TakeLab

Laboratorij za analizu teksta i inženjerstvo znanja
Text Analysis and Knowledge Engineering Lab
Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Unska 3, 10000 Zagreb, Hrvatska

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 4276

Primjena nenadziranog strojnog učenja za akviziciju glagolskih razreda iz korpusa

Filip Čulinović

Zagreb, srpanj 2015.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA ZAVRŠNI RAD MODULA

Zagreb, 13. ožujka 2015.

ZAVRŠNI ZADATAK br. 4276

Pristupnik: **Filip Čulinović (0036472908)**
Studij: Računarstvo
Modul: Računarska znanost

Zadatak: **Primjena nenadziranog strojnog učenja za akviziciju glagolskih razreda iz korpusa**

Opis zadatka:

Glagoli su glavni nosioci značenja rečenice i stoga su od posebnog značaja za semantičku analizu teksta. Pritom su se vrlo korisnima pokazali leksičkosemantički resursi koji glagole grupiraju u sintaktičke i semantičke razrede (npr. FrameNet, VerbNet). Takvi resursi međutim postoje samo za manji broj jezika, a njihova je izrada skupa i dugotrajna. Zbog toga je u literaturi predloženo više postupaka za automatsku akviziciju glagolskih razreda iz korpusa. Većina takvih postupaka temelji se na nenadziranom strojnom učenju odnosno grupiranju.
U okviru završnoga rada potrebno je upoznati se s teorijskom podlogom za grupiranje glagola u glagolske razrede te odgovarajućim jezičnim resursima kao što su FrameNet i VerbNet. Proučiti postupke nenadziranog strojnog učenja, s naglaskom na postupke grupiranja, uključivo i mekog grupiranja, te proučiti postupke za vrednovanje grupiranja. Razraditi postupak za grupiranje glagola iz korpusa na hrvatskome jeziku u glagolske razrede prema sintaktičkim i semantičkim svojstvima glagola, po uzoru na postupak Kawahare i dr. (2014). Izgraditi i ručno označiti odgovarajući skup tekstnih podataka na hrvatskome jeziku za razvoj i ispitivanje postupka. Razviti programsku implementaciju postupka te ga primijeniti na hrvatski web-korpus. Provesti iscrpno eksperimentalno vrednovanje postupka, statističku obradu rezultata te analizu pogrešaka. Radu priložiti izvorni i izvršni kod razvijenog sustava, označene skupove podataka i potrebnu dokumentaciju te citirati korištenu literaturu.

Zadatak uručen pristupniku: 13. ožujka 2015.
Rok za predaju rada: 12. lipnja 2015.

Mentor:

Doc. dr. sc. Jan Šnajder

Djelovođa:

Doc. dr. sc. Tomislav Hrkać

Predsjednik odbora za završni rad modula:

Prof. dr. sc. Siniša Srbljić

Zahvaljujem svojoj obitelji na svoj podršci koju su mi pružili tokom dosadašnjeg školovanja

SADRŽAJ

1. Uvod

U ovom radu opisan je pristup problemu automatske akvizicije glagolskih razreda metodama nenadziranog strojnog učenja. Iako ljudima intuitivan, navedeni problem je težak zadatak u području računalne analize prirodnog jezika. Jedan od pristupa rješavanju je i nenadzirano strojno učenje zbog same opsežnosti prirodnih jezika te je takav pristup i opisan u ovom radu.

Analiza prirodnog jezika (engl. *natural language processing, NLP*) područje je računalne znanosti koje se bavi načinima na koje bismo mogli iskoristiti računala za razumijevanje i manipulaciju prirodnim jezicima. Analiza prirodnog jezika povezna je s velikim brojem znanstvenih disciplina kao što su lingvistika i umjetna inteligencija [2].

NLP sustav može krenuti od razine riječi da bi odredio morfološku strukturu i prirodu riječi kao što su frazemi, značenje riječi i sl. Viša razina kreće od rečenice gdje se određuje poredak riječi, gramatika, značenje cijele rečenice i sl. da bismo došli do konteksta općenitog okoliša ili domene. Dana riječ ili rečenica može imati specifično značenje ili konotaciju u određenom kontekstu ili domeni te može biti povezana s mnogim drugim riječima i rečenicama unutar istog konteksta [2].

Liddy (1998) i Feldman (1999) predlažu da bismo bili u mogućnosti razumijeti prirodne jezike, bitno je razlikovati sljedećih sedam međuzavisnih razina koje ljudi koriste za ekstrakciju značenja iz teksta ili govora: (1) fonetička ili fonološka razina, (2) morfološka razina, (3) leksička razina, (4) sintaksna razina, (5) semantička razina, (6) komunikacijska razina i (7) pragmatična razina.

Za svaku od ovih razina postoji cijelo područje istraživanja koje se bavi rješavanjem tih problema. U ovom radu poseban fokus je na semantičkoj razini i značenju glagola [8, 3].

Glagoli igraju glavnu ulogu u prenošenju značenja rečenice. Zbog toga je poznavanje značenja glagola esencijalno za analizu prirodnog jezika te se u tu svrhu koriste brojni leksički resursi poput glagolskih razreda [6].

U ovom radu značajni su pojmovi semantičkog okvira i glagolskih razreda. Seman-

tički okviri su istovjetni jednom značenju nekog glagola te će zbog toga neki glagoli biti reprezentirani samo jednim semantičkim okvirom dok će mnogi glagoli sadržavati više različitih semantičkih okvira. Glagolski razredi su grupe glagola, točnije semantičkih okvira glagola, koje nose isto ili slično značenje [6].

Uzmimo za primjer četiri rečenice:

- *Danas je pao avion.*
- *Vrane su padale iz leta.*
- *Pao je snijeg.*
- *U Tihi ocean srušio se zrakoplov.*

Tri rečenice sadrže glagol *pasti*, ali samo dvije spadaju u isti semantički okvir. Prva i druga rečenica spadaju u isti okvir jer su značenja prekida leta. Četvrta rečenica je zasebni semantički okvir istog značenja. Zbog toga bi prilikom svrstavanja u razrede te tri rečenice svrstali u isti glagolski razred.

S obzirom da prirodni jezici sadrže tisuće ili čak desetke tisuća glagola, broj parova kojima bi trebalo dodijeliti brojčanu sličnost broji se u milijunima, a potencijalno i u desecima ili stotinama milijuna parova. Kako semantika prirodnih jezika nije strogo definirana, ljudski označivači bi zbog svoje subjektivne procjene mogli značajno odstupati jedni od drugih. Iz tog razloga u ovom radu korišteno je *nenadzirano strojno učenje* koje će na temelju strukture rečenica u kojima su upotrijebljeni glagoli detektira njihovo značenje te ih grupira u slične razrede.

Primjene analize prirodnog jezika su brojne. Koristi se u područjima poput strojnog prevođenja, procesiranju i sumarizaciji teksta prirodnog jezika, korisničkim sučeljima, ekstrakciji informacija iz teksta, raspoznavanju govora, umjetnoj inteligenciji, ekspertnim sustavima i sl.[2]

Cilj koji se želi postići ovim radom je ekstrakcija glagolskih razreda iz korpusa za hrvatski jezik. Dobiveni razredi po sličnosti bili bi od značajne koristi u analizi sentimenta hrvatskog jezika, posebice jer je ovaj rad prvi ovakve vrste za hrvatski jezik.

Ekstrakcija je ostvarena dvostupanjskim grupiranjem. U prvom koraku grupiraju se konteksti glagola iz rečenica u semantičke okvire, a drugom se grupiraju semantički okviri svih glagola u glagolske razrede.

Nastavak ovog rada bavit će se opisom problema i radova slične tematike te kasnije dubljim opisom modela za ekstrakciju glagolskih razreda i detaljima njegove implementacije. Na kraju će biti opisana evaluacija rezultata ovog modela te komentar na dobivene rezultate.

2. Opis problema i slični radovi

Konkretan problem kojim se bavi ovaj rad je kako iz velikog broja rečenica ekstrahirati glagolske razrede. Ljudi tokom cijelog svog života svakodnevno komuniciraju te konstantno uče i nadopunjuju svoje znanje jezika. Time je razumijevanje jezika kod ljudi postalo podsvjesno i intuitivno. Model prikazan u ovom radu poučava računalo kako razlikovati glagole različitog značenja.

Uzmimo za primjer nekoliko rečenica:

- *Moja majka je primila pismo.*
- *U kuću je primio goste.*
- *Primio sam pohvalu od šefa.*
- *Marko je primio rođendanski paket.*

U ovim rečenicama nisu zastupljena sva moguća značenja glagola *primiti*. Ipak, ljudi već prilikom prvog čitanja mogu zaključiti da je glagol *primiti* u prvoj i četvrtoj rečenici istog značenja. Ljudima je također intuitivno jasna suptilna razlika između fizičkog objekta poput pisma i paketa i nematerijalnog čina poput pohvale.

Iz perspektive računala, ove rečenice su u potpunosti različite. Osim zajedničkog glagola, one ne sadrže iste riječi. Iz samog glagola, bez susjednih riječi, nemoguće je odrediti njegovo značenje. Iz tog razloga, da bi bilo moguće odrediti značenje glagola bitno je upamtiti i njegov kontekst – riječi uz koje se pojavljuje.

Ipak, nije nam dovoljno da se samo okolne riječi glagola podudaraju. Time bismo dobili potencijalno beskonačan broj semantičkih okvira dok smo svjesni da ih u stvarnosti ima samo nekolicina po glagolu. Računalo bi trebalo naučiti prepoznati sličan kontekst glagola. Na neki način trebalo bi kvantificirati sličnost riječi te pomoću toga odrediti sličnost riječi *pismo* i *paket*.

Već u ovom koraku možemo prepoznati dio zahtjevnosti ovog problema. Određivanje značenja i sličnosti glagola ovisi o značenju i sličnosti riječi koje se pojavljuju uz njega. Očito je da postoji uzajamna ovisnost između ovih problema.

Da bismo razmotrili sljedeće potencijalne probleme uzmimo za primjer ove rečenice:

- *Mrav je stao.*
- *Stao je na mrava.*
- *Snijeg je pao.*
- *Avion je pao.*

Iako se u prve dvije rečenice glagola *stati* nalaze gotovo iste riječi, one su različitog značenja. Iz ovog primjera možemo zaključiti da nije samo bitno koje riječi se pojavljuju, već i koja je njihova služba u rečenici. Značenje toga je da svaku rečenicu, osim izdvajanja riječi, treba i analizirati i razložiti na sintaksne dijelove što je zadatak sintaksnog parsera. S obzirom da određivanje glagolskih razreda ovisi o sintaksi rečenice, sam uspjeh modela za ekstrakciju glagolskih razreda ovisiti će i o preciznosti korištenog sintaksnog parsera.

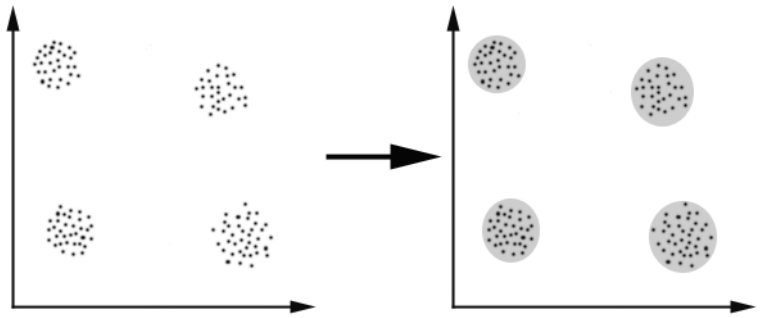
Uspoređujući rečenice glagola *stati* i one glagola *pasti*, dolazimo do sljedećeg problema. Prilikom analize rečenica glagola *pasti* možemo primijetiti da je jedina razlika u subjektu. On je razlikovna jedinica između značenja spuštanja padalina i prekida leta. S druge strane, za glagol *stati* u ovim primjerima nije toliko bitno za značenje glagola tko je subjekt. U tom primjeru (ne)postojanje prijedloga i objekta je razlika u značenju između prekida kretanja i gaženja. Time smo ustanovili da ne samo da je i služba riječi bitna u razlikovanju značenja, već je potrebno i na neki način kvantificirati doprinose elemenata službe riječi značenju glagola. Kako nije moguće unaprijed znati koji elementi službe riječi će najviše doprinijeti značenju glagola, potrebno je definirati neke univerzalne odrednice.

Prilikom analize značenja i grupiranja glagola moramo dakle uzeti u obzir: (1) riječi koje se pojavljuju u rečenici, (2) službu tih riječi u rečenici i (3) doprinos elemenata službe riječi.

2.1. Grupiranje

Grupiranje (engl. *clustering*) je, kao što mu i samo ime kaže, grupiranje sličnih objekata. Uzmimo u obzir kao bazu podataka sva viđenja manjih planeta. Velik broj takvih planeta nalazi se u orbiti između Marsa i Jupitera. Prvi manji planet je Ceres, otkriven 1801. godine od strane Piazzija i Gaussa. Uzimajući u obzir sliku u odnosu na fiksirane zvijezde, moguće je izračunati njihove orbitalne elemente. Astronomi često takve planete smatraju smetnjom prilikom promatranja drugih zanimljivih događaja.

Postoji oko 2000 takvih imenovanih planeta te tisuće opažanja tih, a moguće i drugih planeta. Bitan problem u praćenju manjih planeta je odluka koji od tih viđenja su viđenja istog planeta. Posebice, ako se tvrdi da je pronađen novi planet, mora biti provjereno da to opažanje nije zapazilo već poznati planet. Ovdje se primjenjuje grupiranje. Objekti su opažanja, dok su razredi u koje ih se svrstava imena tih planeta. Lako je opaziti da se takve klasifikacije događaju konstantno u mislima i govoru [4].



Slika 2.1: Grupiranje sličnih elemenata [1]

Grupa ili razred je skupina sličnih elemenata. Osnovni podaci u problemima grupiranja sastoje se od broja odluka o sličnosti skupa objekata. Kao standardan način za reprezentaciju sličnosti je preko skupa udaljenosti između parova objekata. Mnogi algoritmi za grupiranje koriste te udaljenosti te kreiraju grupe objekata unutar kojih su udaljenosti male. Odabir funkcije sličnosti nije manje bitan od odabira varijabli korištenih u algoritmu grupiranja. Razlog tome jest što je struktura grupiranja primitivnija od funkcije udaljenosti. Struktura grupiranja ima pristup udaljenostima među objektima, dok funkcija udaljenosti ima pristup karakteristikama objekata te je zbog toga informiranija o njihovoj prirodi [4].

Jedan od načina izračuna udaljenosti je i *euklidska udaljenost*. Uzmimo za primjer skup podataka u kojemu postoji M objekata i N varijabli. Vrijednost varijable J za objekt I definirana je kao $A(I, J)$.

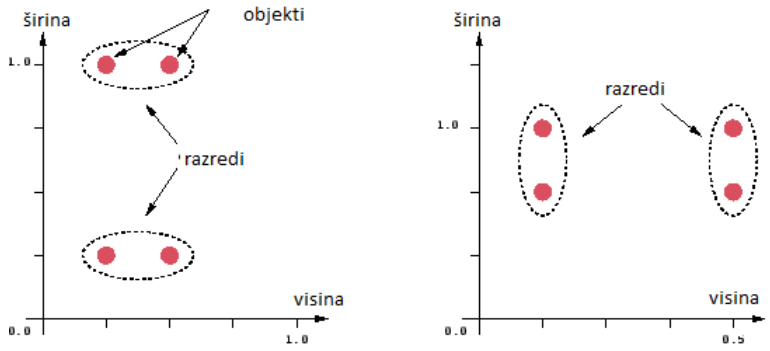
Definicija 2.1. *Euklidska udaljenost između objekata I i K definirana je kao:*

$$D(I,K) = (\sum_{I \leq J \leq N} [\, A(I,J) - A(K,J) \,]^2 \,)^{\frac{1}{2}}$$

.

U jednoj, dvije ili tri dimenzije, rezultat je samo ravna linija – udaljenost između vektora koji odgovaraju objektima *I* i *K*. U stvarnosti, često ćemo naići na slučajeve u

kojima sve varijable nemaju jednak doprinos udaljenosti. Uzmimo za primjer binarni klasifikator koji bi na temelju dobi, visine i težine probao pogoditi spol objekta. Već intuitivno možemo usporediti koliko značajniji bi doprinos dala visina u odnosu na dob ako bi bila mjerena u centimetrima. Također, ako je visina mjerena u metrima, a težina u kilogramima, doprinos težine bio bi drastično veći. U tu svrhu uvode se težine (engl. *weights*). Njihova je uloga povećati ili smanjiti doprinos neke varijable. Ovakva udaljenost nužna je kada su varijable mjerene različitim skalama ili jedinicama. Težine varijabli služe za skaliranje varijabli da bi njihove vrijednosti bile usporedive. Utjecaj skaliranja na grupiranje možemo vidjeti na slici 2.2.



Slika 2.2: Utjecaj skaliranja na grupiranje [1]

Da bismo uveli utjecaj težina na varijable u našem izračunu udaljenosti moramo modificirati dosadašnju formulu. Sada definiramo *težinsku euklidsku udaljenost*.

Definicija 2.2. *Težinska euklidska udaljenost između objekata I i K gdje $W(J)$ označava težinu varijable J definirana je kao:*

$$D(I,K) = (\sum_{I \leq J \leq N} W(J)[\,A(I,J) - A(K,J)\,]^2\,)^{\frac{1}{2}}$$

.

Ovaj oblik udaljenosti nije nužan ako su sve varijable mjerene istom skalom, ali ipak može poslužiti za utjecaj na doprinos varijabli po subjektivnim ili *apriornim* temeljima. Ako bismo za primjer uzeli izbore, vjerojatno bi veća težina bila dana rezultatima koji su bliže sadašnjosti ili bi različite težine bile dodijeljene prošlim lokalnim, državnim i parlamentarnim izborima [4].

Idealno, skaliranje bi trebalo biti učinjeno na način da su varijance unutar razreda približno jednake. Ovdje postoji osnovna zatvorena petlja:

1. Za grupiranje objekata, nužno je izraditi mjeru udaljenosti između objekata.

2. Za definiranje udaljenost, nužno je odrediti težine svake varijable.

3. Za određivanje težina varijable, nužno je poznavati razrede objekata da bi se varijance unutar razreda izjednačile.

Za potrebe ovog rada definirana je sličnost kao varijanta težinske euklidske udaljenosti između početnih i semantičkih okvira.

Markovljevo grupiranje (engl. *Markov cluster algorithm, MCL*) je algoritam grupiranja korišten u modelu razvijenom za potrebe ovog rada. On je efikasan algoritam grupiranja čvorova u grafu temeljen na simulaciji protoka u grafu. U ovom algoritmu veze između čvorova označavaju njihovu sličnost. Bitan pojam korišten u razvoju tog algoritma su *nasumične šetnje* (engl. *random walks*). Općenito, šetnjom se naziva bilo koji niz posječenih čvorova pri čemu postoji veza između susjednih čvorova. Nasumičnom šetnjom se naziva šetnja prilikom koje je vjerojatnost prelaska za svaki čvor V_j koji je povezan s čvorom V_i jednaka. Zbog neovisnosti budućih stanja o prošlima, svaka nasumična šetnja odgovara nekom Markovljevom lancu te iz toga slijedi i ime algoritma.

Također definiramo i Markovljevu matricu M_G čiji stupci odgovaraju normiranim vektorima stupaca matrice sličnosti G. Vjerojatnost da ćemo iz čvora i do čvora j doći preko čvora k iznosi $(M_G)_{i,k} \cdot (M_G)_{k,j}$. Time je vjerojatnost da neka šetnja koja kreće iz čvora i te završava u čvoru j lako izračunljiva kao suma prethodnog izraza za svaki potencijalni međučvor k. Potenciranjem Markovljeve matrice veze unutar grupa postići će veće vrijednosti od veza između grupa. Ovaj postupak je osnova Markovljevog grupiranja te se naziva širenje (engl. *expansion*).

Ovaj efekt se ipak gubi za velik broj koraka u šetnji. Iz tog razloga uvodi se postupak inflacije. Njegova funkcija je dodatno ojačati veze unutar grupa te oslabiti one između grupa. Ostvaren je na način da se elementi matrice potenciraju nekim faktorom k koji je realan nenegativn broj te se zatim svaki stupac ponovo normalizira.

Postupci širenja i inflacije vrše se naizmjence sve dok Markovljeva matrica ne konvergira, tj. dođe u stabilno stanje. Iz završnog stanja vrši se konstrukcija grupa.

2.2. Slični radovi

S obzirom na važnu ulogu glagola u značenju rečenica, glagolski razredi bili su tema mnogih znanstvenih radova.

U prošlosti, definirani su ručno izrađeni glagolski razredi poput Levinovih razreda [7] i njihove ekstenzije *VerbNet* leksikona [10] u kojemu su glagoli organizirani u

razrede prema njihovom semantičkom i sintaktičkom ponašanju.

Postoje mnogi pokušaji za automatsku akviziciju glagolskih razreda. U većini prethodnih radova o ekstrakciji glagolskih razreda prepostavljeno je da su glagoli jednoznačni, što nije realistično s obzirom da mnogi često korišteni glagoli nose više značenja.

Jedan od tih radova je *A general feature space for automatic verb classification* koji tretira sve glagole kao zasebne jedinice podataka te na njih primjenjuje klasifikaciju. Iako je u tom radu opisan pristup koji uzima u obzir sintaktičke značajke kao kriterij za klasifikaciju, i druge brojne značajke su uzete u obzir. Također, opisani pristup koristi klasifikaciju koja je metoda nadziranog strojnog učenja, dok je u ovom radu korišten pristup grupiranja što je nenadzirano strojno učenje. U zaključku je također navedeno da su njihovi eksperimenti pokazali da su sintaktičke značajke najinformativnije i najbitnije za određivanje glagolskih razreda, što je prihvaćeno kao temelj ovog rada [5].

Još jedan sličan rad je rad *Improving Verb Clustering with Automatically Acquired Selectional Preferences* iz 2009. godine. U tom radu opisan je pristup ekstrakciji glagolskih razreda pomoću grupiranja, ali još uvijek nije u obzir uzeta potencijalna višeznačnost glagola te je svaki glagol tretiran kao jedna zasebna jedinica [11].

U Kawahara (2014) opisan je pristup akviziciji glagolskih razreda koji uzimaju u obzir višeznačnost za engleski jezik. U tom radu korišten je dvostupanjski pristup akviziciji koji se temelji na dvostrukom grupiranju. Iz početnih okvira koji reprezentiraju značenje glagola u rečenici grupiranjem se dobivaju semantički okviri koji reprezentiraju općenita značenja glagola. Njihovim grupiranjem dobiveni su glagolski razredi. Ovaj rad temelji se na idejama iz tog članka. Ipak, postoje brojne razlike. Stvorene su strukture podataka koje odgovaraju pravilima hrvatskog jezika te služe kao reprezentacija značenja glagola. Također, korištene su različite metode za izračun sličnosti okvira i druga metoda grupiranja [6].

3. Model za ekstrakciju glagolskih razreda

U ovom poglavlju biti će objašnjena svojstva korpusa korištenog kao temelj ovog modela te pseudokod ostvarenja samog modela.

3.1. Korpus

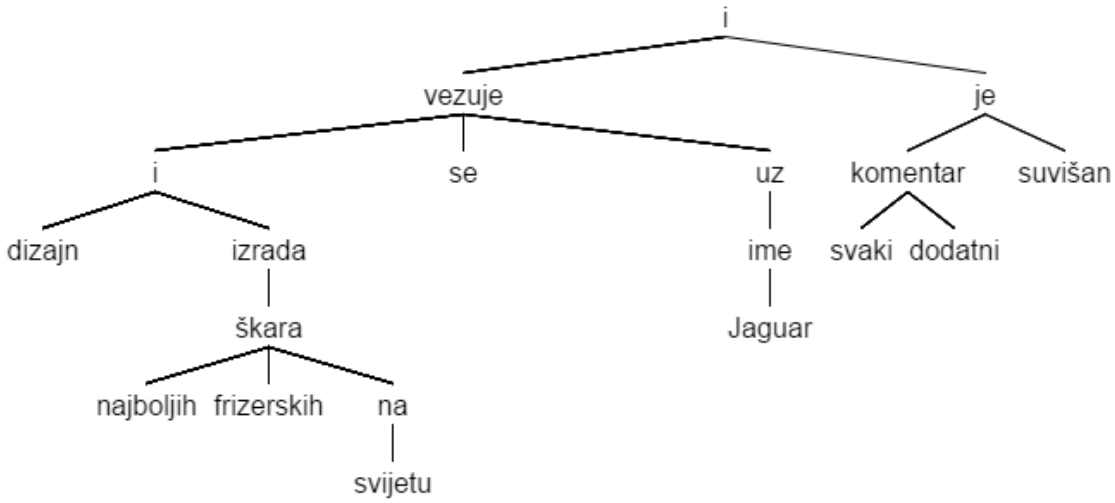
Za ekstrakciju glagolskih razreda potreban je skup rečenica hrvatskog jezika. Korpus korišten za ovaj rad već je analiziran pomoću sintaksnog parsera. Svaki redak korpusa sastoji se od jedne riječ rečenice, njezine leme, vrste i službe te riječi te identifikatora riječi roditelja u stablu.

Uzmimo za primjer rečenicu iz korpusa *"Dizajn i izrada najboljih frizerskih škara na svijetu vezuje se uz ime Jaguar i svaki dodatni komentar je suvišan."*.

Za tu rečenicu pomoću podataka iz korpusa definirano je sintaksno stablo sa slike 3.1. Jasno se mogu razaznati dvije podrečenice od kojih se složena rečenica sastoji te su te podrečenice podstabla glagola koji su predikati tih podrečenica.

Bitna značajka korpusa je to što sadrži leme riječi. Uzimajući u obzir pravila hrvatskog jezika i načine na koje se morfologija riječi mijenja ovisno o situaciji, lema riječi će se pamtiti umjesto same izvorne riječi te služiti kao nositelj značenja prilikom daljnje obrade.

Za prvu podrečenicu usporedno su prikazane riječi i njihove leme u tablici 3.1. Možemo primijetiti kako za neke riječi uopće nije bilo promjene, dok je za druge razlika značajna. Bitno za primijetiti je da su svi glagoli procesom lematizacije prebačeni u infinitiv što će značajno olakšati i ubrzati njihovo raspoznavanje.



Slika 3.1: Sintaksno stablo rečenice iz korpusa

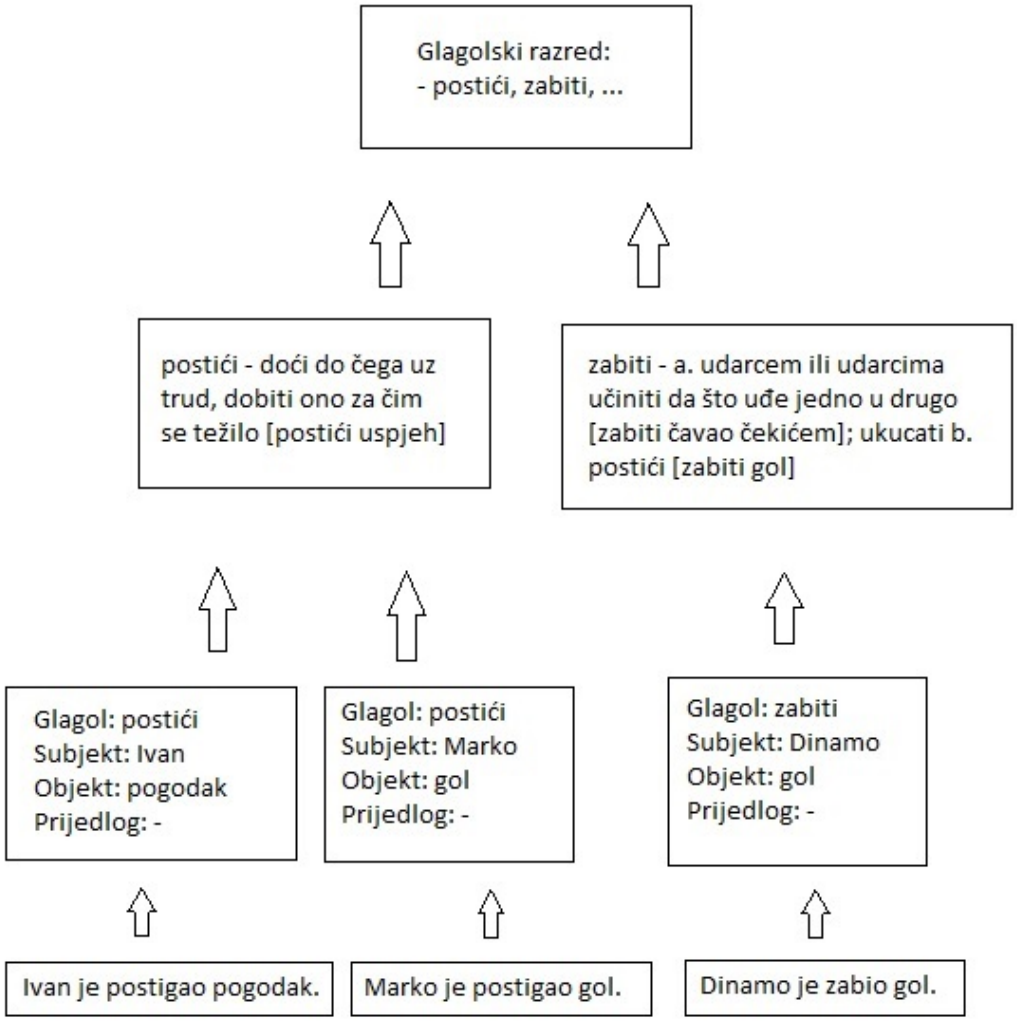
Riječ	**Lema riječi**
dizajn	dizajn
i	i
izrada	izrada
najboljih	dobar
frizerskih	frizerski
škara	škar
na	na
svijetu	svijet
vezuje	vezivati
se	sebe
uz	uz
ime	ime
Jaguar	jaguar

Tablica 3.1: Riječi i njihove leme

3.2. Pseudokod algoritma

Izvedba algoritma ekstrakcije glagolskih razreda prikazana je pseudokodom u Algoritmu 1. Sa slike **??** vidljivi su stupnjevi rada modela.

U pseudokodu se može primijetiti jasna trostupanjska struktura modela za ekstrakciju. U prvom koraku se za svaku rečenicu iz korpusa stvori potreban broj početnih



Slika 3.2: Prikaz stupnjeva u postupku

okvira. U drugom koraku se za svaki glagol grupiraju početni okviri te se spajaju u semantičke okvire. U trećem koraku semantički okviri se grupiraju u glagolske razrede.

3.2.1. Ekstrakcija početnih okvira

Prvi korak u modelu odnosi se na pripremu podataka. Njegov zadatak je prenijeti podatke iz stvarnog svijeta, u ovom slučaju rečenice, u programske strukture kojima je moguće manipulirati – okvire.

Algoritam ekstrakcije početnih okvira prikazan je pseudokodom u algoritmu 2.

Prije nego što se analizira svaka rečenica korpusa zasebno, zabilježeni su svi glagoli i njihov broj ponavljanja u korpusu.

Analiza svake rečenice kreće ispitivanjem njene duljine. Da bi analiza bila jed-

Algoritam 1 Ekstrakcija glagolskih razreda

for all $recenice$ iz $korpus$ **do**
    $okviri \leftarrow izvuciPocetneOkvire(recenica)$
    u $pocetniOkviri[glagol]$ dodaj $okviri$
for all $glagoli$ **do**
    $grupe \leftarrow grupiraj(pocetniOkviri[glagol])$
    for all $grupe$ **do**
        u $semantickiOkviri$ dodaj $spojiOkvire(grupa)$
$razredi \leftarrow grupiraj(semantickiOkviri)$

Algoritam 2 Ekstrakcija početnih okvira

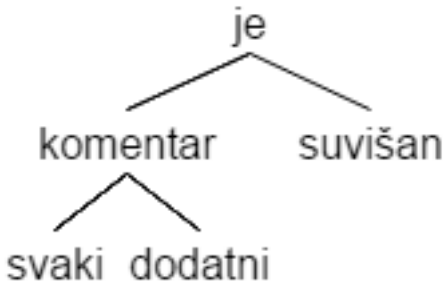
$brojPonavljanjaGlagola = prebrojiPonavljanja(korpus)$
$recenice \leftarrow dohvatiRecenice(korpus)$
for all $recenice$ **do**
    if $duljinaRecenice > maksimalnaDuljina$ **then**
        preskoci
    $stablo = izgradiStablo(recenica)$
    for all $glagoli$ iz $recenica$ **do**
        if $brojPojavljivanjaGlagola[glagol] < minimalanBroj$ **then**
            preskoci
        else
            $podstabloGlagola = izvuciPodstablo(stablo, glagol)$
            $okvir = stvoriOkvir(podstablo)$
            u $pocetniOkviri[glagol]$ dodaj $okvir$

nostavnija, postavljena je gornja granica za maksimalan broj riječi rečenice te se dulje rečenice ne uzimaju u obzir.

Zatim se pomoću podataka iz korpusa o sintaksnom stablu izgradi stablo te rečenice pomoću funkcije *izgradiStablo()*.

Potom se za svaki glagol iz rečenice radi zasebna analiza. Ukoliko je broj pojavljivanja glagola u korpusu manji od zadane donje granice, glagol se preskače. Razlog tome je što za glagole s manjim brojem pojavljivanja postoji mogućnost da nisu uhvaćena sva značenja. Također postoji mogućnost da premalen broj uzoraka ne daje reprezentativnu sliku o korištenju tog glagola i njegovom kontekstu.

Ukoliko je broj pojavljivanja glagola zadovoljavajuć, iz stabla rečenice se izvlači podstablo za glagol pomoću funkcije *izvuciPodstablo()*. Podstablo za rečenicu *"Dizajn i izrada najboljih frizerskih škara na svijetu vezuje se uz ime Jaguar i svaki dodatni komentar je suvišan."* čije je sintaksno stablo prikazano slikom 3.1 i glagol *biti* prikazano je na slici 3.3.



Slika 3.3: Sintaksno podstablo rečenice za glagol *biti*

Iz dobivenog podstabla pomoću funkcije *stvoriOkvir()* kreira se okvir te se ekstrahiraju riječi i poznati podaci. Taj okvir se zatim sprema u listu početnih okvira za taj glagol.

3.2.2. Stvaranje semantičkih okvira

Drugi korak u modelu odnosi se na grupiranje početnih okvira koji simboliziraju kontekst glagola u zasebnim rečenicama. Spajanjem okvira unutar tih grupa dobiveni su semantički okviri koji simboliziraju određeno značenje glagola.

Algoritam stvaranja semantičkih okvira prikazan je pseudokodom u algoritmu 3.

Za svaki glagol ekstrahiran iz korpusa radi se zasebna analiza. Prvo se dohvate svi početni okviri za taj glagol. Zatim se prebroji broj ponavljanja svakog zasebnog

Algoritam 3 Stvaranje semantičkih okvira

```
1: for all glagoli do
2:     pocetniOkviri ← dohvatiPocetneOkvire(glagol)
3:     brojPojavljivanja = prebrojiPojavljivanja()
4:     for all pocetniOkviri do
5:         if brojPojavljivanja[pocetniOkvir] < minimalanBroj then
6:             continue
7:         else
8:             u bitniOkviri dodaj pocetniOkvir
9:     semantickeGrupe ← grupiraj(bitniOkviri)
10:    semantickiOkviri = spojiOkvirePoGrupama(semantickeGrupe)
```

okvira. Potom se filtriraju svi nebitni početni okviri. To je ostvareno na način da se uspoređuje broj pojavljivanja tog okvira s donjom predefiniranom granicom. Razlog tome je što su za potrebe ovog rada zanimljivi upravo oni okviri (kontekst glagola) koji se učestalo pojavljuju, dok smo slobodni odbaciti okvire čiji je kontekst iznimka.

Nakon što je obavljeno filtriranje početnih okvira, semantičke grupe dobijemo funkcijom *grupiraj()*, koja će preko funkcije sličnosti okvira i algoritma grupiranja stvoriti grupe. Semantičke okvire, koji su cilj ovog koraka, dobit ćemo spajanjem početnih okvira unutar svake grupe pomoću funkcije *spojiOkvirePoGrupama()*.

3.2.3. Stvaranje glagolskih razreda

Treći korak ovog modela je ujedno i cilj ovog rada, a to je ekstrakcija glagolskih razreda. Pseudokod trećeg koraka nalazi se u algoritmu 4.

Algoritam 4 Stvaranje glagolskih razreda

```
1: for all glagoli do
2:     u semantickiOkviri dodaj dohvatiSemantickeOkvire(glagol)
3: glagolskiRazredi ← grupiraj(semantickiOkviri)
```

Treći korak najjednostavniji je od dosadašnjih koraka. Dohvate se semantički okviri svih glagola dobivenih u koraku 2. Zatim se odjednom svi semantički okviri grupiraju u razrede. Grupiranjem su dobiveni glagolski razredi koji se sastoje od semantičkih okvira različitih glagola, ali potencijalno i više semantičkih okvira istog glagola za koje je algoritam zaključio da po sličnosti pripadaju u isti razred.

3.2.4. Funkcija grupiraj

Funkcija grupiraj nositelj je glavne funkcionalnosti ovog modela. Njezin zadatak je pretvoriti strukturu podataka okvira u izvor podataka razumljiv algoritmu grupiranja te izvršiti samo grupiranje okvira.

Algoritam 5 Grupiraj

```
1: function GRUPIRAJ(okviri)
2:     matricaSlicnosti = izgradiMatricuSlicnosti(okviri)
3:     grupe = mclGrupiranje(matricaSlicnosti)
4:     return grupe
```

Funkcija *izgradiMatricuSlicnosti()* koristi funkciju *slicnost()* da bi izračunala sličnost dvaju okvira za sve parove iz liste okvira te ih sprema u matricu sličnosti.

Grupiranje pomoću matrice sličnosti ostvareno je algoritmom grupiranja MCL. Algoritam grupiranja MCL kratica je za Markovljev Algoritam Grupiranja (engl. *Markov Cluster Algorithm, MCL*). Taj algoritam je nenadzirani algoritam grupiranja unutar grafova koji se temelji na stohastičkoj simulaciji toka unutar grafova.[12]

Funkcija *slicnost()* izračunava sličnost dva okvira. Sličnost $A(X, Y)$ dva okvira X i Y definirana je kao:

$$A(X,Y) = \sum_{i=1}^{n} W_i * \mathfrak{c}(X(i) \cap Y(i)) \,/\, \min(\mathfrak{c}(X(i)),\ \mathfrak{c}(Y(i)))$$

gdje n označava broj atributa okvira, W_i označava težinu atributa i, a $\mathfrak{c}$ označava kardinalni broj skupa. Sličnost je time definirana kao varijanta *težinske euklidske udaljenosti* gdje je sličnost jednaka sumi skaliranog udjela preklapanja skupova atributa okvira.

4. Implementacija modela

Model za akviziciju glagolskih razreda opisan u prethodnom poglavlju implementiran je u programskom jeziku Python. Jezik je odabran radi njegove jednostavnosti te lakoće prijenosa ideja u programski kod. S obzirom da je određivanje glagolskih razreda kompleksan problem te je za njegovo rješavanje bilo potrebno osmisliti također kompleksan model, jednostavnost jezika bila je velika prednost prilikom odabira alata za implementaciju modela. Određivanje glagolskih razreda nije program koji bi se izvodio svakodnevno ili uopće često. Jednom kad se pokretanjem programa dobiju zadovoljavajući rezultati svrha programa je ispunjena. Kasnije se prilikom semantičke analize koriste samo rezultati implementacije ovog modela, ali se on više ne pokreće osim ako ne dolazi do nekih promjena u modelu ili korpusu. Iz tog razloga malo lošije performanse jezika Python nisu razlog za brigu.

Implementacija modela podijeljena je u četiri izvršne datoteke: (1) countVerbs.py, (2) corpusAnalysis.py, (3) semanticFrames.py te (4) verbClass.py.

Zadaci su podijeljeni u te četiri datoteke s obzirom na lančanu zavisnost. Izvršna datoteka *countVerbs.py* jednostavan je program čiji je zadatak prebrojati i zabilježiti glagol i broj njegovog ponavljanja u zadanom korpusu. Kao takav, neovisan je o sljedećem koraku – ekstrakciji početnih okvira te služi kao izvor podataka za taj korak.

Program *corpusAnalysis.py* snosi odgovornosti ekstrakcije početnih okvira opisanih u modelu. Program koristi podatke dobivene brojanjem glagola kao procjenu je li glagol dovoljno čest da bi njegova analiza dovela do korektnih rezultata.

Drugi korak u modelu – stvaranje semantičkih okvira, zadatak je programa *semanticFrames.py*. On kao ulazni skup podataka koristi okvire dobivene u prethodnom koraku.

Svrha modela i završni korak ostvaren je programom *verbClass.py*. Njegov zadatak je učitati sve semantičke okvire dobivene u drugom koraku te grupiranjem stvoriti glagolske okvire i zapisati rezultate.

4.1. Strukture podataka

Prilikom implementacije modela, prvi problem s kojim se trebalo suočiti je reprezentacija prirodnog jezika. S obzirom da je prethodno objašnjeno da utjecaj na značenje glagola nosi i služba riječi koje se nalaze uz glagol te je njihov doprinos različit, bilo je potrebno stvoriti strukturu podataka koja će omogućiti bilježenje svih tih podataka.

U tu svrhu stvoren je razred *Okvir*. Razred *Okvir* sastoji se od glagola čiji kontekst ili značenje pamti te od skupova objekata, subjekata i prijedloga koji su ekstrahirani iz njegovog konteksta.

Atribut	**Vrijednost**
glagol	čitati
subjekti	Marko
objekti	knjiga
prijedlozi	–

Tablica 4.1: Primjer početnog okvira

Za rečenicu *"Marko čita knjigu."* ekstrakcijom podataka iz podstabla kako je opisano u modelu, dobiven je početni okvir za glagol *čitati* kako je prikazan u tablici 4.1.

Atribut	**Vrijednost**
glagol	čitati
subjekti	Marko
objekti	knjiga
prijedlozi	–

+

Atribut	**Vrijednost**
glagol	čitati
subjekti	Ivan
objekti	pismo
prijedlozi	–

Atribut	**Vrijednost**
glagol	čitati
subjekti	Marko, Ivan
objekti	knjiga, pismo
prijedlozi	–

Tablica 4.2: Spajanje početnih okvira u semantički okvir

Ista struktura podataka pokazala se pogodnom i za prikaz semantičkih okvira. Semantički okviri stvoreni su na način tako da su dva početna okvira spojena po atribu-

tima u jedan objekt razreda *Okvir* koji simbolizira jedno značenje tog glagola. Takvo spajanje prikazano je tablicom 4.2. U njoj možemo vidjeti kako iz dva početna okvira, prvi koje je prethodno spomenut i drugi koji je dobiven iz rečenice *"Ivan čita pismo."*, je dobiven semantički okvir koji obuhvaća kontekste oba početna okvira. Zbog jasnoće primjera, ovdje u okvire nisu spremane leme, već cijele riječi.

Rezultati ovog rada – glagolski razredi – ostvareni su kao skupovi semantičkih okvira za koje je grupiranjem zaključeno da pripadaju istom razredu te im nisu dodijeljena posebna imena.

4.2. Performanse

Iako je prethodno zaključeno da performanse ove implementacije nisu u prvom planu, zadatak se pokazao pogodnim za neka jednostavna, ali značajna poboljšanja u performansama.

S obzirom da je analiza svake rečenice korpusa u potpunosti nezavisna od analize ostalih rečenica, uvedena je višedretvenost čiji je zadatak čitati i obraditi rečenicu. Broj dretvi u programu ovisi o broju jezgara procesora računala na kojem se program izvodi. Time su performanse programa značajno bolje na novijim računalima i posebice na poslužiteljima.

Da bi se olakšala kasnija analiza svakog zasebnog glagola objekti okvira spremani su u zasebne datoteke koje nose imena glagola. Svi objekti su prije spremanja serijalizirani te je tako ušteđen memorijski prostor te ubrzano rukovanje podacima u sljedećim koracima.

Uzimajući u obzir broj glagola te memorijska ograničenja računala, bilo je potrebno ograničiti učestalost pristupa programa memoriji. Spremanje objekata u datoteke nakon svake rečenice značajno bi usporilo rad programa, dok je ispis okvira nakon završene analize kompletnog korpusa spriječen nedovoljnim memorijskim resursima. Zbog toga u ovoj implementaciji, pristup memoriji, tj. ispis svih okvira događa se periodično – svakih 1% obrađenog korpusa.

U drugom koraku algoritma potrebno je bilo odbaciti okvire koji se ne pojavljuju određeni broj puta. U tom koraku, broj okvira je vrlo velik te svako ubrzanje prilikom filtriranja značajno doprinosi performansama ovog koraka. Poboljšanje je ostvareno pomoću strukture podataka rječnik (engl. *dictionary*) jezika Python čime je filtriranje svedeno na $O(n)$.

Glavni element ovog modela je algoritam grupiranja. Poboljšanje u njegovim performansama ostvareno je vezanjem algoritma na biblioteku *numpy* koja veže operacije

s matricama na implementacije u programskom jeziku C te time značajno ubrzava performanse algoritma.

Analiza korpusa reda veličine pedesetak milijuna rečenica potrajala je oko šesnaest sati. Otprilike isto vrijeme bilo je potrebno i za grupiranje početnih okvira svih glagola. Vrijeme utrošeno na obradu semantičkih okvira te ekstrakcije glagolskih razreda bilo je manje od sat vremena.

Sljedeći mogući korak koji bi bio značajan za performanse algoritma bio bi uvođenje višedretvenosti u drugi korak algoritma. S obzirom da se grupiranje početnih okvira u semantičke okvire radi neovisno za svaki glagol, zadatak je pogodan za uvođenje višedretvenosti gdje bi svaka dretva vršila obradu zasebnog glagola.

5. Evaluacija

Za potrebe evaluacije korišten je *hrWaC* korpus. *HrWaC* je web-korpus prikupljen sa stranica *.hr* domena. Trenutno je najveći dostupan korpus na hrvatskom jeziku. Podaci su preuzeti s 14396 domena. Cijeli korpus označen je lemama, morfosintaktičkim opisima te sintaksnom ovisnosti. Korpus se sastoji od 1.9 milijardi jedinica informacija (engl. *tokens*) iz 50940598 rečenica. Na slici 5.1 prikazan je primjer rečenice iz korpusa.[9]

```
1   Sve sve Q   Q   -   3   Aux -   -
2   se  sebe    P   P   -   3   Aux -   -
3   odvijalo    odvijati    V   V   -   0   Pred    -   -
4   unutar  unutar  S   S   -   3   Prep    -   -
5   <num>   24  M   M   -   6   Atr -   -
6   sata    sat N   N   -   4   Adv -   -
7   .   .   Z   Z   -   0   Punc    -   -
```

Slika 5.1: Primjer podataka iz korpusa

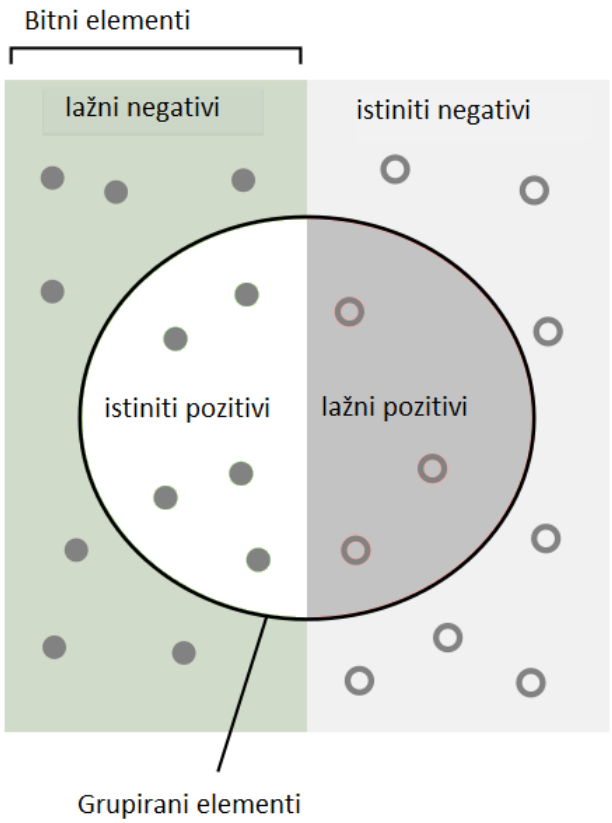
U prvom stupcu nalaze se riječi iz preuzetih rečenica. Ako je analizom utvrđeno da je riječ broj, ona je zamijenjena oznakom *<num>*. Drugi stupac sadrži leme riječi dobivene pomoću CST-ovog lematizatora. Treći stupac sadrži oznake vrste riječi, dok sedmi stupac označava njihovu službu u rečenici. Šesti stupac nosi indeks riječi roditelja u sintaksnom stablu rečenice.[9]

5.1. Preciznost i odziv

Evaluacija rada modela za ekstrakciju glagolskih razreda napravljena je pomoću dvije značajke vrednovanja grupiranja – preciznosti (engl. *precision*) i odziva (engl. *recall*).

Izračun obje karakteristike temelji se na četiri klasifikacije objekata: (1) istiniti pozitivi, (2) lažni pozitivi, (3) istiniti negativi i (4) lažni negativi.

Istiniti pozitivi (engl. *true positive, TP*) su objekti koji su uhvaćeni grupom te njoj uistinu i pripadaju. Lažni pozitivi (engl. *false positive, FP*) su objekti za koje je poznato da ne pripadaju grupi, ali su ipak svrstani u nju postpukom grupiranja. Istiniti negativi (engl. *true negative, TN*) su objekti za koje je poznato da ne pripadaju grupi te ih je algoritam grupiranja svrstao van grupe dok su lažni negativi (engl. *false negative, FN*) objekti za koje je poznato da bi trebali pripadati grupi dok ih je algoritam svrstao van grupe. Ove četiri mogućnosti prilikom grupiranja elemenata prikazane su na slici 5.2.



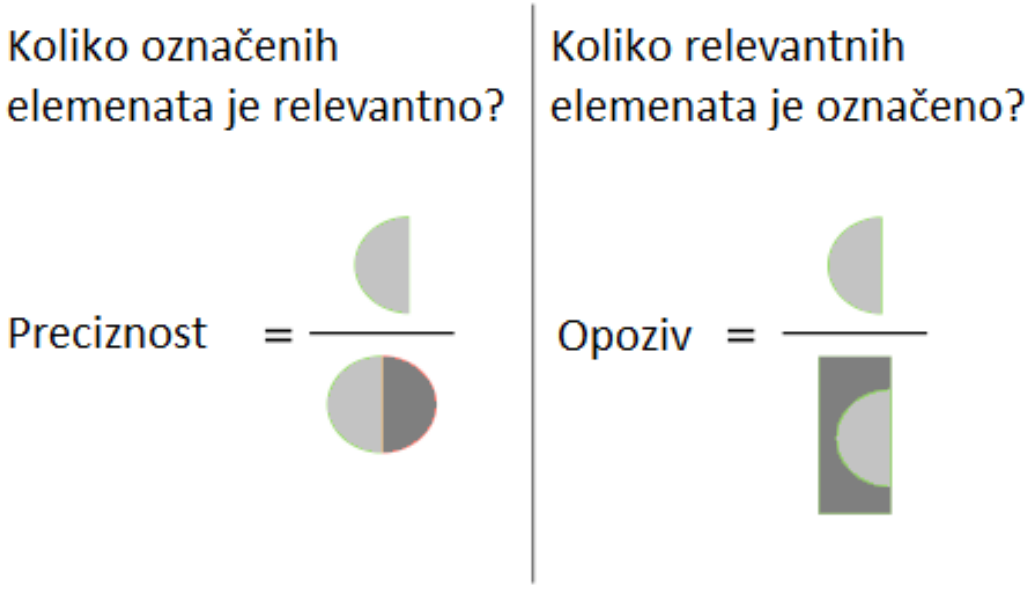
Slika 5.2: Primjer istinitih i lažnih pozitiva i negativa [13]

Preciznost i odziv definirani su sljedećim formulama:

$$Preciznost = \frac{TP}{TP + FP}$$

$$Odziv = \frac{TP}{TP + FN}$$

Njihova vizualna reprezentacija u odnosu na sliku 5.2 prikazana je na slici 5.3



Slika 5.3: Preciznost i odziv [13]

Primjer podataka za testiranje dan je u dodatku A. U njemu se nalaze početni okviri potrebni za testiranje grupiranja u semantičke okvire za glagol *zabiti*. Prvih deset okvira pripada jednom semantičkom okviru, dok drugih deset pripada drugom.

5.2. Rezultati

Primjer ispitnog skupa podataka naveden je u dodatku A. Korištenjem skupova ispitnih podataka poput toga za 14 glagola dobiveni su rezultati za stvaranje semantičkih okvira iz početnih prikazani u tablici 5.1.

Višestruki retci za isti glagol odgovaraju ispitnim primjerima za različita značenja glagola. Grupe za testiranje su one s maksimalnim brojem članova skupa testnih primjera za određeno značenje uz pretpostavku da takva grupa najbolje predstavlja to značenje.

Glagol	**Preciznost**	**Odziv**
dijeliti	0.5	0.818
	0.5	0.818

dobiti	0.5	0.9
	0.5	0.818
odbiti	0.529	0.9
	0.471	0.5
pasti	0.265	0.643
	0.235	0.421
	0.235	0.727
	0.265	0.692
pogoditi	1.0	0.692
postići	0.529	0.818
	0.471	0.727
primiti	0.214	0.9
	0.214	0.818
	0.214	0.75
	0.214	0.75
	0.143	0.75
spustiti	0.5	0.692
	0.5	0.643
srušiti	1.0	0.692
udariti	0.5	0.818
	0.5	0.9
uputiti	0.591	0.813
	0.409	0.692
valjati	1.0	0.333
vrijediti	0.346	0.9
	0.308	0.667
	0.346	0.563
zabiti	0.474	0.9
	0.526	1.0

Tablica 5.1: Rezultati ispitivanja

Iz rezultata je vidljivo da je odziv prilično dobar. Prosječan odziv kroz cijelo testiranje iznosi 0.742. S druge strane preciznost je lošija strana ovog modela. Prosječna

preciznost u koraku stvaranja semantičkih okvira bila je 0.452. Ti rezultati nam govore da iako je model dobro procijenio sličnost između okvira što je vidljivo iz odziva, ipak nije bio dovoljno precizan da razluči razlike između grupa što je vidljivo iz preciznosti. Omjer broja semantičkih okvira dobivenih u odnosu na pretpostavljeni broj značenja glagola prema Anićevu rječniku kroz cijeli ovaj korak iznosi 1.02. Iz toga je vidljivo da su parametri grupiranja podešeni na optimalan način te je algoritam grupiranja kvalitetno odredio broj razreda.

Primjer ovakvog grupiranja je glagol *zabiti*. Algoritam grupiranja smjestio je 9 od 10 ispitnih početnih okvira značenja 1 u istu grupu, ali je u tu istu grupu također stavio i svih 10 ispitnih početnih okvira značenja 2. Iz tablice 5.1 vidljivo je da je zbog toga odziv vrlo dobar, ali je preciznost loša. Način na koji bi se moglo pristupiti rješavanju tog problema bilo bi penaliziranje različitosti prilikom izračuna sličnosti okvira. Trenutno nema negativnih posljedica ako jedan okvir pokriva mnogo širi kontekst od drugog – ukoliko je jedan okvir podskup drugog, sličnost je maksimalna.

Testiranje završnog koraka modela – akvizicije glagolskih razreda izvršeno je na temelju tri ispitna glagolska razreda koji sadrže semantičke okvire dobivene u prethodnom koraku. U ovom koraku za preciznost i odziv ispitani su svi glagolski razredi koji sadrže preklapanje sa ispitnim skupom.

Prosječna preciznost	0.333
Prosječan odziv	1.0

Tablica 5.2: Prosječna preciznost i odziv glagolskih razreda

Iako je ispitni skup podataka bio malen, odmah je vidljivo da je odziv postignuo savršen rezultat što znači da su svi potrebni semantički razredi završili u istoj grupi. S druge strane preciznost je vrlo loša iz čega možemo zaključiti da je model zaključio da su različiti semantički okviri također vrlo slični iako u stvarnosti nisu. Potencijalni pristupi rješavanju tog problema bilo bi penaliziranje različitosti prilikom izračuna sličnosti te optimizacija parametara grupiranja.

Iz korpusa su ekstrahirani početni okviri 5145 glagola. Filtriranjem početnih okvira te grupiranjem u semantičke okvire te odbacivanjem okvira koji se sastoje od premalo okvira dobiveni su semantički okviri za 2229 glagola. Grupiranjem glagola u razrede dobiveno je 2010 glagolskih razreda. Primjerak dobivenog glagolskog razreda nalazi se u dodatku B.

6. Zaključak

U ovom radu izgrađen je model za ekstrakciju glagolskih razreda za hrvatski jezik. Njegova implementacija izvedena je u programskom jeziku Python. Problem ekstrakcije glagolskih razreda težak je problem analize prirodnih jezika te je u ovom modelu pristup rješavanja bio dvostupanjsko grupiranje. Tretiranjem konteksta rečenica kao podatkovne jedinice u prvom koraku te njihovom grupiranju u semantičke okvire dobiveni su rezultati kojima se savladao problem polisemije glagola.

Rezultati ovog modela nisu bili zadovoljavajući. Iako su grupiranjem postignuti dobri rezultati po grupiranju riječi uzimajući u obzir njihovu službu, neke karakteristike sličnosti ipak nisu bile zadovoljavajuće. Iako je uzeto u obzir mogućnost pojavljivanja dvije riječi u dva konteksta, sličnost je prepoznata samo ako je korištena ista riječ.

Za postizanje boljeg rješenja potrebno je izmijeniti neke komponente sustava. Element koji bi najviše utjecao na rezultate bila bi sličnost okvira. Potrebno je modelirati sličnost okvira koja uzima u obzir ne samo podudaranje riječi već i njihovu sličnost. Dodatno poboljšanje algoritma bila bi optimizacija težina varijabli prilikom izračuna sličnosti. Sljedeće moguće poboljšanje bilo bi dodavanje drugih atributa okvirima što bi proširilo upamćeni kontekst u kojemu se glagol pojavljuje. U budućoj implementaciji modela značajno poboljšanje performansi bilo bi dobiveno u drugom koraku algoritma uvođenjem višedretvenosti. Također, prilikom razvoja ovog modela razvijena je i druga funkcija sličnosti okvira temeljena na vektorima sličnosti riječi. Početni rezultati pomoću te funkcije nisu bili zadovoljavajući pa ovom prilikom nije uključena u rad. Ipak, funkcija je obećavajuća te će se izvršiti dodatna optimizacija parametara modela u budućnosti.

S obzirom da je ovo bio rad ovakve prirode za hrvatski jezik, implementacijom poboljšanja iz prijašnjeg odlomka rezultati modela trebali bi biti značajno poboljšani. Za opširnije testiranje modela te optimizaciju težina u funkciji sličnosti i parametara grupiranja potreban je dodatan rad na problemu.

LITERATURA

[1] Clustering image, 2013. URL `http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/`.

[2] Gobinda G Chowdhury. Natural language processing. *Annual review of information science and technology*, 37(1):51–89, 2003.

[3] Susan Feldman. Nlp meets the jabberwocky. *Online*, 23(3):62–72, 1999.

[4] John A Hartigan. Clustering algorithms. 1975.

[5] Eric Joanis, Suzanne Stevenson, i David James. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(03):337–367, 2008.

[6] Daisuke Kawahara, Daniel W Peterson, i Martha Palmer. A step-wise usage-based method for inducing polysemy-aware verb classes. U *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL2014)*, stranice 1030–1040.

[7] Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.

[8] Elizabeth D Liddy. Enhanced text retrieval using natural language processing. *Bulletin of the American Society for Information Science and Technology*, 24(4): 14–16, 1998.

[9] Nikola Ljubešić i Filip Klubička. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. U *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, stranice 29–35, Gothenburg, Sweden, 2014. Association for Computational Linguistics.

[10] Karin Kipper Schuler. Verbnet: A broad-coverage, comprehensive verb lexicon. 2005.

[11] Lin Sun i Anna Korhonen. Improving verb clustering with automatically acquired selectional preferences. U *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, stranice 638–647. Association for Computational Linguistics, 2009.

[12] Stijn Marinus Van Dongen. Graph clustering by flow simulation. 2001.

[13] Walber. Precision and recall, 2014. URL `http://en.wikipedia.org/wiki/Precision_and_recall`.

Primjena nenadziranog strojnog učenja za akviziciju glagolskih razreda iz korpusa

Sažetak

Problem sličnosti riječi težak je problem područja analize prirodnog jezika. U ovom radu opisano je stvaranje glagolskih razreda pomoću metoda nenadziranog strojnog učenja. Korišten je algoritam grupiranja MCL te je rad modela isproban na podacima korpusa *hrWaC*. Implementacija modela je ostvarena u programskom jeziku Python.

Ključne riječi: nenadzirano strojno učenje, analiza prirodnog jezika, glagolski razredi, grupiranje, sličnost riječi, semantika

Acquisition of Verb Classes from Corpus using Unsupervised Machine Learning

Abstract

Word similarity is a difficult natural language processing task. In this paper, verb class acquisition problem is solved using unsupervised machine learning methods. For the clustering parts, MCL algorithm has been used and the model has been tested on data from *hrWaC* corpus. The model described in this paper has been implemented in Python programming language.

Keywords: unsupervised machine learning, natural language processing, verb classes, clustering, word similarity, semantics

Dodatak A
Ispitni okviri glagola "zabiti"

zabiti {’gol’, ’internacionalo’} {’u’, ’od’} {’hat’}
zabiti {’gol’} {’iz’, ’nakon’} {’ronaldo’}
zabiti {’gol’} set() {’ivanschitz’}
zabiti {’gol’} {’s’} set()
zabiti {’autogol’} {’u’} set()
zabiti {’pogodak’} {’na’} set()
zabiti {’koš’} {’uz’} set()
zabiti {’koš’} set() {’aldridge’}
zabiti {’vodstvo’} {’za’} set()
zabiti {’vodstvo’} {’u’, ’za’} {’kosticin’}

zabiti {’glava’} {’u’} set()
zabiti {’prikolica’} {’u’} set()
zabiti {’ograda’} {’u’} {’piquet’, ’sumnja’}
zabiti {’vožnja’} {’u’, ’za’, ’usred’} set()
zabiti {’kuća’} {’u’} set()
zabiti {’on’} {’u’} set()
zabiti {’traktor’} {’u’} set()
zabiti {’automobil’} {’u’, ’ispred’} set()
zabiti {’automobil’} {’u’} set()
zabiti {’spomenik’} {’u’} set()

Dodatak B
Primjerak glagolskog razreda

’dodati’, ’slagati’, ’pripremati’, ’uputiti’, ’ugroziti’, ’zabrinjavati’, ’upisivati’, ’oboljeti’, ’prirediti’, ’zapošljavati’, ’odvojiti’, ’hotati’, ’zadovoljavati’, ’skidati’, ’počiniti’, ’osigurati’, ’ukinuti’, ’narasti’, ’reći’, ’oglasiti’, ’kupiti’, ’koti’, ’primjenjivati’, ’dobivati’, ’paziti’, ’kazati’, ’naučiti’, ’skupiti’, ’privlačiti’, ’uzeti’, ’emitirati’, ’maknuti’, ’razviti’, ’snimati’, ’objavljivati’, ’odigrati’, ’označavati’, ’izbaciti’, ’prethoditi’, ’sugerirati’, ’izvesti’, ’odbiti’, ’postizati’, ’bijeti’, ’podići’, ’lažiti’, ’postignuti’, ’značiti’, ’predvoditi’, ’povlačiti’, ’upustiti’, ’navoditi’, ’tumačiti’, ’napredovati’, ’buditi’, ’poticati’, ’nastati’, ’nastupiti’, ’njati’, ’učvršćivati’, ’spasiti’, ’krenuti’, ’ustajati’, ’činiti’, ’zamoliti’, ’približiti’, ’rješavati’, ’vladati’, ’zagovarati’, ’predlagati’, ’dokazivati’, ’oduševiti’, ’doznati’, ’pobijediti’, ’pokupiti’, ’upati’, ’ukazivati’, ’seliti’, ’snalaziti’, ’koći’, ’pušiti’, ’pomagati’, ’uzvratiti’, ’uvesti’, ’kožiti’, ’očistiti’, ’zabavljati’, ’povezati’, ’proširiti’, ’započinjati’, ’osjeti’, ’isključiti’, ’budeti’, ’zauzimati’, ’dopuštati’, ’boliti’, ’doznavati’, ’izvaditi’, ’ogledati’, ’odrediti’, ’optuživati’, ’podnijeti’, ’spriječiti’, ’pitati’, ’odaći’, ’razmisliti’, ’stati’, ’dijeliti’, ’iznenaditi’, ’objašnjavati’, ’podsjetiti’, ’otići’, ’opravdati’, ’predložiti’, ’probiti’, ’prozvati’, ’spašavati’, ’predstavljati’, ’susreti’, ’svađati’, ’priznavati’, ’spuštati’, ’ubaciti’, ’čuti’, ’pokriti’, ’tužiti’, ’povući’, ’pripadati’, ’podržavati’, ’upravljati’, ’zahtijevati’, ’pokrivati’, ’boti’, ’bježati’, ’ovisiti’, ’poznati’, ’ostati’, ’pokušati’, ’zaključiti’, ’izbjegavati’, ’dovesti’, ’nasmijati’, ’doprinositi’, ’istraživati’, ’jeti’, ’gađati’, ’odati’, ’pjeti’, ’pozdraviti’, ’uhvatiti’, ’ulaziti’, ’voljeti’, ’približavati’, ’obilaziti’, ’istaknuti’, ’pogađati’, ’posjetiti’, ’opravdavati’, ’okušati’, ’proizvesti’, ’zadržavati’, ’isključivati’, ’predati’, ’omogućivati’, ’popeti’, ’izgubiti’, ’upozoriti’, ’mučiti’, ’poboljšati’, ’viditi’, ’silaziti’, ’posuditi’, ’ljubiti’, ’pohvati’, ’shvaćati’, ’putovati’, ’veseliti’, ’otkrivati’, ’natjecati’, ’izostati’, ’usmjeriti’, ’skinuti’, ’konzultirati’, ’okupljati’, ’pomisliti’, ’zaslužiti’, ’tjerati’, ’priopćiti’, ’nastaviti’, ’upućivati’, ’preseliti’, ’vrijediti’, ’naviknuti’, ’popraviti’, ’baciti’, ’raspisati’, ’nastradati’, ’odnositi’, ’sastati’, ’navesti’, ’plesati’, ’predvidjeti’, ’snimiti’, ’zamijeniti’, ’prijeći’, ’odgovarati’, ’suočavati’, ’planirati’, ’glasati’, ’sanjati’, ’poručivati’, ’ubijati’, ’propustiti’, ’odbijati’, ’ignorirati’, ’uspjeti’, ’pomognuti’, ’prenijeti’, ’primijetiti’, ’uvoditi’, ’diviti’, ’otkazati’, ’ostvariti’, ’ojačati’, ’vući’, ’poštovati’, ’izazivati’, ’protiviti’, ’garantirati’, ’zaustaviti’, ’izvijestiti’, ’unositi’, ’stajati’, ’prolaziti’, ’dozvoliti’, ’savjetovati’, ’desiti’, ’osvajati’, ’probuditi’, ’pokušavati’, ’učvrstiti’, ’nagađati’, ’proizlaziti’, ’raskinuti’, ’potvrditi’, ’nadati’, ’prihvaćati’, ’primijeniti’, ’trgovati’, ’pomoći’, ’miješati’, ’kupovati’, ’požaliti’, ’sklopiti’, ’pretvarati’, ’živiti’, ’poslužiti’, ’steći’, ’donesti’, ’izgraditi’, ’posjedovati’, ’dignuti’, ’preminuti’, ’sadržati’, ’slazati’, ’puniti’, ’stjecati’, ’prikupiti’, ’posvati’, ’diplomirati’, ’ljeći’, ’osjećati’, ’pokrenuti’, ’kužiti’, ’operirati’, ’izvršavati’, ’braniti’, ’moliti’, ’nedostajati’, ’prevoziti’, ’pojavljivati’, ’isteći’, ’stupiti’, ’proizvoditi’, ’svladati’, ’uhititi’, ’organizirati’, ’preporučati’, ’posviti’,

'preostajati', 'studirati', 'uključivati', 'slijediti', 'zapaliti', 'zamisliti', 'zalagati', 'pratiti', 'držati', 'prikupljati', 'zaposliti', 'reagirati', 'pući', 'rušiti', 'naručiti', 'očitivati', 'pronaći', 'slaviti', 'soti', 'pogriješiti', 'izdavati', 'nazivati', 'odreći', 'izbiti', 'donirati', 'pjevati', 'potresati', 'prikazivati', 'zaraziti', 'pobiti', 'sjediti', 'ljubaviti', 'javljati', 'raniti', 'kupivati', 'manjkati', 'gostovati', 'podsjećati', 'kati', 'ublažiti', 'donijeti', 'aktivirati', 'vjerovati', 'zaboraviti', 'uplatiti', 'naplatiti', 'osloboditi', 'praviti', 'slomiti', 'zatvarati', 'istražiti', 'svjedočiti', 'surađivati', 'završavati', 'provoditi', 'ugasiti', 'nadoknaditi', 'nemoći', 'obavljati', 'iskazivati', 'dostaviti', 'izaći', 'omogućiti', 'kandidirati', 'kršiti', 'odnijeti', 'uroditi', 'pogoršavati', 'raspravljati', 'uvati', 'promatrati', 'uočiti', 'osvrnuti', 'nazvati', 'spustiti', 'imenovati', 'izvršiti', 'pokvariti', 'ponuditi', 'oslobađati', 'konstatirati', 'preuzimati', 'teći', 'pobjediti', 'odvezati', 'događati', 'plaćati', 'osjetiti', 'tribati', 'prihvatiti', 'kriti', 'plasirati', 'ispunjavati', 'razgovarati', 'angažirati', 'oporaviti', 'promovirati', 'pregovarati', 'kamoliti', 'napisati', 'vezati', 'zarađivati', 'dužiti', 'uspijevati', 'potpisati', 'pristajati', 'izjasniti', 'oduševljavati', 'pretvoriti', 'nemati', 'netati', 'ispuniti', 'trčati', 'odobravati', 'smiriti', 'trpiti', 'smanjivati', 'udariti', 'siti', 'roditi', 'bilježiti', 'ugrađivati', 'nositi', 'odgovoriti', 'obilježiti', 'podignuti', 'obnoviti', 'ohladiti', 'uskladiti', 'nazočiti', 'pronalaziti', 'sastaviti', 'zabilježiti', 'zahvaljovati', 'čuvati', 'odraditi', 'svirati', 'poboljšavati', 'parkirati', 'povezivati', 'slati', 'čekati', 'navijati', 'platiti', 'sastojati', 'komunicirati', 'napominjati', 'pozivati', 'gurnuti', 'propisivati', 'počinjati', 'ukazati', 'ugostiti', 'financirati', 'osuđivati', 'težiti', 'obožavati', 'vratiti', 'ispaliti', 'nemati', 'pomaknuti', 'sjedati', 'dodijeliti', 'viđati', 'dovršiti', 'zaraditi', 'boriti', 'hvati', 'prestajati', 'civiti', 'kositi', 'žaliti', 'posvećivati', 'proteći', 'startati', 'izboriti', 'prepoznati', 'boći', 'stići', 'zabaviti', 'pojasniti', 'ponijeti', 'zaigrati', 'vjenčati', 'utjecati', 'maltretirati', 'sumnjati', 'ponašati', 'prepustiti', 'provaliti', 'provjeriti', 'pustiti', 'učiniti', 'ocjenjivati', 'napadati', 'oprati', 'zadovoljiti', 'poduzimati', 'ispostaviti', 'trošiti', 'primjećivati', 'jačati', 'zastupati', 'okrenuti', 'odlučiti', 'preporučiti', 'uplaćivati', 'umirati', 'koštati', 'upoznati', 'doseći', 'izgorjeti', 'koristiti', 'porasti', 'konzumirati', 'probati', 'istrčati', 'riješiti', 'uvjeriti', 'liječiti', 'upotrebljavati', 'iskazati', 'dvojiti', 'izvući', 'paliti', 'progovoriti', 'položiti', 'pricati', 'gorivati', 'odbacivati', 'dostavljati', 'polagati', 'podnositi', 'krijati', 'najtežiti', 'padati', 'zahvaliti', 'dirati', 'kretati', 'popustiti', 'realizirati', 'razumjeti', 'pridružiti', 'zateći', 'predavati', 'zabranjivati', 'urediti', 'izreći', 'razmatrati', 'prelaziti', 'poželjeti', 'uginuti', 'objasniti', 'stavljati', 'letiti', 'razmišljati', 'sretati', 'pružiti', 'doživjeti', 'odlučivati', 'nemojtiti', 'uzimati', 'odraziti', 'spominjati', 'oduzeti', 'apsorbirati', 'iskoristiti', 'odustajati', 'informirati', 'optužiti', 'dogovarati', 'nuditi', 'prezentirati', 'smijati', 'izvoditi', 'stizati', 'izjednačiti', 'obuhvaćati', 'dobrodoći', 'uloviti', 'zaprijetiti', 'daći', 'izdvajati', 'pristupiti', 'pozvati', 'nabavljati', 'hodati', 'zasluživati', 'zaključivati', 'mariti', 'opustiti', 'doživiti', 'postavljati', 'osnovati', 'trenirati', 'odlaziti', 'govoriti', 'baviti', 'zamišljati', 'luditi', 'otputovati', 'prestati', 'otpadati', 'poslati', 'pogledati', 'složiti', 'zaštititi', 'moljeti', 'patiti', 'ati', 'kombinirati', 'zauzeti', 'postupati', 'potrajati', 'upitati', 'radovati', 'odobriti', 'dovoditi', 'brinuti', 'prouzročiti', 'zaostajati', 'naći', 'uništiti', 'ispadati', 'rađati', 'pokazivati', 'pomati', 'zadužiti', 'iznositi', 'štititi', 'potražiti', 'veseti', 'dolaziti', 'testirati', 'otkriti', 'sadržavati', 'proizaći', 'ležati', 'prenositi', 'glasovati', 'najavljivati', 'očitovati', 'računati', 'visiti', 'izdvojiti', 'procjenjivati', 'naglasiti', 'pohađati', 'obnašati', 'izazvati', 'obraćati', 'pogoditi', 'ponavljati', 'skočiti', 'demantirati', 'poznatti', 'skrenuti', 'pojeti', 'odvesti', 'obnavljati', 'usporediti', 'davati', 'potaknuti', 'udarati', 'predstaviti', 'bilježati', 'dosegnuti', 'birati', 'isplati', 'sakupiti', 'osiguravati', 'skuhati', 'dizati', 'pitatiti', 'ustvrditi', 'oteti', 'uključiti', 'olakšati', 'obuhvatiti', 'uvoziti', 'zatvoriti', 'ukloniti', 'hvaliti', 'strahovati', 'potruditi', 'fotografirati', 'vježbati', 'suočiti', 'kritizirati', 'bolovati', 'gosti', 'napustiti', 'osuditi', 'potvrđivati', 'izraziti', 'nestati', 'pasti', 'defini-

rati’, ’pati’, ’posjeti’, ’prodati’, ’uzrokovati’, ’zadržati’, ’širiti’, ’najaviti’, ’prijetiti’, ’opisati’, ’sudariti’, ’održavati’, ’sjesti’, ’brati’, ’pretrpjeti’, ’dozvoljavati’, ’bojati’, ’određivati’, ’preporučivati’, ’primati’, ’najbržati’, ’dodjeljivati’, ’čestitati’, ’narediti’, ’dugovati’, ’smatrati’, ’prikazati’, ’mjeriti’, ’postupiti’, ’produžiti’, ’naslijediti’, ’zapitati’, ’oštetiti’, ’pozdravljati’, ’slabiti’, ’privući’, ’plaviti’, ’srediti’, ’zaljubiti’, ’preskočiti’, ’blokirati’, ’odbaciti’, ’nestajati’, ’pobjeđivati’, ’jedvati’, ’odustati’, ’pripasti’, ’družiti’, ’nametati’, ’procijeniti’, ’vraćati’, ’prekršiti’, ’kasniti’, ’gasiti’, ’sjeti’, ’povećavati’, ’obećati’, ’silovati’, ’bacati’, ’otvoriti’, ’prebaciti’, ’nastajati’, ’smetati’, ’sukobiti’, ’podrazumijevati’, ’trajati’, ’zaboravljati’, ’izmjenjivati’, ’ulagati’, ’ispasti’, ’pisati’, ’promijeniti’, ’stradati’, ’stupati’, ’započeti’, ’piti’, ’komentirati’, ’smanjiti’, ’ostajati’, ’zovati’, ’iznijeti’, ’vršiti’, ’upisati’, ’okupiti’, ’priključiti’, ’prometovati’, ’zaustavljati’, ’poginuti’, ’držiti’, ’shvatiti’, ’obavijestiti’, ’obratiti’, ’dići’, ’prakticirati’, ’obrađivati’, ’gurati’, ’kontaktirati’, ’ostavljati’, ’proslaviti’, ’naplaćivati’, ’otpjevati’, ’izvlačiti’, ’podizati’, ’izlagati’, ’obraniti’, ’trati’, ’doživljavati’, ’poduzeti’, ’učiti’, ’doputovati’, ’ukrati’, ’promašiti’, ’ticati’, ’djelovati’, ’saznati’, ’suditi’, ’htjeti’, ’morati’, ’stignuti’, ’uložiti’, ’voliti’, ’umrijeti’, ’usvojiti’, ’usuditi’, ’prati’, ’naglašavati’, ’zatražiti’, ’primiti’, ’utvrđivati’, ’naići’, ’posvetiti’, ’posjećivati’, ’tragati’, ’pridonositi’, ’pobjeći’, ’javiti’, ’poručiti’, ’omogućavati’, ’uvažavati’, ’preuzeti’, ’bivati’, ’ozlijediti’, ’zvučiti’, ’ispraviti’, ’proglasiti’, ’brojati’, ’dodavati’, ’odazvati’, ’razbiti’, ’šetati’, ’dospjeti’, ’dočekati’, ’skupljati’, ’upasti’, ’utvrditi’, ’dokazati’, ’formirati’, ’provesti’, ’rasti’, ’potjecati’, ’idati’, ’bistiti’, ’jamčiti’, ’slušati’, ’služiti’, ’popiti’, ’odabrati’, ’odrasti’, ’rezultirati’, ’ispati’, ’izmijeniti’, ’uslijediti’, ’pričati’, ’preostati’, ’isticati’, ’poštivati’, ’izdržati’, ’skužiti’, ’povesti’, ’osvojiti’, ’uspostaviti’, ’ocijeniti’, ’podupirati’, ’buniti’, ’proučiti’, ’opisivati’, ’ući’, ’čivati’, ’provjeravati’, ’očuvati’, ’trebatiti’, ’glasiti’, ’staviti’, ’prodavati’, ’neznati’, ’podržati’, ’varati’, ’izdati’, ’pružati’, ’regulirati’, ’tretirati’, ’jesti’, ’označiti’, ’izražavati’, ’podijeliti’, ’napuniti’, ’oslabiti’, ’spadati’, ’raspolagati’, ’susretati’, ’napraviti’, ’srušiti’, ’predviđati’, ’spomenuti’, ’isplatiti’, ’istjecati’, ’razmotriti’, ’oprostiti’, ’kontrolirati’, ’obojati’, ’opovrgnuti’, ’ubiti’, ’pristići’, ’pojaviti’, ’registrirati’, ’izlaziti’, ’ploviti’, ’ideti’, ’postajati’, ’intervenirati’, ’puštati’, ’vrijeđati’, ’zaprimiti’, ’pamtiti’, ’uraditi’, ’napasti’, ’isplaćivati’, ’uvjeravati’, ’prevezti’, ’odgađati’, ’potrošiti’, ’pristati’, ’propasti’, ’izbrisati’, ’preći’, ’pucati’, ’smjestiti’, ’potpisivati’, ’graditi’, ’sreti’, ’nabaviti’, ’razočarati’, ’nastojati’, ’smiti’, ’izraditi’, ’otpustiti’, ’sviđati’, ’tući’, ’stvarati’, ’viti’, ’pregledati’, ’obznaniti’, ’obići’, ’dinati’, ’kaniti’, ’donositi’, ’zvati’, ’pohvaliti’, ’ispričati’, ’izgledati’, ’katedrati’, ’sjetiti’, ’pretpostavljati’, ’premašiti’, ’odmoriti’, ’ispitati’, ’poduprijeti’, ’tražiti’, ’zadobiti’, ’izići’, ’pričekati’, ’povećati’, ’smjeti’, ’voziti’, ’primjetiti’, ’boraviti’, ’znatiti’, ’pridonijeti’, ’sačuvati’, ’zabraniti’, ’ići’, ’prekinuti’, ’dopustiti’, ’gledati’, ’nametnuti’, ’ostvarovati’, ’tvrditi’, ’dogovoriti’, ’obaviti’, ’privesti’, ’ustanoviti’, ’preliti’, ’razbijati’, ’smirivati’, ’pojačati’, ’izabrati’, ’truditi’, ’pročitati’, ’uređivati’, ’valjati’, ’zahvaljivati’, ’izrađivati’, ’čitati’, ’namjeravati’, ’funkcionirati’, ’lagati’, ’poslovati’, ’pridržavati’, ’gruditi’, ’peći’, ’spremati’, ’mislititi’, ’razvijati’, ’stvoriti’, ’pobrinuti’, ’završiti’, ’prisjetiti’, ’postaviti’, ’napomenuti’, ’nisiti’, ’pokretati’, ’glumiti’, ’poznavati’, ’odmarati’, ’ispitivati’, ’nastavljati’, ’spavati’, ’okončati’, ’analizirati’, ’mijenjati’, ’čuditi’, ’nastupati’, ’skrivati’, ’dominirati’, ’otvarati’, ’odgoditi’, ’odražavati’, ’pokazati’, ’prijaviti’, ’snaći’, ’osmisliti’, ’gubiti’, ’nemojti’, ’uživati’, ’preživjeti’, ’ostaviti’, ’zanimati’, ’ponoviti’, ’odvijati’, ’obećavati’, ’dugati’, ’priznati’, ’poći’, ’okretati’, ’pripremiti’, ’proučavati’, ’prići’, ’proći’, ’izbjeći’, ’kuhati’, ’obilježavati’, ’sjećati’, ’ostvarivati’, ’prisustvovati’, ’upozoravati’, ’razlikovati’, ’nadzirati’, ’zabijati’, ’inzistirati’