UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

Ljubica Peleksić

KLASIFIKACIJA OBOLELIH OD ALCHAJMEROVE BOLESTI NA OSNOVU ANALIZE SPONTANOG GOVORA

master rad

Beograd, 2021.

Mentor:

doc. dr Jelena Graovac, profesor
Univerzitet u Beogradu, Matematički fakultet

Članovi komisije:

prof. dr Gordana Pavlović-Lažetić

doc. dr Jovana Kovačević

Datum odbrane:

porodici

Sadržaj

Glava 1

Uvod

Glava 2

Podaci

U ovom poglavlju će biti predstavljeni podaci koji su korišćeni u radu za dobijanje rezultata, kao i tehnike obrade podataka. koje su vršene nad njima. Podaci korišćnjeni za rešavanje problema klasifikacije obolelih osoba od Alchajmerove bolesti su transkripti slobodnog govora dementnih osoba. Transkripti se prikupljaju kao audio ili video zapisi, a razgovori sa osobama nisu struktuirani niti imaju neki odredjeni sled. Na osobi koja intervjuise je da postavlja pitanja, dok intervjuisani odgovara. Osoba koja intervjuise takodje prati tok razgovora sa pitanjima, te nisu u svakom razgovoru ista. U svrhu poredjenja, pored podataka o osobama obolelim od demencije tipa Alchajmer, bili su potrebni i intervjui sa nedementnim starijim osobama. Intervjui sa osobama obolelim od Alchajmera su sakupljani u obliku video i audio zapisa koji su prikupljeni od osoba koje su koristile usluge dnevnog boravka za obolele od Alchajmera u Novom Sadu, jedine takve ustanove u Srbiji organizovane od strane Udruženja građana Alchajmer. Ovi intervjui su bili prikpljani tokom grupnih razgovora sa obolelim. Intervjui sa nedementnim starijim osobama su prikupljeni zahvaljujući više od 20 volontera studenata Biološkog fakulteta u Beogradu, krajem 2017. godine i u toku 2018. Od audio zapisa razgovora sa osobama obolelim od Alchajmera su kreirani transkripti od strane vise osoba, dok su studenti koji su učestvovali u prikupljanju podataka od dementnih starijih osoba zatim od audio zapisa kreirali tekstualne zapise. Svaki intervju je zapisan na dva načina i to kao originalan u kome je svaka reč napisana tačno onako kako je izgovorena, uključujući ponavljanja, nedovršene reči, greške u izgovoru, i drugi u kome su ispravljene sve greške u izgovoru. Kako su ovi intervjui sprovedeni grupno, prvi korak je bio da se razdvoje razgovori, tako da rečenice koje je izgovorila jedna osoba se nalaze u jednoj tekstualnoj datoteci koja nosi ime te osobe.. Reči osobe koja postavlja pitanja se ne

moraju naći u svakoj datoteci i ne predstavljaju podatke koji se obradjuju. Nakon toga transkripti su pažljivo obradjeni, tako da svaki bude ispravno podeljen na rečenice. Ovaj korak je bio krucijalan kako bi u narednom koraku moglo da se dodje do vrsta reči i lema, za svaku izgovorenu reč.

Transkripti su zapisivani po definisanom protokolu koji ubraja sledeće:

1. Intervju sa jednom osobom se nalazi u datoteci koja nosi ime te osobe
2. Ime obolelog se označava vitičastim zagradama
3. Pitanje postavljeno od strane osobe koja vodi intervju se označava uglastim zagradama
4. Koristi se UFT-8 kodna shema i latinica
5. Koriste se slova sa dijakriticima (č, ć, š, đ,...)
6. Pauze između izgovorenih reči se zapisuju odgovarajućim brojem crtica, gde svaka crtica predstavlja jedadan sekund pauze
7. Brojevi se zapisuju sa crticom izmedju, ako su višecifreni brojev

Nakon što su transkripti bili ispravno zapisani po protokolu, a rečenice ispravno podeljene, takve datoteke su prosledjene na Filološki Fakultet u Beogradu, gde su odradjeni za svaku reč algoritamskim putem vrsta reči i njena lema. Vrste reči biće upotrebljene u okviru rešavanja problema klasifikacije obolelih pacijenata od Alchajmerove bolesti leksičkim metodama. Lema reči se koristi pri računanju vrednosti vokabulara i vokabulara za reči izgovorene jednom. Ove metrike će biti pomenute u daljem radu, u sekciji koja se bavi rešavanjem problema metodama leksičke analize.

2.1 Problem odredjivanja vrsta reči

U daljem tekstu će biti objašnjen problem dodeljivanja vrsta reči sekvenci reči, kao i predstavljena dva algoritma za rešavanje ovog problema, Skriveni Markovljev Model (eng. Hidden Markov Model / HMM) i Uslovljena nasumična polja (eng. Conditional Random Fields/CRF). Proces pridruživanja vrsta reči rečima se naziva anotacija. Vrste reči koje postoje i njigove oznake na engleskom jeziku su sledece:

1. Pridev: ADJ

2. Apozicija: ADP
3. Prilog: ADV
4. Pomoćni: AUX
5. Veznici: CCONJ
6. Član: DET
7. Uzvik: INTJ
8. Imenica: NOUN
9. Broj: NUM
10. Rečca: PART
11. Zamenica: PRON
12. Vlastita imenica: PROPN
13. Znak interpunkcije: PUNCT
14. Veznik: SCONJ
15. Simbol: SYM
16. Glagol: VERB
17. Drugo: X

Metoda odredjivnja vrsta reči je kompleksan problem. Zadatak je takav da dodeljujemo svakoj reči xi u ulaznoj sekvenci reči labelu yi, tako da izlazna sekvenca Y ima istu dužinu kao ulazna sekvenca X. [2]. Reči imaju vise mogućih vrsta reči i zadatak ovog procesa je da se pronadje ispravna vrsta reči za tu situaciju. Za reči se bira takva vrsta koja je najverovatnija. Za mnoge je verovatnoća da pripadaju svim vrstama sem jednoj izuzetno mala, pa je lako odlučiti se.

Jedan od načina da se reši problem odredjivanja vrsta reči je koristeći Skriveni Markovljev model (eng. Hidden Markov Model / HMM). Skriveni Markovljev model je probabilistički sekvencni model koji za sekvencu jedinica izračunava raspodelu verovatnoće po mogućim sekvencama i bira najbolju opciju. Markovljev lanac, model koji nam govori o verovatnoćama sekvenci slučajnih promenljivih, ima pretpostavku

da ako želimo da predvidimo buduće stanje, jedino što je bitno je trenutno stanje. Markovljev lanac se grafički predstavlja grafom, gde su čvorovi grafa stanja, a grane predstavljaju verovatnoće. Suma svih vrednosti grana koje idu iz odredjenog čvora moraju biti 1. Markovljev lanac se oslanja na dogadjaje koji mogu da se posmatraju, dok u slučaju reči to nije moguće. Zato se koriste Skriveni Markovljev Model., koji poseduje skrivene promenljive. Zadatak odredjivanja skrivene sekvence promenljivih na osnovu observacija u modelu se naziva dekodiranje. Skriveni Markovljev Model počiva na dve pretpostavke. Prva je identična kao za Markovljev lanac, dok druga kaže da verovatnoća nekog stanja zavisi samo od stanja koje je proizvelo to stanje i ni jednog više. Algoritam za odredjivanje vrsta reči se sastaji iz matrice koja sadrži verovatnoće da se jedna vrsta reči nalazi posle druge i matrice koja sadrži verovatnoće da se odredjeni tag dodeli odredjenoj reči. Algoritam za dekodiranje za Skriveni Markovljev Model se naziva Viterbi algoritam. Viterbi algoritam prima dve matrice koje smo pomenuli, a vraća putanju kroz stanja Skrivenog Markovljevog Modela koja deeljuje najveću verovatnoću datoj sekvenci. [2].

Algoritam koji je prethodno prikazan ima problem sa nepoznatim recima, vlastitim imenima, akronimima, novim rečima. Algoritam usljovljenih nasumčnih polja nalazi način da iskoristi oredjene odlike reči, npr. veliko slovo ili prefiks ili sufiks reči, što je teško dodati u Skriveni Markovljev Model. Trenira se logaritamski linearan model. U modelu uslovljenih polja računamo verovatnoću svih vrsta reči u sekvenci, a ne pojedinačni vrstu jedne po jedne reči. Svako svojstvo se oslanja na vrstu reči prethodne i sledeće reči i na celu ulaznu sekvencu reči. Za zaključivanje se takodje koristi Viterbi algoritam da bi se odabrala najbolja sekvenca vrsta reči. [2].

2.2 Lematizacija

Proces lematizacije jeste onaj u kome se za svaku reč nalazi njena kanonska forma, tj. lema. U srpskom jeziku, za imenice lema je nominativ jednine, za glagole infinitiv, a za prideve nominativ jednine muškog roda. Takođe, proces lematizacije uključuje vraćanje rodne varijacije reči na njen predjašnji oblik. Postoje dva metoda za rešavanje problema lematizacije. Prvi da se prema obliku reči otklanja sufiks i nalazi njena lema. Ovaj oblik rešavanja ne daje tako dobre rezultate. Drugi pristup uključuje korišćenje skupa podataka koji za svaku reč, za svaku njenu moguću vrstu reči ima odredjenu lemu. Postoji mogućnost i kombinovanja ova dva pristupa.

2.3 Skup podataka pre i nekon obrade

Skup podataka koji će biti korišćen za eksperimente u ovom radu sadrži sakupljene podatke na već opisan način, koji su zatim obradjeni pomenutim tehinikama. Intervjua sa pacijentima obolelim od demencije Alchajmerovog tipa ima 22, koje nazivamo „pozitivni" i nalaze se u fascikli „P". „Negativnih", tj intervjua sa starijim licima koji nemaju utvrdjenu demenciju Alchajmerovog tipa ima 57 i oni se nalaze u fascikli „N". Nakon procesa odredjivanja vrsta reči i lema za svaku reč intervjua, za svaki transkript je kreiran novi dokument koji u sebi sadrzi potrebne informacije. Ako se u transkriptu nalazi rečenica:

Ja sam Bojana.

Onda će se u odgovarajućoj datoteci naći i sledeći redovi, gde u svakom prva reč označava izvorni oblik reči, druga vrstu, a treća lemu.

Ja VERB Ja sam AUX jesam Bojana PROPN Bojana . PUNCT .

Glava 3

Resavanje problema metodama leksicke analize

Glava 4

Resavanje problema metodama masinskog učenja

Glava 5

Rezultati

Glava 6

Zaključak

Glava 7

Test

Ovo je rečenica u kojoj se javlja citat [3]. Još jedan citat [1].

Bibliografija

[1] Yuri Gurevich and Saharon Shelah. Expected computation time for Hamiltonian path problem. *SIAM Journal on Computing*, 16:486–502, 1987.

[2] Daniel Jurafsky and James H. Martin. Sequence labeling for parts of speech and named entities. In *Speech and Language Processing*, 2020.

[3] Petar BLA BLA Petrović and Mika Mikić. Naučni rad. In Miloje Milojević, editor, *Konferencija iz matematike i računarstva*, 2015.

Biografija autora

Ljubica Peleksić (*Beograd, 18. novembar 1993.*) Ljubicina biografija