Obrada prirodnih jezika

Uvod

Obrada prirodnih jezika - NLP

- Nova zanimljiva oblast
- Ali i dosta praktičnog rada
- Puno detalja koji su bitni kao i teoretski koncepti
- Kada se pogledaju radovi i istraživanja u ovoj oblasti – sve je novo i zanimljivo, postoje inovativne ideje
- Ali je u pozadini dosta eksperimenata i rada sa podacima
- Zato je predmet organizovan kao kombinacija teorije i praktičnog rada

Predavanja/vežbe

- Boško Nikolić (nbosko@etf.bg.ac.rs)
- Vuk Batanović (vuk.batanovic@etf.bg.ac.rs)
- Rad na računarima
- Obrada praktičnih skupova podataka – korpus srpskog jezika
- Projekat:
  - Podela na grupe – 4-5 članova
  - Konkretan zadatak i rad, vrlo brzo definisan, tokom celog semestra izrada

Program predmeta

- Mašinsko učenje u obradi prirodnih jezika
- Generativni i diskriminativni modeli
- Modeli sekvenci
- Pregled morfoloških, sintaksnih i semantičkih problema u obradi prirodnih jezika
- Jezički modeli
- Stemovanje i lematizacija, parsiranje
- Klasifikacija tekstova na osnovu tematike i sentimenta
- Semantička sličnost
- Prepoznavanje imenovanih entiteta
- Pregled novih tehnologija

Osnovni NLP pipeline

- Tekst -> Tokenizacija -> Obrada -> Model -> Rezultat
- Tekst - ulazni podaci u sirovom obliku (rečenice, dokumenti, korisnički upiti)
- Tokenizacija - razbijanje teksta na manje jedinice (reči, fraze ili tokeni)
- Obrada - čišćenje i priprema podataka (uklanjanje stop reči, lematizacija, normalizacija)
- Model - primena NLP modela (npr. klasifikacija, prevođenje, analiza sentimenta)
- Rezultat – izlaz sistema (predikcija, oznaka, generisani tekst ili uvid)

Primer: Tokenizacija

- Ulaz: Ne volim ovu knjigu
- Tokeni: [Ne, volim, ovu, knjigu]

Primer: NER (imenovani entiteti)

- Rečenica: Beograd je glavni grad Srbije
- Beograd → Lokacija
- Srbije → Država

Primer: analiza sentimenta

- Ulaz: Ovaj film je bio iznenađujuće dobar
- Izlaz: Pozitivan sentiment

- Ulaz: Ovo je odličan proizvod
- Izlaz: Pozitivan sentiment

- Ne mogu da verujem koliko je ovo dobro
- Pozitivan sentiment uprkos negaciji

Uvod

- NLP (Natural Language Processing) se bavi proučavanjem metoda za računarsku obradu i interpretaciju tekstualnih podataka napisanih na nekom od prirodnih jezika
  - jezika koje ljudi koriste u međusobnoj komunikaciji (npr. srpski, engleski, itd.)
- Njihova struktura i semantika nisu direktno mašinski čitljivi
- Veliki stepen kompleksnosti, nejasnoće i dvosmislenosti/višesmislenosti u izražavanju
  - „Uskoro sledi poskupljenje struje.“
  - „Cena električne energije će ubrzo porasti.“

Uvod

- 1985. godine objavljeno manje od 500 radova
- 2000. godine oko 2000 radova
- 2014. godine skoro 4000 radova
- Elsevier - 2000. godine objavljeno oko 2400 radova iz obrade prirodnih jezika indeksiranih u Scopus bazi
- 2016. godine taj broj prestigao 9800
- Procena za 2025 godinu je izmedju 25000 i 30000

Procena broja radova

Godina	Procena broja radova	Komentar
2015	~3.000 – 5.000	početak deep learning ere
2016	~4.000 – 6.000	stabilan rast
2017	~6.000 – 8.000	nagli skok (DL modeli)
2018	~8.000 – 10.000	širenje NLP primena
2019	~10.000 – 13.000	Transformer revolucija
2020	~12.000 – 15.000	veliki skok (BERT, GPT)
2021	~15.000 – 20.000	eksplozija LLM istraživanja
2022	~18.000 – 25.000	industrija + akademija
2023	~20.000 – 28.000	stabilan visok nivo
2024	~22.000 – 30.000	blagi rast
2025	~25.000 – 30.000	saturacija + konsolidacija

Evolucija NLP metoda

- Rule-based: kontrola, teško održavanje
- Mašinsko učenje: fleksibilnost, potreba za podacima
- Duboko učenje: najbolji rezultati, kompleksnost

Razumevanje jezika

- Ideja o razumevanju prirodnog jezika je stara koliko i ljudsko razmišljanje o računarima i robotima
- Potreba da se uspostavi komunikacija i to čovek pomoću svog jezika
- Naučna fantastika – 1926. Metropolis – robot koji razume
- Želimo da računari rade za nas – onda želimo i da komuniciraju sa nama

Interpretacija računara

- Računari komuniciraju na mašinski način, a ne ljudski
- Nisu računari krivi – u programiranju zahtevamo preciznost, specifično znanje, API i značenje pojedinačnih termina – prirodni jezici nisu takvi
- “Prirodni jezici nisu pogodni za rad sa računarima”
  - XML, semantički veb, meniji, kombo boksovi
  - “Računari ne mogu da komuniciraju kao što ljudi međusobno komuniciraju”

Ciljevi obrade jezika

- Cilj NLP istraživanja može biti **složen** – uzeti bilo koji deo teksta i razumeti ga
- Razumeti poruku koju krije dati tekst
- Imati u realnom vremenu dijalog sa robotom
- Cilj NLP istraživanja može biti i dosta **jednostavniji** - context sensitive spelling correction – Office 2007-08 – jednostavno procesiranje jezika

Početak

- Od pedesetih – period Hladnog rata, pokušaji SAD i SSSR da se automatski prevode dokumenti sa jednog jezika na drugi
- Velika očekivanja, ali i velika razočarenja
- Mala procesna snaga, ali i nerazumevanje kako jezik stvarno funkcioniše
- Složena struktura prirodnih jezika
  - modelovanje pomoću mašina sa konačnim stanjima je neodgovarajuće

Početak

- Sistemi koji su tražili i zamenjivali reči
- Rečenica na ruskom jeziku – prepoznaju se reči i vrši se njihova zamena sa prvom koja je pronađena u rečniku kao mogući ekvivalnet
- Iz sadašnje perspektive mašinskog prevoda – ne tako dobri rezultati
- Veći broj detalja koji otežavaju proces razumevanja i prevođenja
- ALPEC izveštaj vlade SAD 1966 da se prekine rad na mašinskom prevođenju
- Smatra se da se radi o beznadežnom pokušaju i naglasak je na proučavanju osnovnih nauka i boljem razumevanju kako prirodni jezici u osnovi funkcionišu

Nastavak

- Pokušaji izvan SAD – Evropa i Japan
- Mali napredak, 90tih značajan oporavak u SAD
- Kasnih 90tih najviše interesovanja u oblastima parsiranja informacija
- Povećana potreba i interesovanje za mašinsko prevođenje
- Jači računarski sistemi, ali i veće količine podataka – milijarde reči u okviru podataka
- Razvoj računarske lingvistike
- Manja očekivanja – www/portali – interesuje nas osnovni prevod, ne detalji – ne savršen prevod

Dalje

- 90tih i dalje direktno prevođenje
- Tekst se podeli u reči, izvrši se njihova osnovna normalizacija, traži se odgovarajući prevod
- Želja da se dobije bolji rezultat
- Sintaksna analiza – sintaksna struktura za rečenice – sintaksno bazirano statističko mašinsko prevođenje
- Semantička struktura – univerzalna, omogućiti prevod sa jezika na jezik

Razlike

- Pojedinačno različiti jezici imaju sve vrste razlika, različite vrste koncepta i ne ponavljaju se u drugim jezicima
- Eskimi imaju dosta reči za reč sneg
- Opšte rešenje bi obuhvatilo izuzetke i kompleknost svakog pojedinačnog jezika i teško bi se realizovalo
- Zato se ne dekodira na potpuno isti način, već se koriste isti izvori informacija – isti tekst paralelno (UN)

Primena NLP-a

- Chatbot
- Prevod
- Pretraga
- Virtuelni asistenti

Teškoće

- Obrada jezika je inverzni problem – postoje elementi koji su na površini i koji se vide, ali je potrebno pronaći i konstruisati elemente koji čine osnovu
- Zvučni signal ili skup reči – potrebno je izvršiti rekonstrukciju rečenice i otkriti njeno značenje
- A prirodni jezici su veoma, veoma dvosmisleni
- Na primer veliki broj reči može biti i imenica i pridev, dve imenice je moguće spojiti u novu imenicu (mreža računara, ...)

Teškoće

- Ljudsko razumevanje prirodnog jezika suštinski zavisi od složene i suptilne upotrebe konteksta, ali i konteksta cele rečenice, konteksta diskusije u okviru koje je rečenica izrečena, kao i samog realnog sveta
- Tako čovek razume drugog čoveka
- Zato na neki način koristimo faktore verovatnoće i rezonovanje da probamo da simuliramo ljudsko razumevanje prirodnih jezika

Poboljšanje

- Mogućnost da se aproksimira znanje o jeziku i kako se jezik koristi
- Koja vrsta sadržaja je u pitanju, da bi se pravilno interpretirala rečenica
- Dosta se upotrebljavaju modeli verovatnoće
- Imamo neizvesno znanja, podatke – predviđanje ili razumevanje sadržaja

Govor

- Prepoznavanje govora – više ka elektrotehnici
- Najviše okrenuti metodama verovatnoće – kako obraditi signale
- Na sličan način prepoznavanje reči iz govora
- Kako primeniti slične ideje i na ostale faze obrade prirodnog jezika
- AT&T labs, Bell labs sada, i IBM istraživački centar – od govora ka razumevanju prirodnih jezika

Obrada teksta

- Prilikom procesuiranja prirodnog jezika potrebno je prethodno obraditi tekst - normalizacija.
- Normalizacije korišćenjem ekvivalentnih klasa, obrade velikih i malih slova, steming i lematizacija
- Često je potrebno da sva pojavljivanja iste reči imaju pre obrade isti oblik, zbog čega su naročito bitne obrade steming i lematizacija.

Mašinsko učenje

- Svi algoritmi mašinskog učenja imaju dve faze
- Prva faza je faza učenja, u kojoj algoritam na osnovu podataka za koje zna odgovor uči kako da klasifikuje
- Druga faza je predviđanje. U toj fazi algoritam dobija novi, nepoznati tekst i na osnovu trening podataka i određene analize teksta predviđa kojoj klasi dati dokument pripada

Mašinsko učenje

- Nadgledano mašinsko učenje - uvek moraju postojati anotirani trening podaci, na osnovu kojih algoritam uči kako da postupa kada vidi nove podatke, odnosno kako da ih klasifikuje
- Nenadgledano mašinsko učenje - algoritam koji nema podatke sa tačnim odgovorima na kojima može da uči, već radi predviđanje bez predhodnog učenja
- U slučaju nenadgledanog učenja, algoritam pokušava da nađe određenu strukturu neanottiranih podataka i na osnovu toga radi predviđanja.

Klasifikacija teksta

- Klasifikacija teksta predstavlja jedan od osnovnih problema mašinskog procesuiranja čovekovog govora
- Metode mašinske klasifikacije teksta se koriste:
  - u filterima neželjene pošte (spam filters)
  - u prepoznavanju autora teksta, kao i njegovog pola i godišta,
  - prepoznavanju teme teksta,
  - analizi sentimenta rečenica
  - pretraživači interneta

Analiza sentimenta

- Analiza sentimenta rečenica predstavlja specijalan slučaj problema klasifikacije teksta, jer može sadržati samo dve klase - pozitivna ili negativna konotacija:
  - Analiza kvaliteta određenih proizvoda (Google Product Search i Bing Shopping)
  - Komentare sa Twitter-a ili Facebook-a o određenoj temi i određuju raspoloženje korisnika ovih društvenih mreža prema prikupljenim temama
- Analiza sentimenta može imati i više klasa/nivoa (npr. 1-5 zvezdica)
- Ekstrakcija mišljenja (opinion extraction, opinion mining), analiza subjektivnosti (subjectivity analysis, subjectivity mining, sentiment mining).
- Analiza sentimenta daje uvid u stavove ljudi o određenoj temi ili prozvodu - rezultati izbora ili položaj proizvoda na tržištu.

Analiza sentimenta

- Prvi rad u oblasti mašinske sentiment analize se pojavio 1979.
- Počev od 2001. godine je počela nagla ekspanzija oblasti
- Veliki istraživački i komercijalni potencijal
- Faktori koji su uticali na ovu ekspanziju su:
  - Unapređenje metoda mašinskog učenja u procesuiranju prirodnog jezika i pribavljanju informacija
  - Dostupnost podataka na kojima algoritmi mašinskog učenja mogu da budu trenirani (internetu - sajtovi sa recenzijama)
  - Intelektualni izazovi i komercijalna i obaveštajna upotreba

Analiza sentimenta

- Postoje rečenice koje imaju jako pozitivan ili negativan sentiment, ali ga nije moguće utvrditi pomoću neke ključne reči
- Reči koje se pojavljuju u toj rečenici nisu izrazito pozitivne ili negativne, već se mogu javiti u bilo kom kontekstu.
- Zavisnost od konteksta i domena
- Ista rečenica u zavisnosti od konteksta može imati različit sentiment
- Na primer rečenica „Pročitajte knjigu”, u tekstu recenzije knjige imaće izrazito pozitivan sentiment,
- U kontekstu recenzije filma negativan sentiment

Obrada negacije

- Većina algoritama za obradu teksta se oslanja na model vreće reči, gde su reči nezavisno posmatrane i analizirane - rečenice „Volim ovu knjigu” i „Ne volim ovu knjigu” su jako bliske u ovakvim analizama
- NE - suprostavljene klase - ne postoji u klasičnom pribavljanju informacija
- „Ne mogu da verujem kako je ovo dobro” - pozitivna rečenica, iako sadrži negaciju

Ironija i sarkazam

- Ove stilske figure je teško mašinski detektovati.
- Čak i ukoliko rečenice ne sadrže ovakve retoričke alate, opet mogu sadržati određene reči koji će biti protumačeni kao ključne reči određene polarnosti, iako zapravo rečenica ima suprotnu polarnost
- „Film izbegava sve klišee i predvidivosti koje se mogu naći u Holivudskim filmovima”.
- „izbegava” - jak neočekivani reverzer sentimenta, teškoće za obradu

Koncepti

- Raniji razvoj NLP rešenja - ručno sastavljenim skupovima pravila (engl. rule-based NLP)
  - obimno angažovanje eksperata
  - izuzetno težak i spor proces
- Statistički pristupi (engl. statistical NLP), zasnovanim na mašinskom učenju
  - o skup primera u obliku parova (ulaz, tačan izlaz)
  - bez eksplicitnog objašnjavanja veza između konkretnih podataka i željenih predikcija
  - često iste ili slične algoritme moguće koristiti za rešavanje različitih problema
  - Razvoj hardvera i Web 2.0

Koncepti

- Prodor NLP modela zasnovanih na dubokom mašinskom učenju (engl. deep learning)
- Zahtevaju jako velike količine podataka za obučavanje, koje je teško prikupiti za manje jezike
- Poslednjih par godina se može primetiti pomak u razvoju i promovisanju višejezičnih modela
- Broj jezika koje takvi modeli podržavaju može znatno varirati - od samo nekoliko do maksimalno stotinak

Kako funkcioniše ChatGPT

- Transformer arhitektura
- Predviđa sledeću reč
- Korišćenje konteksta

Transformer modeli

- Attention mehanizam
- Paralelna obrada
- Bolje razumevanje konteksta

Transformers

- Transformer-i su generativni modeli dubokog učenja koji su se pojavili u obradi prirodnih jezika i postali su vrlo popularni u poslednjih nekoliko godina.
- Oni predstavljaju vrstu neuronske mreže koja koristi višestruke slojeve pažnje kako bi obradila i predstavila tekstove.
- Umesto tradicionalnih modela koji koriste rekurentne neuralne mreže (RNN) za obradu sekvencijalnih podataka, poput teksta, Transformer modeli koriste slojeve pažnje kako bi obradili sve ulazne reči istovremeno.
- Transformer modeli nisu ograničeni na sekvencijalni pristup obrade teksta, što im omogućava da nauče složene odnose između reči i rečenica u tekstu.
- NLP - automatsko prevodjenje, generisanje teksta, klasifikacija teksta i odgovaranje na pitanja
- I u drugim oblastima – obrada slike
- BERT (Bidirectional Encoder Representations from Transformers)

Primena

- U semantičke probleme u obradi prirodnih jezika spada veliki broj zadataka koji podrazumevaju ili za cilj imaju pravilno razumevanje značenja tekstova:
  - analiza sentimenta i emocija u tekstu,
  - određivanje semantičke sličnosti,
  - detekcija parafraza,
  - odgovaranje na pitanja,
  - dohvatanje informacija,
  - izrada sažetaka teksta,
  - uprošćavanje teksta,
  - mašinsko prevođenje,
  - zaključivanje na prirodnom jeziku

NLP u realnim sistemima – izazovi

- Performanse i latencija
  - Modeli (posebno LLM) su veliki i spori
  - Potreba za optimizacijom (batching, caching, distillation)
- Skalabilnost
  - Obrada velikog broja korisničkih zahteva u realnom vremenu
  - Cloud infrastruktura i distribuirani sistemi
- Kvalitet podataka
  - Šum, greške, različiti formati (ćirilica/latinica, sleng)
  - Nedostatak anotiranih podataka za manje jezike
- Bias i etički problemi
  - Modeli mogu reflektovati pristrasnosti iz podataka
  - Važnost kontrole i evaluacije izlaza
- Robusnost sistema
  - Neočekivani inputi (typo, ironija, sarkazam)
  - Edge-case scenariji
- Integracija u softverske sisteme
  - API dizajn, mikroservisi, monitoring
  - Verzije modela i kontinuirano unapređenje
- NLP nije samo model – već kompletan softverski sistem sa realnim ograničenjima

Naši projekti

- **COMtext.SR**
  - Projekat razvoja osnovnog skupa resursa i alata za automatsku obradu tekstova na srpskom jeziku, kako za ekavicu tako i za ijekavicu, koji će biti javno dostupni pod licencom koja omogućava njihovu upotrebu u bilo koje svrhe, uključujući komercijalne
  - Fokus na domenima tekstova koji do sada nisu razmatrani bilo u akademskim bilo u komercijalnim javno dostupnim resursima i alatima za srpski jezik, kao što su pravno-administrativni, finansijski, medicinski, itd.
  - Projekat sprovodi konzorcijum Inovacionog centra ETF i ReLDI centra za jezičke podatke, uz finansijsku podršku domaćih i stranih kompanija i fondacija
  - Projekat otpočeo 2023. godine - prvi cilj: poboljšanje pretrage teksta u pravno-administrativnim dokumentima
    - To se ostvaruje rešavanjem problema morfosintaktičkog opisivanja i lematizacije tekstova
  - U 2024. godini projekat je usmeren na problem prepoznavanja imenovanih entiteta u pravno-administrativnim tekstovima
  - Više informacija i svi rezultati projekta su javno dostupni: https://github.com/ICEF-NLP/COMtext.SR

Naši projekti

- **STOP - Softver za prevenciju tekstualnih uvreda na srpskom jeziku: Otkrivanje govora mržnje pomoću veštačke inteligencije (Software for Text Offences Prevention in Serbian: AI-driven Hate Speech Detection)**
  - Program za izvrsne projekte mladih istraživača – PROMIS (finansiran od Fonda za nauku Republike Srbije)
  - Novi softverski sistem koji će detektovati govor mržnje na srpskom jeziku i koji će biti od velikog značaja u sprečavanju digitalnog nasilja
  - Formiranje skupova anotiranih podataka i kreiranje softverskog sistema za detektovanje govora mržnje na srpskom jeziku.
  - Realizovaće se nove metode za analizu kratkih tekstova primenom tehnika mašinskog učenja i veštačke inteligencije.
  - Analiziraće se detektovanje govora mržnje u prikupljenim podacima na srpskom jeziku.
  - Novi NLP modeli za srpski jezik

Naši projekti

- **AVANTES - „Nova rešenja u razvoju softvera zasnovana na sličnosti tekstova“**
  - Nacionalni projekat, Fonda za nauku Republike Srbije
  - Smanjiti jaz između prirodnih i programskih jezika
  - Semantička sličnost tekstova različitih dužina (engl. cross-level semantic similarity) – srpski i engleski, dva domena –novinski tekstovi i komentari u programskom kodu
  - Semantička pretraga koda (engl. semantic code search) –pronaći deo koda (modul/klasu/funkciju) koja semantički odgovara upitu na prirodnom jeziku – srpski i engleski

Naši projekti

- **Automatizacija Rapid Integrated Assessment procedure na srpskom jeziku – UNDP projekat**
  - Rapid Integrated Assessment(RIA) –pronalaženje delova pravnih dokumenata (zakona, strategija, akcionih planova...) koji se odnose na ciljeve održivog razvoja UN (Sustainable Development Goals)
  - Ručno sprovođenje RIA procedure zahteva čitanje i analizu hiljada stranica teksta i visok stepen ekspertskog znanja
  - Semantička pretrage pravnih dokumenata Republike Srbije (nestandardan format, ćirilica/latinica, morfologija)
  - Razvijeni sistem primenjiv i za semantičku pretragu drugih vrsta dokumenata