Otvoreni resursi i tehnologije za obradu srpskog jezika

Vuk Batanović, Nikola Ljubešić,

Tanja Samardžić, Maja Miličević Petrović

Otvoreni resursi u obradi prirodnih jezika

- Glavni resursi
  - (Ručno) anotirani korpusi tekstova
  - Alati i modeli za automatsku analizu i označavanje tekstova na prirodnim jezicima
- Otvorenost i javna dostupnost resursa dovode do:
  - Većeg stepena ponovljivosti rezultata istraživanja
  - Šire saradnje istraživača
  - Stimulisanja zajedničkog rada na unapređivanju postojećih resursa

Otvoreni resursi u obradi srpskog jezika

- Projekat ReLDI – *Regional Linguistic Data Initiative* (2015 – 2018)
  - Aktivnosti obuhvatale izradu i promovisanje javno dostupnih resursa za obradu južnoslovenskih jezika
  - Standardizovana metodologija anotacije podataka
  - Svi resursi objavljeni pod slobodnim licencama poput Creative Commons, GPL, i sl.
- ReLDI centar za jezičke podatke
  - NVO registrovana u Srbiji 2018. godine
  - Nastavlja aktivnosti započete u okviru projektnog partnerstva
  - Bliska saradnja sa sličnim centrima i inicijativama u regionu, poput CLARIN centra za južnoslovenske jezike (CLASSLA)

Anotirani korpusi tekstova na srpskom jeziku

- Ručno anotirani korpusi
  - Standardnog jezika
  - Nestandardnog jezika
  - Specijalizovani korpusi
- Veb korpus srpskog jezika
- Izrada mnogih korpusa je koordinisana sa sličnim poduhvatima na drugim jezicima (hrvatskom i slovenačkom)

SETimes.SR

- Ručno anotirani korpus tekstova iz novinskog domena pisanih standardnim srpskim jezikom
- Namenjen obučavanju i evaluaciji računarskih modela na većem broju problema iz obrade prirodnih jezika
- 163 dokumenta, 3891 rečenica, 86726 tokena
- Sadrži sledeće slojeve anotacije:
  - Segmentacija na dokumente, rečenice i tokene
  - Morfosintaktičke oznake po MULTEXT-East i Universal Dependency standardima
  - Leme
  - Sintaksne dependencije po Universal Dependency standardu
  - Imenovani entiteti (PER, DERIV-PER, LOC, ORG, MISC)

ReLDI-NormTagNER-sr

- Ručno anotirani korpus tvitova na srpskom jeziku
- Namenjen prilagođavanju računarskih modela fenomenima koji su česti u nestandardnom jeziku koji se koristi na internetu
- 3748 tvitova, 91 781 tokena
- Sadrži sledeće slojeve anotacije:
  - Segmentacija na rečenice i tokene
  - Normalizacija na nivou reči
  - Morfosintaktičke oznake po MULTEXT-East i Universal Dependency standardima
  - Leme
  - Imenovani entiteti (istih 5 tipova entiteta kao i u *SETimes.SR*)

STS.news.sr

- Specijalizovani ručno anotirani korpus za problem određivanja semantičke sličnosti kratkih tekstova
- 1192 para rečenica iz novinskog domena
- Granulirane ocene sličnosti na skali 0 – 5, dobijene usrednjavanjem ocena petoro anotatora
- Sadržaj korpusa dobijen iz ranijeg *paraphrase.sr* korpusa
  - Namenjen problemu detekcije parafraza
  - Sadrži samo binarne ocene sličnosti zadate od strane jednog anotatora

Serbian Movie Review Dataset – SerbMR

- Izbalansiran korpus filmskih recenzija razvijen za problem određivanja polarnosti sentimenta dokumenata
- Oznake sentimenta dobijene automatski, konverzijom numeričkih ocena zadatih od strane recenzenata
- Dostupan u dve varijante:
  - *SerbMR-2C* – dve klase (pozitivna i negativna) – ukupno 1682 recenzije
  - *SerbMR-3C* – tri klase (pozitivna, neutralna i negativna) – ukupno 2523 recenzije

SentiComments.SR

- Korpus od 3490 kratkih komentara iz domena filmova
- Ručno anotiran oznakama sentimenta koje omogućavaju više nivoa interpretacije
- Može se koristiti za veći broj problema u analizi sentimenta, poput određivanja polarnosti, određivanja subjektivnosti, detekcije sarkazma, itd.

Veb korpus *srWaC*

- Najveći javno dostupni korpus tekstova opšteg tipa na srpskom jeziku
- 555 miliona tokena, preko 25 miliona rečenica, oko 1,3 miliona dokumenata
- Izgrađen prikupljanjem celokupnog sadržaja sa *.rs* domena
- Sprovedene obrade:
  - Uklanjanje duplikata na nivou pasusa
  - Vraćanje dijakritičkih oznaka
  - Automatsko morfosintaktičko označavanje i lematizacija

Alati i modeli za obradu tekstova na srpskom jeziku

- Trenutno su dostupni alati i modeli za sledeće vrste obrade tekstova:
  - Tokenizaciju
  - Vraćanje dijakritičkih oznaka
  - Morfosintaktičko označavanje (po MULTEXT-East i Universal Dependencies standardima)
  - Morfološku normalizaciju
    - Stemovanje
    - Lematizaciju
  - Sintaktičko parsiranje (po Universal Dependencies standardu)
  - Označavanje imenovanih entiteta (PER, DERIV-PER, LOC, ORG, MISC)
  - Određivanje semantičke sličnosti kratkih tekstova

CLASSLA paket alata

- Proširena i izmenjena verzija Stanford *Stanza* biblioteke
- Trenutno najbolje javno dostupno rešenje za:
  - Tokenizaciju
  - Morfosintaktičko označavanje (po MULTEXT-East i Universal Dependencies standardima)
  - Lematizaciju (uz pomoć flektivnog leksikona *srLex*)
  - Sintaktičko parsiranje (po Universal Dependencies standardu)
  - Označavanje imenovanih entiteta (pet već pomenutih vrsta entiteta)
- Paket sadrži odvojene modele za standardni i za nestandardni jezik
  - Modeli za nestandardni jezik su dosta otporni na izostanak dijakritičkih oznaka u tekstu
- Pored srpskog podržava i hrvatski, slovenački i bugarski jezik

Ostali alati i modeli za obradu tekstova na srpskom jeziku

- ReLDI tokenizator – uključen u CLASSLA paket, moguće je i koristiti ga samostalno
- Statistički alat za vraćanje dijakritičkih oznaka koji podržava srpski, hrvatski i slovenački jezik
- *SCStemmers* – paket koji obuhvata 4 algoritma za stemovanje tekstova na srpskom ili hrvatskom jeziku
- *STSFineGrain* – paket modela za određivanje semantičke sličnosti kratkih tekstova
  - Moguće je obučiti modele na osnovu *STS.news.sr* ili nekog sličnog korpusa na bilo kom jeziku

ReLDIanno veb servis

- Omogućava jednostavno korišćenje jezičkih alata za srpski, hrvatski i slovenački jezik
- Uključuje alate za:
  - Tokenizaciju
  - Morfosintaktičko označavanje
  - Lematizaciju
  - Sintaktičko parsiranje
  - Označavanje imenovanih entiteta
- Servisima se može pristupiti putem veb aplikacije ili kroz Python biblioteku
- Veb aplikacija podržava različite formate fajlova: TXT, DOCX, PDF i ZIP

Dalji razvoj otvorenih resursa i tehnologija za obradu srpskog jezika

- Osim tokenizatora, alati koji se trenutno nalaze u pozadini *ReLDIanno* servisa su starija generacija otvorenih alata – radi se na njihovoj zameni CLASSLA paketom
- Proširivanje skupa otvorenih resursa novim vrstama anotacija (npr. koreferentnih odnosa)
- Verovatan budući prelazak na alate zasnovane na velikim jezičkim modelima poput BERT-a

Slobodan pristup opisanim resursima

- ReLDI centar za jezičke podatke: https://reldi.spur.uzh.ch/hr-sr/resursi-i-alati
- CLARIN.SI repozitorijum: https://www.clarin.si
- Veb servis *ReLDIanno*: http://www.clarin.si/services/web
- CLASSLA (https://www.clarin.si/info/k-centre) paket alata:
  - GitHub: https://github.com/clarinsi/classla-stanfordnlp
  - Pypi: https://pypi.org/project/classla

Kontakt podaci

- Vuk Batanović, Inovacioni centar Elektrotehničkog fakulteta u Beogradu, vuk.batanovic@ic.etf.bg.ac.rs
- Nikola Ljubešić, Institut Jožef Stefan, Ljubljana, nikola.ljubesic@ijs.si
- Tanja Samardžić, Univerzitet u Cirihu, tanja.samardzic@uzh.ch
- Maja Miličević Petrović, Univerzitet u Beogradu – Filološki fakultet, m.milicevic@fil.bg.ac.rs