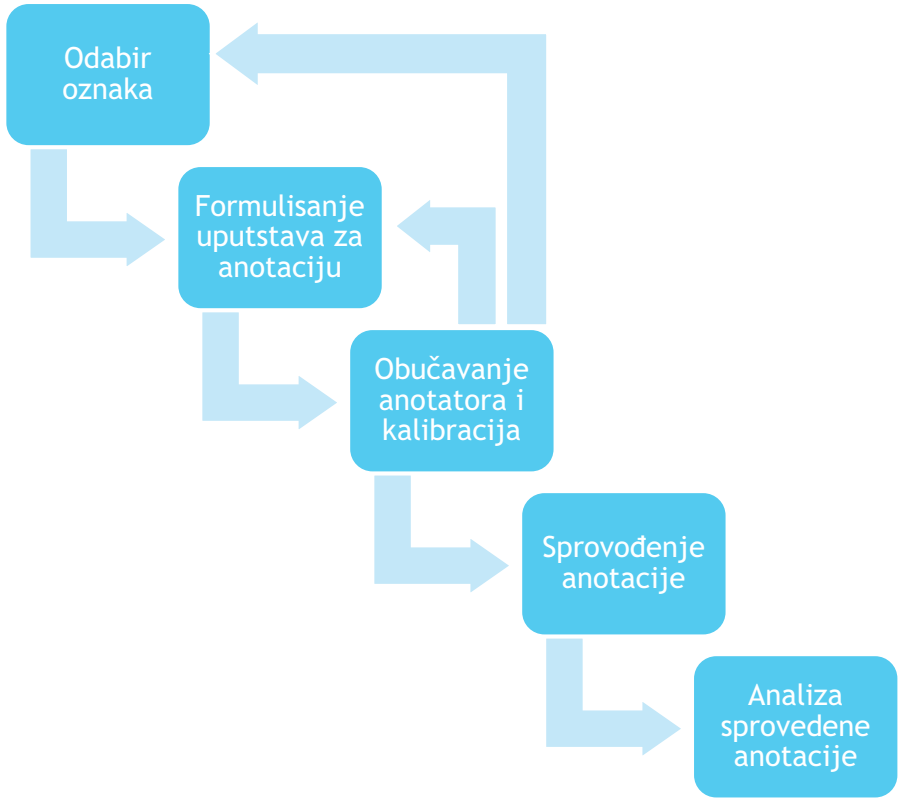
- Obeležavanje prikupljenih podataka korišćenjem određenog sistema označavanja vezanog za problem koji se rešava
  - Za neke zadatke (npr. analizu sentimenta) anotaciju mogu da vrše i ljudi koji nisu domenski eksperti
  - Za druge zadatke (npr. parsiranje) sprovođenje anotacije zahteva ekspertsko lingvističko znanje
- U oba slučaja insistiranje na kvalitetu anotacije je izuzetno važno da bi primena statističkih modela imala smisla
  - Pravljenje dovoljno detaljnih i jasnih uputstava za anotaciju
  - Odabir kvalitetnih anotatora

Anotacija podataka

- Gornja granica za performanse bilo kog računarskog sistema je stepen slaganja anotatora
  - Ako ljudi ne mogu da se slože šta je tačan odgovor u X% slučajeva, onda nema smisla očekivati da računarski sistem performira preko tog nivoa
  - Slaganje anotatora sa samim sobom – provera konzistentnosti anotacije
- Česte metrike za merenje stepena saglasnosti anotatora
  - Procentualna saglasnost anotatora
    - Problem: ne uzima u obzir stepen slučajne saglasnosti anotatora
  - Koenov (*Cohen's*) *kappa* i Skotov (*Scott's*) *pi* koeficijent
    - Uzimaju u obzir stepen slučajne saglasnosti anotatora
    - Koriste se za merenje saglasnosti u obeležavanju kategorijskih oznaka tj. klasa
    - Namenjeni određivanju saglasnosti dva anotatora

Anotacija podataka

- Česte metrike za merenje stepena saglasnosti anotatora
  - Flajsov (*Fleiss'*) *kappa* koeficijent
    - Proširenje Skotovog *pi* koeficijenta na više anotatora
  - Pirsonov (*Pearson*)/Spirmanov (*Spearman*) koeficijent korelacije
    - Koriste se za određivanje korelacije između dva skupa realnih numeričkih oznaka
    - Pirsonov koeficijent definiše korelaciju između samih numeričkih vrednosti
    - Spirmanov koeficijent definiše korelaciju između rangova numeričkih vrednosti
  - Kripendorfov (*Krippendorff's*) *alpha* koeficijent
    - Opšta i preporučena metrika za izražavanje stepena (ne)saglasnosti anotatora
    - Primenjiv na proizvoljan broj anotatora
    - Primenjiv na širok spektar različitih tipova oznaka: binarne, kategorijske, ordinalne, intervalske,…



Dijagram koraka u fazi anotacije podataka

Slika preuzeta iz: V. Batanović, *Metodologija rešavanja semantičkih problema u obradi kratkih tekstova napisanih na prirodnim jezicima sa ograničenim resursima*, Elektrotehnički fakultet, Univerzitet u Beogradu, 2020.

Koraci u fazi anotacije podataka

- 1. Odabir oznaka za obeležavanje podataka, zajedno sa njihovim definicijama i interpretacijama
  - Korišćenje standardizovanih oznaka za posmatrani problem, ako postoje
  - Razvoj novog sistema označavanja
- 2. Formulisanje uputstava za anotaciju
  - Izgled i obim uputstava može primetno varirati u zavisnosti od konkretnog problema koji se razmatra, usvojenog sistema oznaka, i konceptualnih odluka vezanih za željeni nivo detaljnosti uputstava
  - Detaljnija uputstva najčešće omogućavaju viši stepen saglasnosti anotacionih odluka između anotatora, tj. veću konzistentnost anotacije
    - Podiže kvalitet anotiranih resursa, po cenu nešto sporijeg procesa anotacije

Koraci u fazi anotacije podataka

- 3. Obučavanje anotatora i kalibracija
  - Upoznavanje anotatora sa uputstvima (i alatom za anotaciju, ako se koristi)
  - Kalibracija – anotacija probnog podskupa prikupljenih podataka
    - Razrešavanje nedoumica anotatora i dorada uputstava
- 4. Sprovođenje anotacije
  - Anotatori obično rade individualno
  - Jednostruka vs. višestruka anotacija
    - Finalne oznake u višestrukoj anotaciji kod kategoričkih ili strukturiranih oznaka – proces razrešavanja neslaganja (engl. *curation*)
- 5. Analiza sprovedene anotacije
  - Razmatranje konzistentnosti anotacije
  - Statistička analiza anotiranih podataka

Koraci u fazi anotacije podataka

- U praksi neretko dolazi do preklapanja između nekih od koraka
  - Obično se dešava kod složenijih anotacionih projekata, gde je teško unapred, pre detaljnog uvida u podatke, dizajnirati sasvim adekvatan skup oznaka i dovoljno sveobuhvatan set uputstava za anotaciju
  - Može doći i do cikličnog povezivanja prva tri koraka
- Ponekad se anotacija podataka može preskočiti ili automatizovati
  - Postojeći indikatori u prikupljenom sadržaju se koriste kao oznake
    - Npr. numeričke ocene ili broj zvezdica u opisima/recenzijama se koriste kao oznake sentimenta
  - Ograničenja u pogledu fleksibilnosti oznaka – oznake u potpunosti određene prisutnim indikatorima
  - Automatizovane oznake mogu i da unesu šum u podatke

Formulisanje, obučavanje i evaluacija modela

- Statistički modeli zasnovani na nadgledanom mašinskom učenju
  - Ulaz modela: anotirani prikupljeni podaci
- Razmatranje različitih (varijanti) modela
- Razmatranje različitih tehnika pretprocesiranja podataka
- Tačan oblik modela zavisi od problema koji se razmatra, kao i od konceptualnog pristupa rešavanju tog problema