Predlog projekta iz predmeta
„Sistemi za istraživanje i analizu podataka“

Katarina-Glorija Grujić Andrija Cvejić Aleksandar Cvejić
E2-102/2019 E2-101/2019 E2-99/2019

1. Definicija problema:

Cilj projekta je automatsko labeliranje pravnih propisa (zakoni, pravilnici i slično) Republike Srbije. Neophodno je prvo prikupiti podatke, zatim izvršena filtracija nad prikupljenim propisima. Takođe, u okviru teksta propisa treba identifikovati delove rečenice. Delovi koji se tiču referenci prema drugim dokumentima posebno obratiti pažnju, jer dati propisi ne sadrže standardni format citiranja ili nisu jasno iskazani citati. Zatim sve identifikovane delove anotirati pomoću Akoma Ntoso standarda za anotaciju legalnih dokumenata.

2. Motivacija:

Iako se uglavnom propisi bave jednom temom, nekada nije dovoljno pročitati samo jedan propis za kreiranje jasne slike o regulativama, time dolazimo u situaciju da svaki građanin treba da u glavi drži u obzir više stotina propisa, gde većina ljudi nisu sposobni za to. U Republici Srbiji prilikom oglašavanjem pravnih propisa na internetu, dokumenti nisu anotirani u mašinsko čitljivom formatu. Prednost pravljenja mašinskog čitljivog dokumenta je mogučnost referenciranja (hyper linked) na druge dokumente, time dobijamo brz pristup spomenutim dokumentima. Pored toga, zakoni postaju obimniji i svakog dana se povećavaju, iz tog razloga nephodno je postojanje lakše pretrage nad njima. Takođe, pored silnih propisa neophodno je postojanje složenog mehanizma za filtraciju, gde se informacije prikazuju po raznim kategorijama, domenima i sadržajima čoveku razumljivim.

3. Relevantna literatura:

[1] de Andrade, G. C., de Paiva Oliveira, A., & Moreira, A. (2019). Ontological Semantic Annotation of an English Corpus Through Condition Random Fields. Information, 10(5), 171. https://doi.org/10.3390/info10050171

- Zadatak rada:

  Upotreba mašinskog učenja za semantičko anotiranje zasnovano na ontologijama. Konkretno, napravljen je model koji može da klasifikuje odabrane vrste najvišeg nivoa ontologije. Zahteva se odgovarajući nivo tačnosti da bi se moglo primenjivati u zadatku automatske semantičke anotacije.

- Metodologija:

  Prvobitno je selektovan koprus sa odrađenim specifikacijama. Odnosno, da ima veliko pokriće i dimenziju. Korišćen je korpus “Open American National Corpus” koji je predložen u relevatnoj literaturi ovog rada. OANC sadrži 8293 dokumenata koji u sebi pored teksta sadrže anotacije od: leksičkog, morfološkog i sintaksne prirode (npr. *eng. part-of-speech*). Anotacija je vršena pomoću automatskog alata, koji unosi greške u sistem.

  Izvršeno je korekcija nad korpusom od strane autora, nakon čega je izvršena standardizacija dokumenata. Koriste se dve faze u anotiraju. Prva faza koristi anotator baziran na pravilima (koji labelira u odnosu na klase iz korpusa), dok druga faza je korekcija anotacija nakon pregledanje dokumenata.

  Nakon pripremljenog korpusa, korišćen je CRF (Conditional Random Feilds) pristup.
- Validacija:

  Preciznost, odziv i F-1 mera.
- Rezultat:

  Evaluacija je vršena nad klasifikacijom od 11 klasa. Postignuti rezultati rada su sledeći: preciznost od 94.1%, odziv od 87.7% i F-1 mera od 90.7%
- Mišljenja:

  Rad se bavi istom temom, samo drugim domenom. Možemo iskoristiti isti pristup rešavanju problema nad našim problemom.

[2] Skeppstedt, M., Kvist, M., Nilsson, G. H., & Dalianis, H. (2014). Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. Journal of Biomedical Informatics, 49, 148–158. https://doi.org/10.1016/j.jbi.2014.01.012

- Zadatak rada:

  Rad ima dva zadatka. Prvi da poredi koliko dobro se NER (*name entity recognition*) koji je obučen na engleskom jeziku radi na švedskom jeziku. Drugi zadatak rada je da se izvrši kategorizacija između 4 kategorije. Podaci su beleške doktora, a klase koje su pokušavali da nađu sledeće: *Disorder, Finding, Pharmaceutical Drug, Body Structure.*

- Metodologija:

  Korpus je prvobitno anotiran i evaluiran. Nakon čega su odgovarajući atributi izabrani za NER model. Model je treniran koristeći izabrane atribute, nakon čega je izvršena evaluacija na validacionim podacima.

  Rad vrši pregled prethodnih modela koji koriste SVN, CRF i rule-based pristupe, odnosno kombinacija sa CRF i nekih od njih. Autori su se odlučili da koriste samo CRF, jer je ukazano da najbolje rezultate daje, doduše otvaraju pitanje da li korišćenje dodatne metode na izlaz od CRFa.

- Validacija:

  F1 mera
- Rezultat:

  Dobijen rezultat je očeivan od strane autora, odnosno da sistem radi bolje na engleskom jeziku. Greške su uzrokovane lošim radom NER modela, greške leminizatora medicinskih termina i manjak švedskog vokubulara. F1 mere dobijene za tražene klase su sledeće:
  - Disorder (poremećaj): 0.81
  - Finding (nalaz): 0.69
  - Pharmaceutical Drug (farmaceutska lek): 0.85
  - Body Structure (struktura tela): 0.85
  - Disorder + Finding: 0.78

- Mišljenja:

  Bavi se sličnim problemom, ista metadologija rešavanja problema. Rad razmatra upotrebe drugih metoda za rešavanja problema. Naime SVN koji dostiže rezultate sličim trenutno najboljim CRFu.

4. Skup podataka:

Skup podataka je napravljen putem automatskog skidanja (web scraping) sa web sajta. Podaci se nalaze na sajtu http://pravno-informacioni-sistem.rs/ koji sadrži propise Republike srbije. Inicijalni skup podataka se odnosi nad zakonima, nad kojima je isprobana ispravnost predloženih metodologija. Skup podataka u trenutku izrade rada sadrži 10 hiljada propisa od kojih su 723 zakona.

5. Metodologija:

Skup podataka je neophodno labelirati. Labeliranje će biti izvršeno ručno od strane članova tima. Zatim, podatke je neophodno pretprocesirati. Za pretprocesiranje

podataka (*lemmatization, stemming*), koristiće se biblioteka ReLDI (REGIONAL LINGUISTIC DATA INITIATIVE) koja sadrži rečnik srpskih reči.

Za anotiranje teksta biće isprobani algoritmi CRF (eng. *Conditional random field*) [1], SVM (eng. *Support vector machine*) [2]. Pored toga, biće isprobani drugi klasifikatori, kao što su KNN (eng. *K-nearest neighbour*), naivni Bajes (eng. *Naive Bayes*).

6. Metod evaluacije:

Skup podataka biće podeljen u train i test skup u razmeri 80:20. Koristiće se mera tačnosti kao metod evaluacije, koja predstavlja odnos broja tačno predviđenih primera i ukupnog broja primera. Obučavanje će biti izvršeno za više različitih train i test skupova podataka.

Rezultati će biti predstavljeni grafičkim putem i u vidu matrice konfuzije (eng. *confusion matrix*).