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

DIPLOMSKI RAD br. 1324

Kompozicijska distribucijska semantika temeljena na modelu leksičke funkcije

Zoran Medić

Zagreb, lipanj 2016.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA DIPLOMSKI RAD PROFILA

Zagreb, 11. ožujka 2016.

Predmet: **Strojno učenje**

DIPLOMSKI ZADATAK br. 1324

Pristupnik: **Zoran Medić (0036465179)**
Studij: Računarstvo
Profil: Računarska znanost

Zadatak: **Kompozicijska distribucijska semantika temeljena na modelu leksičke funkcije**

Opis zadatka:

Računalna semantika ima važnu ulogu u sustavima za obradu i razumijevanje prirodnoga jezika. Distribucijski semantički modeli značenje riječi prikazuju kontekstnim vektorima u višedimenzijskom vektorskom prostoru. Kompozicijska distribucijska semantika bavi se izgradnjom prikaza značenja višerječnih fraza u vektorskome prostoru.
U radu je potrebno proučiti i opisati postojeće distribucijske semantičke modele i modele kompozicijske distribucijske semantike te postupke njihove izgradnje i vrednovanja. Proučiti modele temeljene na tenzorskoj algebri, model leksičke funkcije Baronija i Zamparellija (2010) te praktični model leksičke funkcije (PLF) Papernoa i dr (2014). Proučiti proširenja modela predložena u radu Gupta i dr. (2015). Razviti implementaciju modela PLF za hrvatski jezik. Izgraditi odgovarajući ispitni skup podataka za provjeru modela. Razmotriti nadogradnju modela temeljenu na relaciji semantičke inkluzije između fraza. Provesti konačno vrednovanje modela na zadatku semantičke kompozicije duljih fraza te razmotriti i druge mogućnosti vrednovanja modela. Radu priložiti izvorni i izvršni kod razvijenog sustava, skupove podataka i potrebnu dokumentaciju te citirati korištenu literaturu.

Zadatak uručen pristupniku: 18. ožujka 2016.
Rok za predaju rada: 1. srpnja 2016.

Mentor:

Doc. dr. sc. Jan Šnajder

Djelovođa:

Doc. dr. sc. Tomislav Hrkać

Predsjednik odbora za diplomski rad profila:

Prof. dr. sc. Siniša Srbljić

Zahvaljujem mentoru dr. sc. Janu Šnajderu, doc., na pomoći i potpori tijekom izrade rada, kao i tijekom cijelog studija.

Zahvaljujem i svim studentima TakeLab-a koji su pomogli pri označavanju podataka potrebnih za izradu rada.

Posebno zahvaljujem svojoj obitelji, majci Marijani, ocu Željku i sestri Josipi na bezuvjetnoj ljubavi, podršci i razumijevanju koje mi pružaju kroz cijeli život.

SADRŽAJ

1. Uvod

Računalna semantika područje je u obradi prirodnoga jezika koje proučava različite načine prikaza jezičnih jedinica te korištenje takvih prikaza u drugim zadacima obrade prirodnoga jezika. Neki od problema kojima se bavi računalna semantika su izgradnja prikaza značenja riječi, razrješavanje anafore, razrješavanje višeznačnosti i automatsko zaključivanje.

Distribucijski semantički modeli značenje riječi prikazuju kontekstnim vektorima u višedimenzijskom vektorskom prostoru. Temelje se na statističkim značajkama riječi u velikim jezičnim korpusima. Brojna istraživanja u protekla dva desetljeća pokazala su da se takav način prikaza značenja riječi može uspješno koristiti u raznim zadacima obrade prirodnoga jezika. Uspješnost distribucijskih modela na razini riječi potaknula je razvoj složenijih modela, koji koristeći vektorske prikaze riječi modeliraju značenje fraza koje ih sadrže.

Kompozicijska distribucijska semantika bavi se izgradnjom prikaza značenja višerječnih fraza u vektorskome prostoru. U tom postupku naglasak je upravo na kompoziciji riječi u frazi, budući da različite jezične konstrukcije, premda sastavljene od istih riječi, mogu imati i različito značenje. Kompozicijsko-distribucijski modeli razlikuju se prema načinu pristupa kompoziciji vektora riječi u višerječnim frazama, odnosno operacijama vektorske algebre između vektora riječi u frazi. Jednostavne vektorske operacije poput zbrajanja ili množenja vektora, iako uspješne s obzirom na svoju jednostavnost, nisu se pokazale dovoljno efikasnima kada su u pitanju višerječne fraze. Zbog toga se sve više istražuju različiti pristupi kompoziciji vektora riječi u frazi, kako bi se pronašao praktičan i uspješan model za prikaz značenja višerječnih fraza u vektorskome prostoru.

U okviru diplomskog rada istraženi su trenutno poznati distribucijsko semantički modeli, kao i modeli kompozicijske distribucijske semantike, prvenstveno modeli temeljeni na tenzorskoj algebri, model leksičke funkcije (Baroni i Zamparelli, 2010) te model praktične leksičke funkcije – PLF (Paperno et al., 2014). Posebno je detaljno proučen PLF-model, koji se uz svoju praktičnost istaknuo i vrlo dobrim rezultatima.

Za taj model razmotrena su i neka od mogućih proširenja temeljena na relaciji semantičke inkluzije između parova fraza sastavljenih od sinonima, odnosno hiponima i hiperonima. Implementiran je i PLF-model za hrvatski jezik te vrednovan na skupu višerječnih fraza u zadatku određivanja semantičke sličnosti višerječnih fraza. Razmotreni su i drugi oblici primjene modela u okviru drugih zadataka računalne semantike, uključujući zadatke semantičke kompozitnosti i devijantnosti.

U nastavku rada dan je pregled u literaturi opisanih distribucijsko semantičkih modela, nakon čega slijedi poglavlje s posebnim pregledom nekoliko značajnijih modela kompozicijske distribucijske semantike. U četvrtom poglavlju detaljnije je opisan praktični model leksičke funkcije, uključujući izgradnju modela i provedeno vrednovanje istog. Peto poglavlje sadrži opis nadogradnji modela predstavljenih u radu (Gupta et al., 2015) te razmotrene nadogradnje temeljene na relaciji semantičke inkluzije među vektorima određenih imenica. U šestom poglavlju opisana je izgradnja modela za hrvatski jezik, zajedno s analizom rezultata vrednovanja modela na višerječnim frazama na hrvatskome jeziku. Sedmo poglavlje sadrži opise drugih načina mogućeg vrednovanja modela, s obzirom na vrstu problema u kojoj se model primjenjuje, uključivo probleme semantičke kompozitnosti i semantičke devijantnosti. Osmo poglavlje sadrži zaključak, zajedno s prijedlozima za daljnji rad. Na samom kraju rada nalaze se dodatci u kojima su dani skupovi izraza na hrvatskom jeziku, korišteni pri vrednovanju modela u radu.

2. Distribucijski semantički modeli

Distribucijski semantički modeli temelje se na kontekstnim vektorima u višedimenzijskom prostoru, pomoću kojih je prikazano semantičko značenje pojedine riječi. Budući da vektori opisuju kontekst u kojem se riječ koristi, slični vektori pojedinih riječi podrazumijevaju i njihovo slično značenje.

Sama ideja prikaza riječi pomoću kontekstnih vektora pojavila se nakon uspješne primjene sličnog modela na prepoznavanje sličnosti dokumenata. Naime, Salton et al. (1975) u svom su radu opisali razvoj modela za određivanje sličnosti dokumenata na temelju riječi koje su u njima pojavljuju. Dokumente su prikazali u matrici u kojoj je svaki stupac jedan dokument, a retci su riječi koje se u dokumentima pojavljuju. Za svaki dokument prebrojana su pojavljivanja pojedine riječi u njemu, te su u retcima matrice naznačeni brojevi pojavljivanja određene riječi u određenom dokumentu.

Iako je takav način prikaza sadržaja dokumenta prilično jednostavan, budući da se niti poredak riječi, kao ni jezične konstrukcije poput fraza i rečenica ne uzimaju u obzir, matrični zapis pojmova u dokumentima (engl. *term-document matrix*) ipak se pokazao znatno uspješnim u pronalasku međusobno sličnih dokumenata na temelju njihovih vektora (stupaca u matrici). Intuitivno obrazloženje tog uspjeha leži u tome da tema dokumenta bitno utječe na raspon vokabulara koji autor koristi kada o toj temi piše, pa će samim time i dokumenti o sličnim temama imati i sličan skup riječi korišten u njima. Osim intuitivnog, ponudili su i statističko obrazloženje, takozvanu statističku semantičku hipotezu (engl. *statistical semantics hypothesis*) koja kaže da statistički uzorci korištenja pojedinih riječi mogu određivati o čemu ljudi govore. Ako hipotezu primijenimo na dokumente, značilo bi to da dokumenti sa sličnim vektorima imaju slično značenje.

Deerwester et al. (1990) svoje su istraživanje umjesto na sličnost dokumenata usmjerili na sličnost riječi, odnosno sličnost redaka u prethodno opisanoj matrici. Inspirirani matricom riječi u dokumentima, zaključili su da se u takvoj matrici osim dokumenata mogu naći i druge jezične jedinice koje sadrže kontekst u kojem se riječ pojavljuje: od paragrafa, rečenica i fraza do samih riječi ili običnih slijedova znakova.

	Ana	*jesti*	*ukusan*	*čokoladan*	*kolač*
Ana		1	1		
jesti	1		1	1	
ukusan	1	1		1	1
čokoladan		1	1		1
kolač				1	1

Tablica 2.1: Matrica supojavljivanja s prozorom veličine dva.

Kontekst u kojem se riječ pojavljuje može biti definiran također na više načina: pomoću prozora riječi, sintaktičke ovisnosti riječi u tekstu, ali i kompliciranijih načina poput kombiniranja gramatičke ovisnosti i selektivnog odabira pozicija riječi. Ideja i dalje ostaje ista – slični vektori imaju i slično značenje, ali ovaj put riječ je o vektorima riječi, a ne dokumenata.

Statističko obrazloženje ovog pristupa nazvano je u lingvistici distribucijskom hipotezom (engl. *distributional hypothesis*), koja kaže da riječi korištene u sličnim kontekstima imaju i slično značenje. Ta hipoteza pokazala se poprilično točnom, te su na njoj utemeljeni brojni distribucijski semantički modeli.

Primjer jednostavnog distribucijskog semantičkog modela je model temeljen na prozoru riječi kojim se obuhvaća kontekst pojedine riječi u tekstu (Lund i Burgess, 1996). Model se temelji na "prozoru" koji prolaskom kroz tekst prikuplja riječi koje se pojavljuju u okolini trenutačno promatrane riječi. Tim skupom okolnih riječi definira se kontekst u kojem se svaka riječ pojavljuje. Veličina prozora može varirati, ovisno o specifičnosti konteksta koji se želi definirati. Također, moguće je okolnim riječima pridati određene težine, ovisno o tome koliko su one udaljene od trenutačno promatrane riječi. Konačni rezultat prolaska tako definiranog prozora kroz tekst je matrica supojavljivanja u kojoj su retci i stupci sastavljeni od istog skupa riječi, dok vrijednosti u matrici predstavljaju broj pojavljivanja određene riječi unutar prozora njoj odgovarajuće. Primjer takve matrice za frazu "*Ana jede ukusan čokoladan kolač*" nalazi se u tablici 2.1.

Grefenstette (1994) je u svom radu opisao distribucijski model temeljen na sintaktičkoj ovisnosti riječi u tekstu. Opisani model složeniji je od modela temeljenog na prozoru riječi, budući da umjesto običnog supojavljivanja riječi razmatra sintaktičku ovisnost između riječi neovisno o njihovoj poziciji u rečenici. Ovisnosti koje se razmatraju su subjekt–glagol, član–imenica, pridjev–imenica i sl. Konačna matrica na

	(subj, *Ana*)	(pred, *jesti*)	(atr, *ukusan*)	(atr, *čokoladan*)	(obj, *kolač*)
Ana		x			
jesti	x				x
ukusan					x
čokoladan					x
kolač		x	x	x	

Tablica 2.2: Matrica sintaktičkih ovisnosti temeljena samo na izravnim sintaktičkim vezama između riječi u rečenici.

kraju je sastavljena od redaka koji predstavljaju riječi u tekstu te stupaca koji predstavljaju sintaktičku ovisnost pojedinih riječi iz teksta. Moguće je više načina popunjavanja elemenata u matrici, od pukog označavanja pojavljivanja sintaktičke ovisnosti za neku riječ, preko broja takvih pojavljivanja u svim promatranim jedinicama teksta, do množenja tog broja s određenim koeficijentom, ovisno o jakosti sintaktičke ovisnosti između dviju riječi. Primjer matrice u kojoj su samo naznačena pojavljivanja sintaktičkih ovisnosti u frazi "*Ana jede ukusan čokoladan kolač*" nalazi se u tablici 2.2.

Iako se uvođenjem sintaktičkih ovisnosti u izgradnju distribucijskih modela postiže veća ekspresivnost vektora koji opisuju pojedinu riječ, i dalje su takvi modeli ograničeni na modeliranje značenja samo jedne riječi. U problemu modeliranja značenja višerječnih izraza ovako jednostavni modeli nemaju izražen uspjeh, budući da u takvim izrazima nisu bitne samo riječi koje ga čine, već je puno značajniji odnos među njima. Kompozicijski distribucijski modeli, opisani u sljedećem poglavlju, osim distribucije riječi u izrazu uzimaju u obzir i njihov međusoban odnos (kompoziciju izraza), te ovisno o njemu modeliraju značenje cijele fraze.

3. Modeli kompozicijske distribucijske semantike

Modeli kompozicijske distribucijske semantike usmjereni su na prikaz značenja višerječnih izraza u vektorskom prostoru, s posebnim naglaskom na međusoban odnos riječi u izrazu. Obični distribucijski modeli temelje se na vektorima riječi koji opisuju kontekst riječi ekstrahiran iz određenog jezičnog korpusa. Ukoliko bismo željeli na isti način ekstrahirati vektor nekog višerječnog izraza, potrebno bi bilo ekstrahirati vektor iz korpusa za svaki izraz koji želimo prikazati kontekstnim vektorom. S obzirom na bogatstvo izraza koje svaki jezik ima, ovakav pristup je prilično neprikladan i neučinkovit, budući da bi matricu svih mogućih višerječnih izraza bilo praktički nemoguće izgraditi, jer se svaki višerječni izraz niti ne mora nalaziti u pojedinom korpusu. Upravo je to primjer problema rijetkosti (engl. *sparsity problem*), koji se često javlja kod statističkih pristupa obradi prirodnog jezika, Problem rijetkosti očituje se u rijetkim pojavljivanjima određenih izraza u korpusu, zbog kojih se sami vektorski prikaz izraza ne može odrediti u potpunosti, budući da zbog malog broja pojavljivanja izraza nije sasvim jasno definiran kontekst u kojem se izraz koristi. Kao rješenje problema modeliranja višerječnih izraza kontekstnim vektorima nameću se kompozicijski distribucijski modeli, koji primijenjujući razne algebarske operacije na vektore pojedinih riječi dobivene distribucijskim modelima, konstruiraju vektore cjelokupnih višerječnih izraza.

Osnovni modeli kompozicijske distribucijske semantike opisani su radu (Mitchell i Lapata, 2008). Riječ je o modelima koji se temelje na vektorskom zbroju i umnošku pojedinih vektora riječi koje čine jedan višerječni izraz, ali i nekim složenijim operacijama među tim vektorima.

Model temeljen samo na zbroju ili umnošku pojedinih dimenzija vektora najjednostavniji je primjer kompozicijsko distribucijskog modela. U slučaju aditivnog modela, vektor višerječnog izraza računa se kao zbroj svih vektora riječi koje ga čine, dok se u slučaju multiplikativnog modela konačni vektor računa množenjem vrijednosti

u pojedinim dimenzijama svih vektora riječi u izrazu. Primjerice, za aditivni model, kontekstni vektor izraza "*pas lovi mačku*" bio bi jednak zbroju vektora riječi "*pas*", "*loviti*" i "*mačka*":

$$\overrightarrow{\text{pas lovi mačku}} = \overrightarrow{\text{pas}} + \overrightarrow{\text{loviti}} + \overrightarrow{\text{mačka}}$$

Međutim, ovakav pristup i dalje ne uzima u obzir kompoziciju izraza, pa bi tako i vektor drugačijeg izraza, sastavljen od istog skupa riječi, svejedno imao isti konačni vektor. Primjerice vektor izraza "*mačka lovi psa*" bio bi jednak vektoru izraza "*pas lovi mačku*", iako ti izrazi imaju različito semantičko značenje:

$$\overrightarrow{\text{mačka lovi psa}} = \overrightarrow{\text{mačka}} + \overrightarrow{\text{loviti}} + \overrightarrow{\text{pas}}$$

S obzirom na to, Mitchell i Lapata (2008) predlažu težinski aditivni model u kojem bi se vektori dobiveni u međurezultatima zbrajanja ili množenja množili s prethodno definiranim težinama, ovisno o svojoj poziciji u izrazu. Time bi se doprinos vektora riječi konačnom vektoru izraza mijenjao ovisno o poziciji riječi u izrazu.

Predložene verzije modela vrednovali su na ručno označenom skupu izraza sastavljenih od imenice (subjekta) i glagola (predikata). Za svaki par imenice i glagola, označivači su imali ponuđena po dva dodatna glagola, izabrana na temelju razine sličnosti s osnovnim glagolom u WordNet-u.[1] Označivači su među njima trebali označiti u kojoj je mjeri svaki od njih sličan osnovnoj frazi i to brojevima od 1 do 7 (od najmanje do najviše sličnih). Vrednovanje je provedeno računajući kosinusnu mjeru sličnosti između izraza s dodatnim glagolom i osnovnog glagola. Očekivani rezultat bila je veća mjera u slučaju glagola kojeg je najviše označivača označilo kao sličnijeg između dva ponuđena. Također, izračunali su i Spearmanov faktor korelacije (ρ) između oznaka označivača i kosinusne mjere sličnosti između vektora dobivenih pojedinim modelom.

Pokazalo se da jednostavan aditivni model najlošije razlučuje slične od manje sličnih parova izraza, dok su multiplikativni i težinski model postigli bolje rezultate. Spearmanov faktor korelacije za aditivni model iznosio je 0.09, dok su za multiplikativni i težinski model postignute vrijednosti od 0.17, odnosno 0.19.

Ovi rezultati ukazali su na veliki potencijal koji u sebi kriju multiplikativni i težinski modeli. Činjenica da su i jednostavnim množenjem vrijednosti u pojedinim dimenzijama vektora, te pridavanjem težina takvim umnošcima, postignuti dobri rezultati, potaknula je razvoj modela temeljenih na kompliciranijim algebarskim operacijama – tenzorskoj algebri.

[1] WordNet je leksička baza engleskoga jezika, sadrži skupove sinonima za pojedinu riječ: `http://wordnet.princeton.edu`

3.1. Model leksičke funkcije

Prethodno opisani modeli oslanjaju se na osnovne algebarske operacije između vektora koji predstavljaju značenja pojedinih riječi u izrazu. Međutim, niti jedan od tih jednostavnih modela ne uzima u obzir intuitivnu ideju, standardnu u teorijskoj lingvistici, da je semantička kompozicija složenija od obične težinske sume ili umnoška vektora. Općenito, jedna od riječi u izrazu, primjerice pridjev, ponaša se kao funkcija koja utječe na drugu riječ (primjerice imenicu), te tako mijenja njeno značenje. Ovakav pristup kompozicijskoj semantici, motiviran formalnom semantikom prirodnog jezika, potaknuo je razvoj *modela leksičke funkcije* (engl. *lexical function model*) (Baroni i Zamparelli, 2010).

3.1.1. Opis modela

U modelu leksičke funkcije, argumenti funkcije su vektori pojedinih riječi (primjerice vektor imenice na koju se odnosi pridjev), dok su funkcije koje primaju argumente predstavljene tenzorima, pri čemu je red tenzora određen brojem argumenata koje funkcija prima. Tako su tenzori koji predstavljaju djelovanje pojedinih pridjeva na imenice reda 2, odnosno matrice, dok su tenzori tranzitivnih glagola (glagola koji opisuju odnos između subjekta i objekta) reda 3.

Baroni i Zamparelli (2010) u svom su radu opisali efikasan način izračuna matrica koje opisuju djelovanje pridjeva na imenice uz koje se nalaze. Postupak treniranja matrica temelji se na modelu linearne regresije i metode parcijalnih najmanjih kvadrata (engl. *partial least squares*). Na ulazu modela nalaze se vektori imenica ekstrahiranih iz korpusa, dok su na izlazu vektori izraza pridjev–imenica, također ekstrahirani iz korpusa. Dakle, treniranjem matrice modelira se preslikavanje odgovarajućeg vektora imenice u njemu odgovarajući vektor izraza pridjev–imenica, čime se efektivno opisuje djelovanje pridjeva na vektora imenice u vektorskom prostoru, a time i na semantičko značenje tako dobivenog izraza.

Metoda parcijalnih najmanjih kvadrata opisana je u (Ng, 2013), a temelji se na dekompoziciji matrice ulaznih vektora ($\mathbf{X}$) i matrice izlaznih vektora ($\mathbf{Y}$) analizom glavnih komponenti (engl. *principal component analysis*). Korištenjem tog postupka, matrice $\mathbf{X}$ i $\mathbf{Y}$ definiraju se kao:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathbf{T}}$$

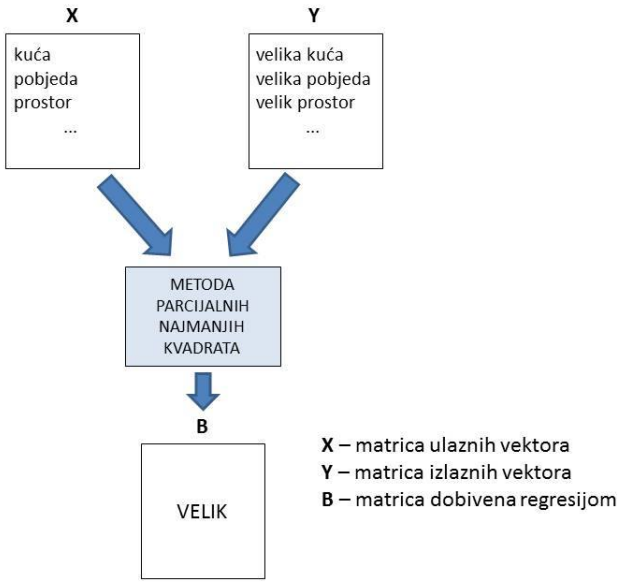
gdje su $\mathbf{T}$ i $\mathbf{U}$ projekcijske matrice matrica $\mathbf{X}$ i $\mathbf{Y}$ ili rezultatske matrice (engl. *score*

matrices), a matrice $\mathbf{P}$ i $\mathbf{Q}$ matrice glavnih komponenti (engl. *principal component matrices*) (Ng, 2013). Ideja metode parcijalnih najmanjih kvadrata je u zajedničkoj petlji postupno obavljati dekompoziciju obiju matrica ($\mathbf{X}$ i $\mathbf{Y}$), pri čemu se određuju linearne kombinacije varijabli iz obiju matrica, maksimizirajući njihovu kovarijancu. Na taj način se, nakon prolaska kroz matrice, dobiva skup jednadžbi koje opisuju povezanost ulaznih varijabli iz matrice $\mathbf{X}$ i izlaznih varijabli iz matrice $\mathbf{Y}$, pri čemu koeficijenti iz tih jednadžbi zapravo određuju matricu koja opisuje takvo preslikavanje.

Spomenutom metodom parcijalnih najmanjih kvadrata računa se matrica koja opisuje preslikavanje n-dimenzionalnih vektora imenica ekstrahiranih iz korpusa u vektore parova pridjev–imenica, također ekstrahirane iz korpusa. Takva matrica zapravo modelira značenje pridjeva, budući da je kreirana pomoću kontekstnih vektora svih parova pridjev–imenica ekstrahiranih iz korpusa. Primjerice, za izračun matrice pridjeva *velik*, ekstrahiraju se iz korpusa svi parovi pridjev-imenica u kojem se kao pridjev pojavljuje pridjev *velik*, te ti parovi predstavljaju izlaz modela linearne regresije, koji se želi predvidjeti (npr. *velika kuća*, *velika pobjeda*, *velik prostor*). Kao ulaz modela koriste se odgovarajući ekstrahirani vektori imenica koje se pojavljuju u pojedinoj kombinaciji pridjev–imenica (npr. *kuća*, *pobjeda*, *prostor*). Matrica dobivena postupkom treniranja veličine je $n \times n$, te je u njoj opisano značenje pridjeva u kombinaciji s odabranom imenicom. Vektor koji opisuje značenje izraza pridjev–imenica dobiva se množenjem matrice odgovarajućeg pridjeva i vektora imenice koja se uz njega koristi. Primjer izgradnje matrice za pridjev *velik* dan je na slici 3.1.

Isti postupak kasnije je primijenjen i na izračun matrica za netranzitivne glagole i određene članove (engl. *determiners*). Grefenstette et al. (2013) proširili su model i na tranzitivne glagole, izgradivši za njih tenzore reda 3, u kojima su sadržane obje funkcije glagola – djelovanje glagola na subjekt i na objekt. Postupak određivanja vrijednosti tenzora za pojedini glagol također se temelji na regresiji, međutim u ovom slučaju regresija se provodi u više koraka.

U prvom se koraku određuju matrice koje opisuju djelovanje izraza glagol–objekt na subjekt i to na sličan način kao i u slučaju matrice za pridjeve – koristeći model linearne regresije uz metodu najmanjih kvadrata. Međutim u ovom slučaju izlazni vektori nisu sastavljeni samo od dvije riječi, već od tri. Primjerice, za računanje matrice koja opisuje djelovanje izraza *jesti meso*, izlazni vektori su *pas jesti meso*, *dječak jesti meso*, *gost jesti meso*, dok su ulazni vektori vektori imenica koje predstavljaju subjekt u spomenutim izrazima: *pas*, *dječak*, *gost*. Dobivena matrica zapravo opisuje značenje izraza *jesti meso* u kombinaciji s raznim subjektima. Na isti način formuliraju se matrice za sve ostale kombinacije objekata koji se mogu pojaviti uz glagol *jesti*



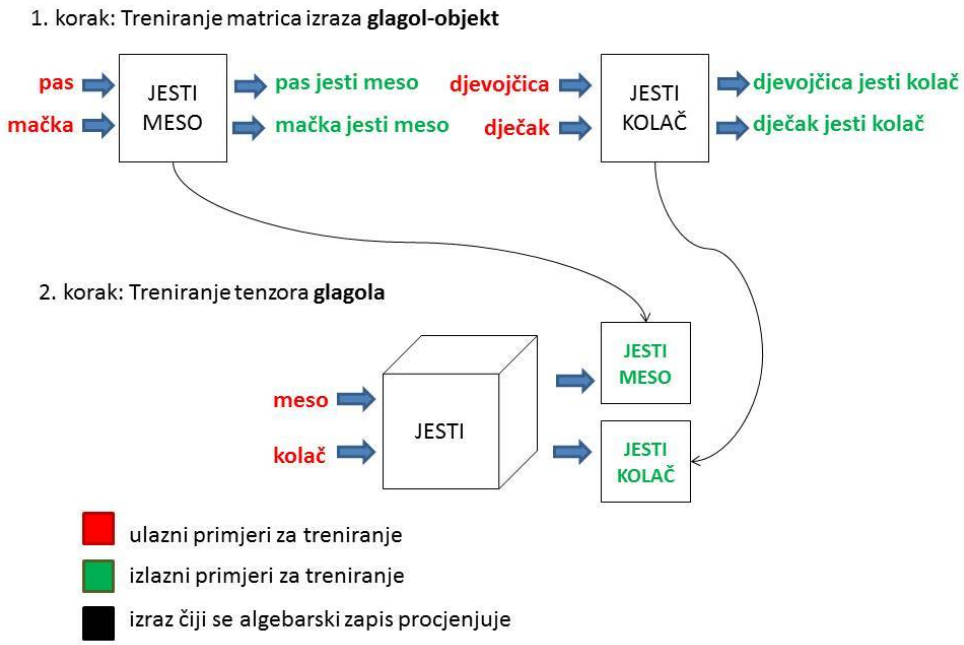
Slika 3.1: Računanje matrice za pridjev *velik* korištenjem metode parcijalnih najmanjih kvadrata. Ulazne matrice su matrice $\mathbf{X}$ i $\mathbf{Y}$ koje sadrže ulazne i izlazne vektore značenja, dok je matrica $\mathbf{B}$ izlazna matrica dobivena metodom parcijalnih najmanjih kvadrata, u kojoj su sadržani parametri koji opisuju djelovanje pridjeva "*velik*" na riječi uz koje se nalazi.

(*jesti kolač*, *jesti ručak* i drugi).

U drugom koraku određuje se tenzor koji opisuje potpuno djelovanje glagola i na subjekt i na objekt. Tenzor se također određuje postupkom linearne regresije uz metodu najmanjih kvadrata, međutim u ovom slučaju izlazni podatci modela nisu vektori, već matrice oblika glagol–objekt. Primjerice, pri treniranju tenzora za glagol *jesti*, izlazni podatci su matrice *jesti meso*, *jesti kolač*, *jesti ručak*, dok su ulazni podatci vektori objekata iz navedenih matrica: *meso*, *kolač*, *ručak*. Tim postupkom određuje se tenzor kojim se, množenjem tenzora vektorom objekta uz koji se glagol koristi, može lako izračunati matrica koja opisuje djelovanje tog para glagol–objekt uz razne subjekte. Daljnjim množenjem dobivene matrice i vektora subjekta, dobiva se konačni vektor koji opisuje značenje cijele fraze.

Na slici 3.2, opisan je postupak treniranja tenzora za tranzitivan glagol *jesti*.

Svi opisani modeli temeljeni na modelu leksičke funkcije postigli su značajno bolje rezultate u usporedbi s ostalim tada dostupnim modelima. Model opisan u radu Baroni i Zamparelli (2010), u kojem je predstavljen model matrica koje sadržavaju značenja pojedinih pridjeva, pokazao se najuspješnijim u modeliranju vektora značenja para pridjev-imenica koji su najsličniji stvarnim vektorima pridjev-imenica ekstrahiranim iz



Slika 3.2: Računanje tenzora za glagol *jesti* koristeći linearnu regresiju u dva koraka.

korpusa. Tenzorski model tranzitivnih glagola iz rada Grefenstette et al. (2013) također se pokazao znatno boljim od dotad predloženih modela, i to u problemu identificiranja sličnosti između višerječnih fraza. Naime, model je ispitan na ručno označenom skupu podataka koji je sastavljen od parova izraza oblika subjekt–glagol–objekt, u kojem su označivači za svaki par izraza označili u kojoj su mjeri slični. Tenzorski model pokazao se najboljim u predviđanju sličnosti parova izraza u istoj mjeri kao i označivači – Spearmanov faktor korelacije između sličnosti dobivene tenzorskim modelom i oznaka označivača iznosio je 0.32, dok je sljedeći najbolji rezultat postignut multiplikativnim modelom iznosio 0.25.

3.1.2. Problemi pri primjeni modela na rečenice

Iako je pristup temeljen na leksičkoj funkciji dosta uspješno riješio probleme prikaza semantičkog značenja manjih višerječnih izraza, sam postupak nije praktično primjenjiv na duže višerječne izraze kao što su rečenice.

Najveći problem predstavljaju veličine vektora, matrica i ostalih tenzora u kojima su sadržana značenja pojedinih riječi. Naime, ako su vektori imenica veličine 300 dimenzija, matrice pridjeva trebale bi sadržavati 300^2 vrijednosti, dok bi broj vrijednosti tenzora tranzitivnih glagola bio $300^3 = 27.000.000$. Očito je da je i za samo te vr-

ste riječi broj vrijednosti koje se trebaju procijeniti modelom regresije prilično velik i neprikladan za efikasnu pohranu i korištenje pri izračunu semantičkih reprezentacija duljih rečenica. S tako velikim dimenzijama javlja se i problem rijetkih podataka (engl. *data sparseness*) za riječi koje se rijetko pojavljuju u korpusu. Primjerice, u slučaju treniranja tenzora za prethodno spomenuti glagol *eat*, potrebno je imati dovoljan broj različitih izraza glagol–objekt, pri čemu se svaki od tih izraza pojavljuje uz dovoljan broj različitih imenica kao subjekata, za koje je također poželjno da se dovoljno često pojavljuju u korpusu kako bi imale dovoljno jasno definirane vektore. Očito nije razumno očekivati da će za svaki glagol koji se pojavljuje u korpusu ovi uvjeti biti zadovoljeni, pa je samim time i kvaliteta izgrađenih reprezentacija značenja riječi upitna, odnosno ovisna o frekvenciji pojavljivanja pojedinih riječi u korpusu. Problem je još izraženiji kod onih vrsta riječi koje imaju više od dva argumenta, primjerice prilozi koji mijenjaju značenje netranzitivnih glagola bili bi predstavljeni tenzorima s ukupno $300^4 = 8.100.000.000$ vrijednosti.

Drugi problem predstavlja činjenica da se riječima istog značenja, a korištenima u različitim sintaktičkim ulogama, pridaju različite reprezentacije značenja koje međusobno nisu usporedive. Primjerice, glagol *jesti* može biti korišten i u tranzitivnom i u netranzitivnom obliku, pri čemu je za tranzitivan oblik njegovo značenje prikazano u obliku tenzora, dok je u slučaju netranzitivnog oblika korištena matrica. Očito je da ne postoji jedinstven zapis u kojem je prikazano kompletno značenje glagola *jesti*, a pogotovo ako uzmemo u obzir da glagol može biti i u drugom obliku (pasivu). Glagolske imenice, kao što je *uništenje*, bile bi prikazane u vektorskom obliku, dok bi za njih odgovarajući glagoli, u ovom slučaju glagol *uništiti*, bili prikazani u obliku tenzora, pri čemu bi opet semantičko značenje riječi bilo razdvojeno u najmanje dva zapisa.

Posljednji problem predstavlja modeliranje značenja riječi u slučaju kada se one koriste u onom obliku koji nije pronađen u korpusu. Naime, model leksičke funkcije prikazuje značenja samo onih konstrukcija koje se u korpusu pojavljuju, pri čemu ne postoji rješenje za modeliranje semantičkog značenja onih konstrukcija koje nisu u takvom obliku prisutne u korpusu. Primjerice, ukoliko je u korpusu glagol *jesti* pronađen samo u tranzitivnom obliku (npr. *pas jede meso*), model leksičke funkcije za njega bi izgradio samo tenzor reda 3 kojim bi modelirao značenje cjelokupnog izraza sastavljenog od subjekta, glagola i objekta. Međutim, ako želimo modelirati značenje fraze u kojoj je glagol *jesti* u netranzitivnom obliku (npr. *pas jede*), ne postoji matrica koja bi nam taj odnos subjekta i objekta opisala, budući da se glagol *jesti* u korpusu pojavljuje samo u tranzitivnom obliku, pa u postupku treniranja reprezentacija riječi takav odnos subjekt i glagola *jesti* nije predviđen. Ovaj problem je sam po sebi neiz-

bježan, budući da je naivno očekivati da će se u korpusu svaka riječ u jeziku pojaviti u svakom mogućem obliku u kojem se može upotrijebiti.

4. Praktični model leksičke funkcije – PLF

Problemi prikaza značenja riječi tenzorima viših redova, prvenstveno veličina tenzora i problem rijetko popunjenih matrica i tenzora uzrokovanih rijetkim pojavljivanjem određenih izraza u korpusu, pokušali su riješiti Paperno et al. (2014) opisavši u svom radu *praktičan model leksičke funkcije*. U svom modelu željeli su i dalje različito predstavljati značenje riječi ovisno o sintaktičkoj uporabi pojedine riječi u izrazu, ali i uspostaviti poveznicu između različitih uporaba riječi, primjerice između tranzitivnih i netranzitivnih oblika glagola. Također, željeli su i svaku riječ prikazati zajedničkim osnovnim načinom, primjerice vektorom, kojim bi bilo opisano osnovno značenje svake riječi.

Uspjeli su izgraditi model temeljen na modelu leksičke funkcije i ideji riječi kao funkcija koje djeluju na ostale riječi u izrazu. Svoj model nazvali su praktičnim modelom leksičke funkcije (engl. *practical lexical function model*) – PLF. U PLF-modelu, pojedina riječ nije predstavljena samo vektorom ili samo tenzorom višeg reda, već je svaka riječ predstavljena svojim vektorom, te dodatno matricama, čiji broj ovisi o broju argumenata na koje riječ kao funkcija djeluje. Primjerice, tranzitivni glagoli predstavljeni su vektorom, matricom koja opisuje djelovanje glagola na subjekt, te još jednom matricom koja opisuje djelovanje glagola na objekt. Množenjem matrica i odgovarajućih vektora riječi argumenata u izrazu dobivaju se vektori čiji zbroj predstavlja konačan vektor značenja cjelokupnog izraza.

4.1. Kompozicijska semantika u PLF-modelu

U PLF-modelu sve su riječi predstavljene pomoću vektora, dok je riječima koje se ponašaju kao funkcije prema drugim riječima u jeziku dodijeljen odgovarajući broj matrica. Općeniti oblik prikaza semantičkog značenja riječi može se definirati kao

n-torka sastavljena od jednog vektora i n matrica:

$$\langle \vec{x}, \mathbf{X_1}, ..., \mathbf{X_n} \rangle$$

Broj matrica u n-torci određen je brojem riječi na koje određena riječ može djelovati kao funkcija. Svaka matrica odgovara jednoj funkciji riječi, te svaka riječ ima onoliko matrica koliko argumenata može primiti kao funkcija: imenice nemaju niti jednu matricu, pridjevi i netranzitivni glagoli po jednu, dok tranzitivni glagoli imaju po dvije matrice – jednu za subjekt, jednu za objekt.

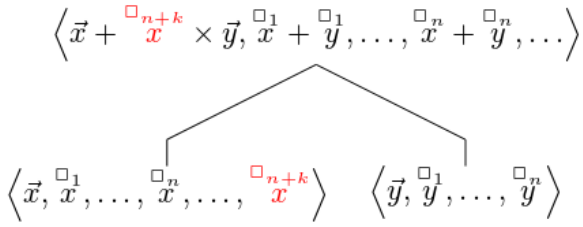
Praktični model leksičke funkcije temelji se na dva pravila kompozicijske semantike:

1. pravilu primjene funkcije za računanje sa strukturama različitih dimenzija i
2. pravilu simetrične kompozicije za računanje sa strukturama istih dimenzija.

Ta dva pravila kombiniraju dva uspješna modela kompozijske semantike: model leksičke funkcije i jednostavan aditivni model.

Prvo pravilo odnosi se na postupak računanja konačnog vektora pri djelovanju riječi kao funkcije na svoje argumente. Primjerice, u slučaju tranzitivnih glagola, matrica objekta tranzitivnog glagola množi se s vektorom imenice koja predstavlja objekt u izrazu, dok se matrica subjekta glagola množi s vektorom subjekta (imenice).

Na slici 4.1 prikazana je općenita primjena pravila u slučaju riječi s različitim brojem pridijeljenih matrica. Kompozicija opisana na slici odnosi se na dvije riječi, prikazane vektorima $\vec{x}$ i $\vec{y}$, pri čemu svaka od njih ima i određeni broj matrica uz vektor, koje opisuju djelovanje te riječi kao funkcije na riječi u njenoj okolini. Riječ x ima $n + k$ matrica, dok riječ y ima n matrica. U kompoziciji opisanoj na slici, riječ x ponaša se kao funkcija koja djeluje na riječ y (koja, dakle, predstavlja argument funkcije). Pravilo kompozicije PLF-modela kaže da se posljednja matrica riječi koja djeluje kao funkcija (matrica $n + k$ riječi x) množi s vektorom riječi y, a preostale matrice riječi x i y se zbroje (matrice uključivo do n) ili se višak matrica neke riječi prenese u skup matrica dobivenog izraza (matrice od $n + 1$ do $n + k$ riječi x). Na taj način grade se matrice i vektori značenja dobivenog izraza, koje opisuju njegovo djelovanje na riječi koje se mogu naći uz njega. Takvim postupkom kompozicije algebarskih zapisa značenja riječi dobiva se n-torka zapisa koje se dalje može rekurzivno primijenjivati na ostale riječi na koje dobiveni izraz može djelovati. Upravo u tome se očituje sposobnost PLF-modela da korištenjem kompozicijskih pravila, iz zapisa značenja riječi u izrazu izgradi zapis značenja veće jezične jedinice – izraza ili rečenice.



Slika 4.1: Primjena pravila primjene funkcije za računanje sa strukturama različitih dimenzija, preuzeto iz rada (Paperno et al., 2014).

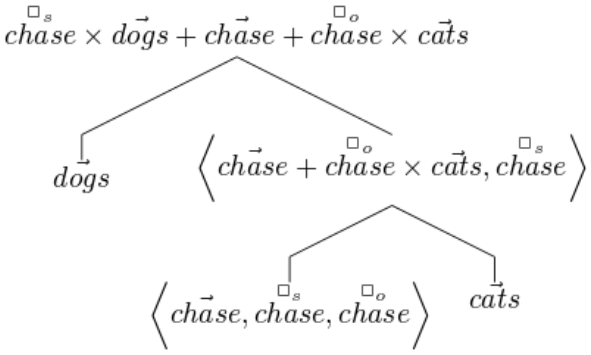
pjevati: $\overrightarrow{pjevati}$, $\mathbf{M}_{pjevati}$	plesati: $\overrightarrow{plesati}$, $\mathbf{M}_{plesati}$
pjevati i plesati:	$\overrightarrow{pjevati} + \overrightarrow{plesati}$, $\mathbf{M}_{pjevati} + \mathbf{M}_{plesati}$

Tablica 4.1: Primjeri simetrične kompozicije

Drugo pravilo odnosi se na slučaj računanja konačnog vektora između riječi čije su reprezentacije iste veličine: dva vektora, dvije matrice, i dr. Pravilo nalaže da se takvi oblici značenja riječi jednostavno zbroje u konačan vektor značenja izraza. Ovo pravilo posebno dolazi do izražaja u slučaju izraza u kojima nije potpuno jasan odnos između riječi funkcije i riječi argumenta. Primjerice u izrazu *pjevati i plesati* nije jasno koja od riječi je funkcija, a koja argument, budući da PLF-model veznicima ne pridaje semantičko značenje, već ih smatra "praznim" elementima koji nemaju svoj vektor (ili tenzor) značenja. Samim time, poštujući pravilo simetrične kompozicije, konačan zapis značenja izraza *pjevati i plesati* bio bi jednak zbroju vektora glagola *pjevati* i *plesati* koji bi činio vektor značenja izraza, te zbroju njihovih matrica kao drugom elementu koji opisuje djelovanje glagola na svoje argumente. Primjer takve simetrične kompozicije dan je u tablici 4.1.

Navedena pravila ukazuju na činjenicu da je PLF-model zapravo dorađena verzija jednostavnog aditivnog modela, u kojoj se vektori riječi argumenata množe matricama riječi funkcija, pri čemu se mijenja značenje riječi argumenata u ovisnosti o značenju riječi funkcije, te se tako dobiveni vektori zbrajaju u konačan vektor značenja izraza. Primjer korištenja pravila o kompoziciji PLF-modela u izračunu vektora značenja izraza *dogs chase cats* (*psi love mačke*), dakle sastavljenog od tranzitivnog glagola, nalazi se na slici 4.2. U prvom koraku računa se vektor koji opisuje izraza glagol–objekt (*chase cats*) te se tako dobiveni vektor pribraja vektoru koji opisuje značenje glagola. Nakon toga, slijedi računanje vektora izraza subjekt–glagol (*dogs chase*) koji se pribraju prethodno dobivenom zbroju dva vektora. Konačni vektor opisuje znače-

nje cijelog izraza, a sastavljen je od zbroja tri vektora: vektora glagola, vektora izraza glagol–objekt i vektora izraza subjekt–glagol.



Slika 4.2: Primjena pravila semantičke kompozicije na tranzitivan izraz. Preuzeto iz rada Paperno et al. (2014)

Važno je napomenuti da se u ovako definiranoj kompoziciji modela svi vektori dobiveni umnoškom matrice i vektora značenja pojedinih riječi normaliziraju na duljinu 1, kao i zbroj vektora dobivenih množenjem matrice i vektora s vektorom odgovarajućeg pridjeva ili glagola. Primjerice, pri izračunu konačnog vektora značenja izraza pridjev–imenica, umnožak matrice pridjeva i vektora imenice normalizira se na duljinu 1, nakon čega se zbroje tako normaliziran vektor i vektor pridjeva te se i tako dobiveni zbroj vektora normalizira na duljinu 1. Dobiveni normalizirani zbroj predstavlja vektor značenja izraza pridjev–imenica. Normalizacija vektora predstavlja svojevrsno povećanje povjerenja u sami model, budući da se kompozicija izraza temelji na zbroju vektora značenja dijelova izraza, pri čemu je duljina vektora od velikog utjecaja na izgled konačnog vektora. Normalizacijom vektora dijelova izraza na istu duljinu postiže se jednak utjecaj značenja dijelova izraza u značenju cjelokupnog izraza.

Praktični model leksičke funkcije uspješno je riješio i probleme koji su bili izraženi u predloženom osnovnom modelu leksičke funkcije. Prvenstveno se to odnosi na problem visoke dimenzionalnosti zapisa semantičkog značenja pojedinih riječi – tenzori reda tri ili više. U PLF-modelu, najveći zapisi su matrice, čiji broj raste linearno u ovisnosti o broju argumenata riječi, a ne eksponencijalno kao u slučaju osnovnog leksičko funkcijskog modela. Takav matrični zapis funkcija riječi dovodi do jednostavnije i efikasnije procjene parametara matrice, njihovog pohranjivanja i u konačnici računanja s tim vrijednostima.

Kao posljedica takve arhitekture, kako bi se modelirao tenzor višeg reda u slučaju riječi funkcija koje primaju više od jednog parametranije više potrebno provoditi postupak linearne regresije u više koraka. Primjerice, za tranzitivne glagole odvo-

jeno se treniraju matrični prikazi za subjekt glagola pomoću vektora subjekta i parova subjekt–glagol, te odvojeno matrice za objekt glagola pomoću vektora objekata i parova glagol–objekt. S takvim pristup očekivano je da će se u korpusu pojaviti dovoljan broj različitih kombinacija parova subjekt–glagol i glagol–objekt, koji u ovakvom pristupu ne moraju nužno biti dio istog izraza subjekt–glagol–objekt, čime se riješava problem rijetkih podataka pri treniranju višedimenzionalnog prikaza značenja glagola.

Prikazi semantičkog značenja riječi u PLF-modelu uključuju osnovni vektorski prikaz za svaku riječ u jeziku, čime se omogućuje usporedba riječi istog značenja korištenih u različitim sintaksnim oblicima u korpusu (*uništiti* – glagol, *uništenje* – imenica).

Odvojeno modeliranje funkcija pojedinih riječi za različite argumente olakšava i računanje konačnog vektora izraza ovisno o samoj konstrukciji izraza. Primjerice, za glagole koji mogu biti i tranzitivni i netranzitivni, pri računanju konačnog vektora izraza koriste se samo one matrice čiji se argumenti u izrazu pojavljuju. Drugim riječima, ako je glagol u izrazu korišten u netranzitivnom obliku, dakle bez objekta, u računanju konačnog vektora koristiti će se samo matrica subjekta glagola, dok se u modelu leksičke funkcije u takvim slučajevima koristio cijeli tenzor reda tri, pri čemu je rezultat množenja bila matrica.

Kao rješenje problema korištenja riječi u sintaktičkoj ulozi u kojoj ista nije pronađena u korpusu, Paperno et al. (2014) predlažu jednostavne načine izostavljanja odgovarajućih matrica, odnosno dodavanja jediničnih matrica u spomenutim slučajevima. Primjerice, ukoliko se glagol koji je u korpusu pronađen samo u tranzitivnom obliku, koristi u netranzitivnom obliku, matrica objekta jednostavno se ispusti iz računanja konačnog vektora značenja izraza. S druge strane, ukoliko se glagol koji je u korpusu pronađen samo u netranzitivnom obliku, upotrijebi u tranzivitnom izrazu, kao matrica objekta može se upotrijebiti jedinična matrica, čime se signalizira neznanje o utjecaju tog glagola na objekt uz koji se koristi.

Sveukupno gledano, PLF-model predstavlja proširenje modela leksičke funkcije koji zadržava dobre odlike modela i riješava naglašene nedostatke istoga. U njemu se zadržava lingvistički motivirana ideja o semantičkoj kompoziciji kao djelovanju pojedinih riječi kao funkcija na riječ uz koju se nalaze, ali je ta ideja ostvarena na praktičniji i pristupačniji način.

4.2. Izgradnja modela

Paperno et al. (2014) izgradili su i ispitali model za engleski jezik. Kao korpus za ekstrakciju vektora riječi i parova različitih kombinacija riječi (subjekt–glagol, glagol–

objekt, pridjev–imenica) korišteni su:

- ukWaC – korpus sastavljen od web-stranica domene .uk,[1] lematiziran, te su u njemu označene gramatičke kategorije riječi (engl. *POS-tagged*), kao i međuovisnosti između riječi u pogledu uloge riječi u izrazu (engl. *dependency based parsing*);
- Wikipedia dump[2] – korpus sastavljen od stranica engleske Wikipedije iz 2009. godine, također predobrađen istim alatima kao i ukWaC;
- British National Corpus[3] – korpus sastavljen od izraza pisanog i govornog engleskoga jezika, prikupljenih iz različitih izvora, prethodno predobrađen istim alatima kao i prethodna dva korpusa.

Združeni korpus, koji čine navedena tri dijela, sastoji se od oko 2.8 milijarde riječi.

Za prikaz vektora riječi izgrađena je matrica supojavljivanja (engl. *co-occurrence matrix*) sastavljena od 30,000 redaka i isto toliko stupaca, koju čini 30,000 najfrekventnijih riječi u združenom korpusu. Dobiveni vektori supojavljivanja transformirani su postupkom izračuna pozitivne uzajamne zajedničke informacije (engl. *positive pointwise mutual information*) u mjere povezanosti pojedinih riječi u matrici. Nakon toga, matrici je postupkom singularne dekompozicije (engl. *Singular Value Decomposition – SVD*) smanjena dimenzionalnost sa $30,000 \times 30,000$ na 300×300. Tako dobiveni vektori sastavljeni od 300 vrijednosti, dodatno su normalizirani na duljinu 1. Osim vektora riječi, iz korpusa su na isti način ekstrahirali i vektore parova subjekt–glagol, glagol–objekt i pridjev–imenica, za sve glagole i objekte iz skupa podataka na kojem je model ispitivan. Iz skupa tako ekstrahiranih vektora za treniranje matrica modela korišteni su vektori samo onih izraza sa frekvencijom pojavljivanja u korpusu većom od 5, čime se želi izbjeći korištenje nedovoljno dobro definiranih vektora.

Nakon što su dobiveni svi vektori imenica, pridjeva i glagola, kao i svi parovi odgovarajućih kombinacija među njima, trenirali su matrice za pridjeve, te matrice za subjekte i objekte glagola. Za treniranje matrica korišten je postupak L2-regularizirane regresije (Ridge regresija) uz korištenje generalizirane unakrsne provjere.

4.2.1. L2-regularizirana regresija

L2-regularizirana regresija najčešće je korištena varijanta linearne regresije u slučajevima loše postavljenih problema (engl. *ill-posed problems*). Loše postavljenim proble-

[1] Korpus ukWaC dostupan je na adresi: `wacky.sslmit.unibo.it`

[2] Opisan i dostupan na službenim stranicama Wikipedije: `en.wikipedia.org`

[3] Dostupan je na adresi: `www.natcorp.ox.ac.uk`

mima smatraju se oni problemi u kojima je u postupku treniranja potrebno izračunati inverz matrice, ali matrica čiji se inverz računa nije kvadratna, pa nema inverz, ili je kvadratna, ali nije konzistentna (primjerice, ne sastoji se od različitih redaka). Pregled Ridge regresije dan u nastavku napravljen je prema postupku opisanom u Alpaydin (2010).

Kako bi se što bolje procijenila željena vrijednost za koju se regresijski model trenira, uvodi se regularizacijski faktor kojim se sprječava prenaučenost modela. Funkcija pogreške takvog modela regresije jednaka je:

$$E = \frac{1}{2}\sum_{i=1}^{N}(\mathbf{w}^{\mathbf{T}}\mathbf{x}^{(i)} - y^{(i)})^2 + \frac{\lambda}{2}\left\|\mathbf{w}\right\|^2 \tag{4.1}$$

Vektor $\mathbf{w}$ predstavlja težine regresijskog modela koje se određuju postupkom treniranja, vektor $\mathbf{x}$ ulazni vektor modela, dok je vrijednost y izlazna vrijednost koja se procijenjuje (za dani ulazni vektor). Vrijednost λ predstavlja regularizacijski parametar kojim se sprječava prenaučenost modela. Naime, porastom regularizacijskog parametra povećava se pogreška modela na skupu za treniranje, ali se ujedno povećava i sposobnost predviđanja modela na dotad neviđenim primjerima.

Ovako definirana funkcija pogreške je derivabilna te se rješenjem u zatvorenoj formi mogu procijeniti težine modela ($\mathbf{w}$). Deriviranjem izraza 4.1 po vektoru $\mathbf{w}$, dobiva se rješenje u zatvorenoj formi:

$$\mathbf{w} = (\mathbf{\Phi}^{\mathbf{T}}\mathbf{\Phi} + \lambda\mathbf{I})^{-1}\mathbf{\Phi}^{\mathbf{T}}\mathbf{y} \tag{4.2}$$

Matrica $\mathbf{\Phi}$ predstavlja dizajn matricu – matricu ulaznih vektora modela, na koje su prethodno primijenjene odgovarajuće bazne funkcije za transformaciju ulaznih vrijednosti, dok je matrica $\mathbf{I}$ jedinična matrica. Izraz $(\mathbf{\Phi}^{\mathbf{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathbf{T}}$ naziva se pseudoinverz matrice, te je zapravo generalizacija standardnog inverza matrice, koji se koristi u slučajevima kada standardni inverz određene matrice nije moguće odrediti (loše postavljen problem).

Vrijednost parametra λ određuje se nekim od postupaka unakrsne provjere modela, primjerice postupkom unakrsne provjere na odvojenom skupu podataka za ispitivanje. U tom postupku za konačnu vrijednost parametra odabire se ona vrijednost za koju je pogreška modela na skupu za ispitivanje najmanja, čime se omogućuje što točniji izračun izlaznih vrijednosti modela za prethodno neviđene ulazne primjere, spriječavajući prenaučenost modela na primjere iz skupa za treniranje.

Opisani postupak Ridge regresije odnosi se na slučajeve u kojima se na izlazu modela očekuje samo jedna brojčana vrijednost (y je skalar). U PLF-modelu izlaz regresijskog modela je vektor. Dakle, u tako definiranom problemu izmijenjena je i

jednadžba koja opisuje rješenje regresijskog problema u zatvorenoj formi:

$$\mathbf{W} = (\mathbf{\Phi^T \Phi} + \lambda \mathbf{I})^{-1} \mathbf{\Phi^T Y} \tag{4.3}$$

U jednadžbi 4.3 težine modela više nisu prikazane vektorom nego matricom $\mathbf{W}$, kao i izlazni vektori u matrici $\mathbf{Y}$.

Kako bi se uspješno primijenio postupak regresije na procjenu parametara matrice PLF-modela, problem se dekomponira na zasebne regresijske probleme, pri čemu je broj regresijskih modela određen brojem stupaca matrice izlaznih vektora $\mathbf{Y}$. Na taj se način u svakom koraku postupka procjenjuje jedan stupac matrice $\mathbf{W}$ koja predstavlja težine modela. Važno je naglasiti da se u takvom prolasku kroz manje regresijske probleme u svakom od problema koristi isti regularizacijski parametar λ, koji se na kraju optimizira na razini cjelokupne matrice $\mathbf{W}$ sastavljene od stupaca težina, koristeći postupak generalizirane unakrsne provjere opisan u nastavku.

4.2.2. Generalizirana unakrsna provjera

Unakrsna provjera modela obavlja se kako bi se utvrdio generalizacijski parametar λ pomoću kojeg bi model što uspješnije predviđao izlazne vrijednosti primjera koji nisu prethodno viđeni. Općenito, unakrsna provjera modela obavlja se na skupu podataka koji je različit od skupa na kojem je model učen. Time se provjerava upravo sposobnost predviđanja izlaza modela na dotad neviđenim primjerima. Osim takvog pristupa, jedan od ostalih mogućih je i pristup izostavljanja jednog primjera za unakrsnu provjeru (engl. *leave-one-out cross-validation*), u kojem se model trenira na $n - 1$ primjera za treniranje (n je ukupan broj primjera za treniranje), te se njegova uspješnost provjerava na izostavljenom primjeru. Prolaskom kroz cijeli skup primjera za treniranje, odnosno izostavljanjem novog primjera iz skupa za treniranje u svakom prolasku, određuje se parametar λ koji omogućuje najbolju generalizaciju modela.

Generalizirana unakrsna provjera aproksimacija je pristupa temeljenog na izostavljanju jednog primjera za unakrsnu provjeru. Koristi se u linearnoj regresiji u kojoj se funkcija gubitka definira pomoću kvadratnog gubitka. Opisana je u radu (Golub et al., 1979). Temelji se na matrici $\mathbf{S}$ (u literaturi znanoj i kao *hat matrix*) kojom se, opisuje odnos između stvarnih izlaznih vrijednosti primjera za treniranje i vrijednosti dobivenih regresijskim modelom:

$$\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$$

Vektor $\hat{\mathbf{y}}$ sadrži izlazne vrijednosti dobivene modelom, a vektor $\mathbf{y}$ stvarne izlazne vrijednosti primjera za treniranje. U slučaju matrica PLF-modela, nije riječ o vektorima

$\hat{\mathbf{y}}$ i $\mathbf{y}$, već o matricama $\hat{\mathbf{Y}}$ i $\mathbf{Y}$, budući da su izlazne vrijednosti modela zapravo vektori. Zbog toga je u tom slučaju njihov odnos opisan kao:

$$\hat{\mathbf{Y}} = \mathbf{S}\mathbf{Y}$$

Matrica $\mathbf{S}$ definirana je na sljedeći način:

$$\mathbf{S} = \mathbf{\Phi}(\mathbf{\Phi^T}\mathbf{\Phi} + \lambda\mathbf{I})^{-1}\mathbf{\Phi^T}$$

gdje je matrica $\mathbf{\Phi}$ dizajn matrica ulaznih vektora modela. Zapravo je matrica $\mathbf{S}$ jednaka umnošku dizajn matrice $\mathbf{\Phi}$ i pseudoinverza dizajn matrice u kojoj je uključen i regularizacijski faktor λ, pri čemu su njene dimenzije jednake $n \times n$, gdje je n broj ulaznih vektora modela.

U radu (Golub et al., 1979) detaljno je prikazan dokaz kojim je potvrđeno da se pomoću ovako definirane matrica može prilično uspješno procijeniti broj parametara modela, i to pomoću zbroja elemenata na dijagonali kvadratne matrice $\mathbf{S}$, koji predstavlja efektivan broj parametara modela. Koristeći tu procjenu izvedena je i aproksimacija generalizirane unakrsne provjere koja se definira kao:

$$\text{GCV}(\hat{\mathbf{Y}}) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\|\mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i}\|}{1 - \text{trace}(\mathbf{S})/N}\right]^2 \tag{4.4}$$

gdje je N broj ulaznih vektora modela, $\|\mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i}\|$ euklidska udaljenost između stvarnog izlaznog vektora primjera i i vektora dobivenog modelom, dok je vrijednost $\text{trace}(\mathbf{S})$ jednaka zbroju vrijednosti na glavnoj dijagonali matrice $\mathbf{S}$, odnosno predstavlja efektivan broj parametara modela. Pomoću ovako definiranog izraza za unakrsnu provjeru vrednuju se parametri regularizacije modela (koji su ugrađeni u matricu $\mathbf{S}$) i odabire se onaj regularizacijski parametar za kojeg izraz 4.4 daje minimalnu vrijednost. Model se treniran s određenim brojem regularizacijskih parametara postupkom dekompozicije regresijskog problema na više manjih regresijskih problema, pri čemu se za svaku tako dobivenu matricu odredi vrijednost jednadžbe 4.4, te se u konačnici odabere onaj regularizacijski parametar za koji je vrijednost jednadžbe minimalna.

Generalizirana unakrsna provjera tako se temelji na minimizaciji izraza 4.4 u smislu minimalne prosječne kvadratne pogreške modela uzimajući u obzir vrijednosti na glavnoj dijagonali matrice $\mathbf{S}$. Ovako definirana unakrsna provjera predstavlja brz i efikasan način izračuna regularizacijskog parametra u modelima linearne regresije (primjerice u opisanom modelu Ridge regresije korištene u PLF-modelu) kojim se aproksimira postupak unakrsne provjere izostavljanjem jednog primjera.

4.3. Vrednovanje modela

Izgrađeni model vrednovan je na nekoliko različitih skupova podataka sastavljenim od višerječnih izraza. Izrazi u skupovima podataka sastavljeni su na različite načine kako bi se obuhvatili široki aspekti kompozicijske semantike u višerječnim izrazima.

Prva dva skupa podataka sastavljena su od parova izraza oblika pridjev–imenica–glagol–pridjev–imenica (*anvan1* i *anvan2*), čija je konstrukcija opisana u radovima (Kartsaklis et al., 2013) i (Grefenstette, 2013). Skupovi podataka sastavljeni su od višerječnih izraza, pri čemu je u svakom paru odgovarajućih izraza izmijenjen glagol. Oba skupa izraza sastoje se od po 200 parova izraza. Glagoli u izrazima odabrani su iz skupa najčešćih višeznačnih glagola u engleskome jeziku, dok su ostale riječi u izrazu odabrane prema učestalosti pojavljivanja uz odabrani glagol. Primjerice, imenice koje označavaju subjekte odabrane su iz skupa najčešćih subjekata određenog glagola, dok su pridjevi koji se odnose na subjekt odabrani iz skupa najčešćih pridjeva određene imenice. Isti postupak primijenjen je i na objekte glagola, odnosno njihove pridjeve.

Parovi su sastavljeni tako da su glagoli u kojima se izrazi razlikuju slični ili pak potpuno nepovezani. Za svaki od parova izraza, označivači su označili u kojoj mjeri su za njih ta dva izraza semantički slična, i to na ljestvici od 1 do 7, gdje oznaka sličnosti 1 označava potpuno nepovezano semantičko značenje dvaju fraza, dok oznaka 7 označava potpuno isto semantičko značenje fraza. Primjerice jedan par izraza za označavanje su izrazi: "*young woman filed long nails*" (*mlada žena je oblikovala duge nokte*) i "*young woman smoothed long nails*"' (*mlada žena je uglađivala duge nokte*), pri čemu je dani par izrazito semantički sličan, dok bi primjer izrazito semantički nepovezanog izraza prvom izrazu bio: "*young woman registered long nails*" (*mlada žena je registrirala duge nokte*). Glagoli *file* i *register* nalaze se u tim izrazima budući da ipak imaju slično značenje, ali ne u ovom kontekstu (primjer sličnog značenja je u slučaju predavanja ili podnošenja nekog dokumenta, pri čemu se mogu koristiti oba glagola). Vrednovanje modela obavljeno je računanjem Spearmanovog koeficijenta korelacije između oznaka označivača i kosinusne sličnosti vektora izraza dobivenih određenim modelom.

Treći skup podataka za vrednovanje (*tfds*) sastoji se od rečenica i njihovih parafraza. Za svaku rečenicu dan je određeni broj parafraza koje imaju slično značenje kao i početna rečenica, ali se poredak riječi ili neke od riječi u rečenici razlikuju. Osim parafraza s istim semantičkim značenjem, dan je i određen broj rečenica s istim skupom korištenih riječi, ali različitim poretkom kojim je i semantičko značenje promijenjeno. Primjerice, za originalnu frazu "*A man plays an acoustic guitar*" (*Mu-*

Model	anvan1	anvan2	tfds	msrwid	onwn
Aditivni model	8	22	-0.2	78	66
Multiplikativni model	8	-4	-2.3	77	55
Model leksičke funkcije	15	30	**5.9**	-	-
PLF-model	**20**	**36**	2.7	**79**	**67**
State-Of-Art	22	27	11.4	87	75

Tablica 4.2: Rezultati vrednovanja različitih modela u radu (Paperno et al., 2014)

škarac svira akustičnu gitaru), parafraza sa sličnim značenjem bila bi: "*A man plays guitar*" (*Čovjek svira gitaru*), dok bi primjer parafraze s istim skupom riječi ali različitim značenjem bila fraza "*A guitar plays a man*" (*Gitara svira muškarca*). Ukupan broj rečenica u navedenom skupu podataka iznosi 157, od kojih je svaka povezana s prosječno 8 parafraza s istim semantičkim značenjem i prosječno 17 parafraza s promijenjenim semantičkim značenjem. Vrednovanje ovog skupa podataka provedeno je t-standardiziranim testom prosječne razlike kosinusne sličnosti originalne fraze sa semantički sličnim parafrazama i kosinusne sličnosti originalne fraze sa semantički nepovezanim parafrazama. Očekivani rezultat uspješnog modela je da su razlike između tih sličnosti što veće.

Preostala dva skupa podataka (*msrvid* i *onwn*) preuzeta su iz javno dostupnih izvora, te su također kao i prva dva skupa ručno označena od strane označivača u zadataku semantičke sličnosti između dviju fraza. Rečenice u ovim skupovima podataka nisu sastavljene prema sintaktičkom pravilu kao u slučaju *anvan* skupova, već su to rečenice slobodne forme. Prvi skup podataka sastoji se od 750 parova rečenica pisanih slobodnim stilom, pri čemu je za svaki par petoro označivača ručno označilo sličnost između rečenica. Drugi skup podataka sastoji se od 561 para izraza, također pisanih slobodnim stilom, te označenih od strane 5 označivača u zadatku sličnosti ponuđenih izraza. Vrednovanje ovih skupova rečenica obavljeno je računanjem Pearsonovog faktora korelacije između sličnosti dobivene modelom i sličnosti koju su za pojedine parove izraza naznačili označivači.

Tablica 4.2 sadrži rezultate vrednovanja PLF-modela na predstavljenim skupovima podataka, uz usporedbu rezultata sa rezultatima jednostavnog aditivnog (*add*), multiplikativnog (*mult*) i modela leksičke funkcije (*lf*). Također, prikazani su i u tom trenutku najbolji rezultati (engl. *state-of-art*) za pojedine skupove podataka. Iz rezultata je vidljivo da je PLF-model postigao značajno bolje rezultate kod 4 od 5 navedenih

skupova podataka. Uz to, na skupu *anvan2* PLF-model postigao je i bolje rezultate od dotad najboljih za taj skup – dobiven je Spearmanov faktor korelacije iznosa 36, dok je prethodno najbolji postignuti iznosio 27. Osobito je primjetna razlika u rezultatima PLF-modela u odnosu na rezultate modela add i mult u slučaju skupova *anvan1* i *anvan2*. Znatno boljim rezultatima, a uz neznatno duže vrijeme izvođenja, PLF-model ističe se kao izrazito prikladna alternativa za jednostavne aditivne i multiplikativne modele.

5. Nadogradnja PLF-modela

Predstavljeni model praktične leksičke funkcije postigao je znatno bolje rezultate u odnosu na jednostavne aditivne i multiplikativne modele, ali i kompleksniji model leksičke funkcije. Uspjeh je još i veći uzimajući u obzir jednostavnost modela.

Ipak, i ovako definiran model sadrži određene nedostatke koje su u svom radu prepoznali i pokušali riješiti Gupta et al. (2015). Njihova zamjerka modelu odnosi se na sam postupak treniranja matrica modela, koji nije konzistentan s načinom korištenja modela pri izračunu vektora višerječnih fraza. Nekonzistentnost modela pokušali su riješiti načinima opisanim u poglavlju 5.1.

U okviru diplomskog rada razmotreni su i načini poboljšanja modela s obzirom na maksimizaciju sličnosti između parova sinonima te relaciju semantičke inkluzije između parova hiponima i hiperonima riječi. Rezultati nadogradnje skupa za treniranje PLF-modela s obzirom na maksimizaciju sličnosti između sinonima predstavljeni su u poglavlju 5.2, dok je nadogradnja modela s obzirom na relaciju semantičke inkluzije između parova hiponima i hiperonima opisana u poglavlju 5.3.

5.1. Prilagodbe PLF-modela

Gupta et al. (2015) u svom su radu prepoznali nedostatak PLF-modela koji se očituje u nekonzistentnosti u postupku treniranja matrica u odnosu na postupak izgradnje konačnog vektora koristeći dobivene matrice.

Naime, u postupku treniranja matrica značenja za pridjeve, odnosno glagole, ulazni vektori modela su vektori značenja pojedinih imenica, a izlazni vektori su vektori oblika pridjev–imenica, odnosno glagol–objekt ili subjekt–glagol. Matrica dobivena takvim postupkom treniranja izravno modelira značenje pojedinog pridjeva ili glagola, budući da se množenjem primjerice vektora pojedine imenice matricom pojedinog pridjeva, dobiva vektor pridjev–imenica, odnosno vektor imenica na koji je svojim značenjem djelovao pridjev. S druge strane, u postupku izračuna konačnog vektora pridjev–imenica u ovom slučaju, originalni PLF-model predlaže da se tako dobive-

nom vektoru umnoška matrice pridjeva i vektora imenice, pribroji i vektor značenja pridjeva. Takvim postupkom povećava se utjecaj značenja pridjeva na konačni vektor značenja izraza, budući da je semantičko značenje pridjeva prisutno i u matrici koja opisuje njegovo djelovanje na imenicu, te u vektoru značenja pridjeva koji se na kraju pridodaje dobivenom umnošku. Jednake zamjerke odnose se i na preostale slučajeve kompozicije: subjekt–glagol i glagol–objekt, u kojima je također prisutan dvostruki utjecaj značenja jedne riječi na konačni vektor značenja izraza, u ovom slučaju to je glagol.

Kao moguće rješenje ove nekonzistentnosti, Gupta et al. (2015) predlažu dvije prilagodbe modela. Prva prilagodba odnosi se na fazu treniranja modela, odnosno izmijenjen način treniranja matrica koje opisuje djelovanje pojedinog pridjeva ili glagola. Primjerice, u slučaju treniranja matrice subjekta glagola, njihov prijedlog svodi se na treniranje matrice koja umjesto vrijednosti vektora para subjekt–glagol, predviđa razliku između vrijednosti vektora subjekt–glagol i vektora glagola. Takav pristup može se definirati na sljedeći način:

$$\mathbf{V}_{subj} = \arg\min_{\mathbf{M}} \sum_{n \in subj(v)} \|\mathbf{M} \times \overrightarrow{n} - (\overrightarrow{nv} - \overrightarrow{v})\|^2$$

gdje su $\overrightarrow{n}$ i $\overrightarrow{v}$ vektori imenice, odnosno glagola, $\overrightarrow{nv}$ vektor para subjekt–glagol, te $\mathbf{M}$ matrica koja opisuje djelovanje određenog glagola na subjekt. Množenjem tako dobivene matrice i vektora imenice te zbrajanjem dobivenog vektora i vektora glagola, smanjuje se naknadni utjecaj vektora glagola u konačnom vektoru, budući da je isti izuzet iz postupka treniranja matrice za glagol, oduzimanjem njegove vrijednosti od vektora para subjekt–glagol. Analogni postupci primijenjuju se i na treniranje matrica za parove pridjev–imenica, te glagol–objekt.

Druga prilagodba modela odnosi se na fazu primjene, odnosno ispitivanja modela. Problem je i u ovom slučaju dvostruki utjecaj vektora pridjeva ili glagola na konačni vektor značenja izraza. Međutim, u ovom pristupu, postupak treniranja matrica ostaje isti kao u originalnom PLF-modelu, a mijenja se postupak izgradnje konačnog vektora višerječnog izraza, u kojem se vektoru dobivenom umnoškom matrice pridjeva ili glagola i vektora imenice ne dodaje vektor pridjeva, odnosno glagola. Na taj način samo izgrađena matrica ima utjecaj na konačni vektor, što je u skladu s postupkom treniranja matrica, budući da se postupkom regresije izravno predviđaju vrijednosti vektora parova pridjev–imenica, subjekt–glagol ili glagol–objekt.

Pri vrednovanju predloženih prilagodbi modela, koristili su iste skupove podataka kao i pri vrednovanju originalnog PLF-modela (*anvan1* i *anvan2*), kako bi rezultati bili

međusobno usporedivi. Rezultati vrednovanja prikazani su u tablici 5.1. Vrednovanje je pokazalo da se korištenjem prilagodbe u fazi primjene modela postižu bolji rezultati nego u načinu predloženom u originalnom PLF-modelu. S druge strane, prilagodbom u fazi treniranja postignuti su dosta lošiji rezultati u odnosu na referentne.

Loši rezultati prilagodbe u fazi treniranja objašnjeni su činjenicom da su sami vektori koje matrica pokušava predvidjeti znatno rijeđi od vektora imenica koje se nalaze na ulazu modela. Drugim riječima, matrica pokušava predvidjeti vektore koji su lošije definirani od samih ulaznih vektora. Samim oduzimanjem vektora pridjeva ili glagola od vektora izraza na koji oni djeluju, treniranje matrice svodi se na predviđanje vrijednosti vektora pridjeva ili glagola iz vrijednosti vektora imenice ($\overrightarrow{nv} - \overrightarrow{v} \approx -\overrightarrow{v}$, u slučaju matrica glagola). Takvo predviđanje vrijednosti vektora glagola samo po sebi je težak problem, dok s druge strane uopće ne riješava problem prikaza značenja izraza.

Bolji rezultati u odnosu na referentne, postignuti prilagodbom faze primjene modela, mogu se objasniti standardnom ravnotežom između visoke generalizacije i visoke varijance modela (engl. *bias-variance tradeoff*). U slučaju originalno predloženog načina primjene PLF-modela, prisutna je o visoka generalizacija u modelu, koju uvodi vektor pridjeva ili glagola u izrazu, što je posebno korisno u slučajevima kada matrice pojedinih pridjeva ili glagola nisu dovoljno dobro definirane, pa je takvim pristupom značenje izraza definirano pretežno značenjem pridjeva ili glagola u izrazu. S druge strane, u slučaju predložene prilagodbe faze primjene modela, riječ je o visokoj varijanci u modelu, budući da se pri izračunu konačnog vektora izraza, model oslanja samo na umnožak trenirane matrice i vektora imenice iz izraza, neovisno o tome koliko je dobro matrica definirana. Zapravo, u slučaju prilagodbe faze primjene, model se može smatrati "prenaučenim" na primjere iz korpusa.

Predloženi načini prilagodbe modela ukazali su na nedostatke u modelu, koji se poglavito odnose na sam postupak treniranja matrica, ali i na činjenicu da je taj postupak znatno ovisan o frekvencijama primjera za treniranje, budući da će samo u slučaju visokofrekventnih primjera matrice biti dobro definirane i sadržavati potpuno semantičko značenje pojedinih pridjeva, odnosno glagola. Na tragu problema ravnoteže između visoke generalizacije i visoke varijance modela, Gupta et al. (2015) predložili su u okviru daljnjeg rada na modelu eksperimentiranje s težinama pojednih matrica, kao potencijalan način kontroliranja generalizacije, odnosno varijance modela.

Model	anvan1	anvan2
PLF-model	20.6	35.2
PLF-model prilagodba treniranja	3.8	17
PLF-model prilagodba ispitivanja	22.1	35.4

Tablica 5.1: Rezultati vrednovanja prilagodbi PLF-modela u radu (Gupta et al., 2015)

5.2. Nadogradnja temeljena na maksimizaciji sličnosti između sinonima

U okviru diplomskog rada predloženi su i vrednovani načini nadogradnje skupa za treniranje modela temeljeni na maksimizaciji semantičke sličnosti između izraza koji sadrže sinonime. Ovakva nadogradnja motivirana je činjenicom da bi par višerječnih izraza istog sintaksnog oblika sastavljen od međusobnih sinonima trebao imati slično semantičko značenje. Primjerice, izrazi poput "*vrijedan poklon*" i "*vrijedan dar*" sastavljeni su od istog pridjeva, ali različitih imenica, koje su pak sinonimi. Unatoč razlici u imenicama, semantičko značenje navedenih izraza je slično, upravo zbog činjenice da su imenice u izrazima sinonimi. Zamjenom i pridjeva u izrazu odgovarajućim sinonimom, primjerice u izrazima "*skupocjen poklon*" ili "*skupocjen dar*", jasno je da je semantičko značenje i tih izraza slično značenju početnog izraza "*vrijedan poklon*", pa je samim time opravdan i prijedlog nadogradnje skupa za treniranje modela s obzirom na sličnost između izraza koji sadrže međusobne sinonime.

Predložena nadogradnja može se ostvariti jednostavnim postupkom izmjene ulaznih i izlaznih primjera u postupku treniranja matrica PLF-modela, točnije proširenjem skupa ulaznih i izlaznih primjera vektorima izraza koji sadržavaju sinonime postojećih primjera za treniranje. Naime, kako bi se maksimizirala sličnost matrica riječi koje predstavljaju međusobne sinonime, u postupku treniranja matrica za pojedine pridjeve i glagole kao izlazni vektori modela uključuju se i vektori izraza koji umjesto trenutnog pridjeva ili glagola sadrže njihove sinonime. Primjerice, u postupku treniranja matrice za pridjev "*vrijedan*" uključuju se svi izrazi oblika pridjev–imenica u kojima je kao pridjev korišten pridjev "*vrijedan*", ali i svi izrazi u kojima je umjesto pridjeva "*vrijedan*" korišten pridjev sinonim pridjeva "*vrijedan*", primjerice "*dragocjen*", "*skup*" i drugi. Kako bi se ipak sačuvao kontekst u kojem se određeni pridjev ili glagol koristi, ne dodaju se svi mogući izrazi koji sadrže sinonime trenutnog pridjeva ili glagola, već samo oni izrazi koji sadržavaju iste imenice ili imenice sinonime onih imenica koje su

korištene u kombinaciji s originalnim pridjevima ili glagolima. Primjerice, u slučaju izraza "*vrijedan poklon*", skupu za treniranje dodaju se izrazi "*skupocjen dar*" i "*dragocjen dar*", ali ne i izraz "*dragocjen izvor*", budući da "*izvor*" nije sinonim imenice "*poklon*" (što ne znači da se taj izraz na kraju neće dodati u konačni skup za treniranje, jer je moguće da je u početnom skupu postojao i izraz sa sinonimom imenice "*izvor*").

Za pronalazak sinonima pridjeva, glagola i imenica korišten je WordNet.[1] WordNet je velika leksička baza podataka engleskoga jezika. Sadrži podatke o značenjima riječi grupirane u skupove sinonima riječi prema različitim kontekstima u kojima se one koriste. Skupovi značenja sadrže sinonime različitih riječi, te je time definiran graf značenja riječi koji se sastoji od semantičkih veza među njima: veza između hiperonima i hiponima, antonima, sinonima i drugih odnosa. Skupovi su poredani prema učestalosti korištenja pojedine riječi u određenom kontekstu, pa tako prvi skup sinonima predstavlja njenu najčešću uporabu, dok zadnji skup predstavlju najrjeđu uporabu riječi. Pretraživanjem baze podataka moguće je na lak način pristupiti sinonimima pojedinih pridjeva, imenica i glagola te pomoću njih odrediti izraze koji sadrže sinonime riječi sadržanih u određenim izrazima.

U postupku maksimizacije sličnosti između izraza odabrana su dva temeljna pristupa. U prvom pristupu kao izrazi sinonimi (izrazi s riječima koje predstavljaju sinonime riječi iz početnog izraza, izabrane na prethodno opisan način) korišteni su izrazi sastavljeni od sinonima prisutnih isključivo u prvom skupu značenja riječi prema WordNetu. Time je skup mogućih izraza sinonima ograničen samo na one sinonime koji predstavljaju najčešće korišteno semantičko značenje riječi. U drugom pristupu kao izrazi sinonimi odabrani su oni izrazi koji su sastavljeni od sinonima riječi koji imaju visoku međusobnu vrijednost semantičke sličnosti Leacocka i Chodorowa (engl. *Leacock and Chodorow similarity – LCH*). LCH-sličnost u prvom koraku određuje broj veza između dva skupa značenja, koje označavaju podvrstu ("*is-a*" veze) u WordNet-ovom grafu povezanosti značenja. Taj broj se zatim skalira s obzirom na najduži postojeći niz takvih veza za ta dva skupa značenja u WordNet-ovom grafu te se konačna mjera sličnosti definira kao negativan logaritam skalirane vrijednosti. Kao početna granična vrijednost za prihvaćanje izraza kao semantički sličnog početnom izrazu izabrana je LCH sličnost u iznosu od 2.5. Maksimalna LCH sličnost za neke pojmove u WordNet-ovoj mreži skupova značenja iznosi 3.583 i to za međusobnu sličnost između istog skupa značenja (Warin i Volk, 2004), pa je samim odabirom vrijednosti manje od te ostvarena mogućnost odabira skupova značenja s dovoljnom razinom sličnosti u od-

[1] WordNet je leksička baza engleskoga jezika, sadrži skupove sinonima za pojedinu riječ: `http://wordnet.princeton.edu`

nosu na početni skup. Izabrana vrijednost može se smatrati hiperparametrom modela, koji bi se mogao optimirati na odvojenom skupu za ispitivanje. Riječi koje imaju toliku ili veću mjeru sličnosti sa trenutno obrađivanom riječi, dodane su u skup za treniranje slijedeći prethodno definirana pravila.

Na ovako definirane osnovne načine dodavanja sinonima skupu za treniranje, primijenjeni su i još neki kriteriji odabira, kako bi se dobili potencijalno bolji rezultati na ispitnim skupovima. Naime, dodavanjem svih izraza sinonima na načine koji su prethodno opisani unosi se veliki šum u skup za treniranje, prvenstveno zbog mogućeg dodavanja u skup za treniranje onih izraza koji nisu dovoljno frekventni u pojavljivanju u korpusu, čime je zapravo njihov vektorski prikaz nedovoljno definiran. Treniranjem modela s takvim vektorima postoji velika mogućnost da će se model previše prilagoditi upravo tim vektorima, premda ni sami vektori tih izraza ne opisuju dovoljno dobro semantičko značenje istih. Time se smanjuje uspješnost modela u modeliranju značenja osnovnog pridjeva ili glagola za koji se matrica trenira, što svakako nije nešto što bi se htjelo postići ovom nadogradnjom. S tim u vidu, na prethodno definirane načine primijenjen je i kriterij odabira samo onih izraza koji se u korpusu pojavljuju barem 100 puta ili više. Također, budući da su s načinom dodavanja izraza samo iz prvog skupa značenja preliminarno postignuti bolji rezultati negoli koristeći LCH-sličnost, na taj način dodavanja primijenjen je i kriterij dodavanja samo onih izraza koji se u korpusu pojavljuju približno frekventno kao i njima odgovarajući izrazi sinonimi. To znači da su za pojedini izraz u skupu za treniranje dodani izrazi sinonimi samo u slučaju da se ti izrazi u korpusu pojavljuju s približno istom frekvencijom (dozvoljena je razlika od 10%) kao i originalni izrazi, ali bez kriterija o frekvenciji pojavljivanja većoj od 100. Tim postupkom osigurano je da su vektori novih parova u skupu za treniranje podjednako dobro definirani kao i njima odgovarajući originalni parovi. U tablici 5.2 dan je pregled broja dodanih vektora u skup za treniranje za svaki od predloženih načina proširenja skupa, zajedno s pregledom broja ulaznih vektora bez predloženih nadogradnji.

Kako bi se usporedili rezultati predloženih nadogradnji skupa za treniranje s originalnim rezultatima PLF-modela na ispitnim skupovima, izgrađen je PLF-model za engleski jezik prema uzoru na onog opisanog u radu (Paperno et al., 2014). Za izgradnju modela korišten je reducirani korpus sastavljen od korpusa ukWaC i engleske Wikipedije (uz njih nije korišten i British National Corpus kao u navedenom radu). Iz združenog korpusa ekstrahirani su vektori supojavljivanja za 30.000 najčešćih riječi (imenica, pridjeva i glagola) u tako definiranom korpusu, kao i vektori supojavljivanja izraza pridjev–imenica, subjekt–glagol i glagol–objekt, za sve pridjeve i glagole koji

	PLF	PLF fst	PLF lch	PLF fst freq	PLF lch freq	PLF fst 10%
AN vektori	338,508	46,893	27,119	1,402	897	2,494
SV vektori	150,386	46,177	48,540	2,860	2,861	3,149
VO vektori	77,771	37,675	39,398	3,274	3,282	2,295

Tablica 5.2: Ukupan broj dodanih vektora izraza u skup za treniranje matrica PLF-modela (prema kategorijama AN – pridjev–imenica, SV – subjekt–glagol i VO – glagol–objekt) za svaku od predloženih , zajedno s brojem vektora korištenih u reprodukciji originalnih rezultata PLF-modela.

se nalaze u ispitnom skupu izraza. Nad ekstrahiranim vektorima primijenjeni su postupci izračuna pozitivne zajedničke informacije, smanjenja dimenzionalnosti vektora (na dimenziju 300), te normalizacije vektora na duljinu 1, prema uzoru na izgradnju originalnog PLF-modela. Matrice pojedinih pridjeva i glagola trenirane su također prema uzoru na originalan PLF-model – koristeći L2-regulariziranu regresiju uz generaliziranu unakrsnu provjeru. Osim reprodukcije PLF-modela, reproducirane su i obje prilagodbe modela predložene u radu (Gupta et al., 2015), kao i jednostavan aditivni i multiplikativni model korišteni u radu (Paperno et al., 2014) za usporedbu s PLF-modelom. Model je izgrađen u programskom jeziku Python, pri čemu je treniranje matrica modela obavljeno korištenjem javno dostupne biblioteke scikit–learn.[2] Kao ispitni skupovi izraza korišteni su javno dostupni skupovi *anvan1* i *anvan2*. Rezultati vrednovanja reproduciranog PLF-modela i njegovih prilagodbi na navedenim ispitnim skupovima prikazani su u tablici 5.3, zajedno s referentnim rezultatima iz originalnih radova.

Reproducirani rezultati u znatnoj mjeri odgovaraju rezultatima iz referentnih radova, ali ipak postoje neke razlike među njima. Prvenstveno se to odnosi na reprodukciju PLF-modela s izmijenjenom fazom ispitivanja, koja je u slučaju *anvan1* skupa izraza postigla lošiji rezultat (19.24) u odnosu na originalni PLF-model (19.37). U referentnim radovima ta prilagodba modela postiže bolje rezultate (22.1) u odnosu na originalni PLF-model (20). Mogući uzrok takvih rezultata izostanak je jednog dijela korpusa pri reprodukciji rezultata (British National Corpus), iako veličina tog korpusa nije velika, pa je samim time umanjen njegov značaj. S druge strane, moguće je da su i vektori supojavljivanja ekstrahirani na drugačije načine od onih korištenih u radovima, pa su time izostavljeni određeni konteksti izraza iz korpusa koji bi bolje definirali sami

[2]Biblioteka scikit-learn sadrži često korištene algoritme strojnog učenja implementirane u programskom jeziku Python: `http://scikit-learn.org/`

Model	anvan1	anvan2
Aditivan	5.67 (8)	23.92 (22)
Multiplikativan	6.49 (8)	-4.22 (-4)
PLF-model	**19.37** (20)	**39.99** (**36**)
PLF-model prilagodba treniranja	3.17 (3.8)	19.56 (17)
PLF-model prilagodba ispitivanja	19.24 (**22.1**)	39.43 (35.4)

Tablica 5.3: Rezultati vrednovanja reproduciranih modela u obliku Spearmanovog koeficijenta korelacije između sličnosti dobivene modelom i oznaka sličnosti označviača za pojedini izraz. Rezultati su dani u odnosu na rezultate dobivene u referentnim radovima (vrijednosti u zagradama) nad istim skupovima podataka

vektor značenja izraza.

Rezultati ispitivanja predloženih nadogradnji skupa za treniranje modela na skupovima *anvan1* i *anvan2* korištenima pri prvotnom vrednovanju PLF-modela, dani su u tablici 5.4. Rezultati ukazuju na to da predložene načini nadogradnje skupa za treniranje ipak nisu dovoljno dobro definirani, sudeći po rezultatima ispitivanja nadogradnji na navedenim skupovima. Niti jedna od navedenih nadogradnji nije postigla bolje rezultate od osnovnog PLF-modela na *anvan* skupovima podataka (vrijednosti za PLF-model su dobivene prethodno opisanom reprodukcijom postupka izgradnje modela iz originalnog rada). Nadogradnja temeljena na dodavanju izraza sinonima samo iz prvog skupa značenja riječi (PLF fst) postigla je bolje rezultate od nadogradnje temeljene na LCH sličnosti skupova sinonima (PLF lch). Štoviše, dodatan kriterij dodavanja samo dovoljno frekventnih izraza sinonima u skup za treniranje (PLF fst freq, PLF lch freq) pokazao se kao opravdan izbor, budući da su u oba slučaja nadogradnji bolji rezultati ostvareni s uključivanjem tog kriterija. Takvo ponašanje modela može se objasniti upravo prethodno navedenim razlogom uključivanja tog kriterija u postupke nadogradnje – nedovoljno dobro definirani vektori uzrokovati će loše definirane težine matrica modela, budući da će se model truditi obuhvatiti i takva loše definirana značenja pojedinih izraza.

Još jedan razlog lošijih rezultata nadogradnji skupa za treniranje je činjenica da su u skup za treniranje dodani vektori imenica koje se već nalaze u tom skupu, čime model zapravo pokušava naučiti preslikavanje istog vektora imenice u različite vektore izraza u kojima je ta imenica sadržana. Primjerice, pri treniranju matrice značenja za pridjev "*vrijedan*", kao izlazni vektori koriste se svi vektori oblika pridjev–imenica u kojima je pridjev upravo pridjev "*vrijedan*", dok su ulazni vektori modela vektori imenica koje se

Model	anvan1	anvan2
PLF	19.37	39.99
PLF fst	18.11	35.82
PLF lch	10.98	34.47
PLF fst freq	18.38	34.74
PLF lch freq	11	33.32
PLF fst 10%	18.32	34.05

Tablica 5.4: Rezultati vrednovanja nadogradnji skupa za treniranje PLF-modela na temelju maksimizacije sličnosti sinonima

u nalaze u tim izrazima – "*dar*", "*poklon*", "*poklon*" i dr. U slučaju nadogradnje skupa za treniranje modela izrazima koji sadržavaju sinonime riječi u izrazu, skup izlaznih vektora nadopunjuje se vektorima izrazima koji sadržavaju sinonime, kao što su, u slučaju pridjeva "*vrijedan*", izrazi "*dragocjen poklon*", "*dragocjen dar*" ili "*skupocjen poklon*". Budući da svakom izlaznom vektoru odgovara jedan ulazni vektor, potrebno je sa svakim dodavanjem vektora izraza pridjev–imenica u izlazni skup vektora dodati i njemu odgovarajući vektor imenice u skup ulaznih vektora. U slučaju dodavanja prethodno navedenih izraza, to bi značilo dodavanje vektora imenica "*poklon*", "*dar*" i "*poklon*" u ulazni skup vektora. Iz navedenih postupaka vidljivo je da će se u skup ulaznih vektora dodati duplikati vektora koji se u tom skupu već nalaze – vektor imenice "*poklon*" pojaviti će se u tom skupu barem 2 puta, prema navedenom postupku. Samim time, model ne može ispravno modelirati niti jedan vektor izraza koji sadrže imenicu čiji se vektor u ulaznom skupu vektora pojavljuje više puta, budući da se minimizacijom kvadratne pogreške pokušava ispravno modelirati oba vektora izraza, što je uz ovakav postupak nemoguće ako oni nisu isti. S druge strane, završni kriterij odabira izraza sinonima, temeljen na dodavanju samo približno jednako frekventnih izraza u skup za treniranje (PLF fst 10%), pokazao se kao manje uspješniji kriterij od ostalih, sudeći po dobivenim rezultatima. Očito je iz toga da samo frekvencija ne bi trebala biti isključivi kriterij za dodavanje vektora u skup za treniranje, već bi se mogle iskoristiti i druge informacije o vektorima ekstrahiranim iz korpusa prije negoli se dodaju u skup za treniranje – primjerice, kao kriterij bi se mogla uzeti i kosinusna sličnost vektora izraza koji sadrže sinonime, pri čemu bi se u skup izraza za treniranje dodali samo oni vektori čija je međusobna kosinusna sličnost dovoljno velika.

Predloženi načini nadogradnje skupa za treniranje pokazali su da se problemu mora

pristupiti opreznije i s detaljnije razrađenim pristupom dodavanja pojedinih izraza sinonima u skup za treniranje. Očito je da dodavanje izraza u skup za treniranje mora biti opravdano kvalitetom samih vektora koji se dodaju, odnosno stupnjem definiranosti tih vektora, dok se istodobno mora paziti da se sam postupak treniranja modela ne otežava dodavanjem znatno različitih izlaznih vektora za iste ulazne vektore.

5.3. Nadogradnja temeljena na relaciji semantičke inkluzije

Nadogradnja PLF-modela temeljena na relacije semantičke inkluzije između hiperonima i hiponima riječi motivirana je sličnim razlozima kao i nadogradnja temeljena na maksimizaciji sličnosti između sinonima. U postupku maksimizacije sličnosti sinonima opravdano je bilo dodavati izraze sinonime u skup za treniranje matrica pojedinih pridjeva i glagola, kako bi matrice značenja sinonima bile što sličnije. Međutim u slučaju relacije semantičke inkluzije, maksimizaciju sličnosti značenja hiperonima i hiponima nije moguće ostvariti tako jednostavnim postupkom zbog činjenica da ni semantičko značenje hiperonima i hiponima nije u potpunosti jednako, premda su donekle slična (primjerice, semantičko značenje hiperonima "*građevina*" u odnosu na značenju njemu odgovarajućeg hiponima "*kuća*"). Jednostavno dodavanje parova izraza hiperonima u skup za treniranje matrice odgovarajućeg hiponima dovelo bi do maksimizacije sličnosti značenja pojedinih hiperonima i hiponima, ali bi se time izgubila prirodno opravdana razlika u njihovim značenjima, zbog koje dvije riječi upravo i jesu hiperonim i njemu odgovarajući hiponim, a ne par sinonima. Zbog toga je potrebno definirati drugačiji način maksimizacije sličnosti između takvih parova riječi, kojim bi se uspješno maksimizirala njihova sličnost, dok bi se istodobno i očuvala razlika u istoj.

5.3.1. Mjere razine semantičke inkluzije među vektorima

Semantička inkluzija u semantičkom vektorskom prostoru motivirana je distribucijskom hipotezom inkluzije (engl. *distributional inclusion hypothesis*) (Lenci i Benotto, 2012), koja kaže da je velik broj istaknutih vrijednosti (prema apsolutnoj vrijednosti) vektora izraza semantički užeg značenja (hiponima) u istoj mjeri istaknut i u vektorima značenja njima odgovarajućih izraza semantički šireg značenja (hiperonimima). Samim time, dva vektora koja su povezana relacijom semantičke inkluzije, dakle par

vektora sastavljen od vektora hiponima i vektora hiperonima, imati će u pojedinim dimenzijama svojih vektora jednake vrijednosti, čime je određeno njihovo slično semantičko značenje. Što je veći broj dimenzija u kojima vektori imaju jednake vrijednosti, veća je i razina semantičke inkluzije među njima.

Ovako definiran odnos između hiponima i hiperonima sličan je odnosu između vektora sinonima izraza. Međutim, razlika je u tome što se u vektorima sinonima ne mogu jasno odrediti dimenzije vektora koje predstavljaju istaknute vrijednosti obaju vektora koje definiraju njihovo uže, odnosno šire semantičko značenje. Vektori sinonima međusobno su slični, ali ne postoji pravilo kojim bi se definirao skup dimenzija koji bi jednog od vektora definirao kao onog s užim, a drugog kao onog sa širim semantičkim značenjem – njihova semantička značenja naprosto su slična (ako ne i ista).

Vodeći se prvenstveno načelom broja dimenzija s jednakim vrijednostima vektora, definirane su standardne mjere razine semantičke inkluzije među vektorima riječi (Lenci i Benotto, 2012). U radu je predstavljen pregled u literaturi korištenih mjera za određivanje razine inkluzije među riječima, te su iste vrednovane na skupu podataka sastavljenom od parova hiperonima i hiponima, ali i parova sinonima te parova nasumično izabranih riječi. Cilj ovakvog vrednovanja mjera bio je otkriti u kolikoj mjeri su mjere za inkluziju sposobne razlikovati stvarne parove hiperonima i hiponima u odnosu na sinonime i nasumično izabrane riječi. Kao najuspješnije mjere za inkluziju pokazale su se mjere $ClarkeDE$ i $invCL$.

$ClarkeDE$ mjera za inkluziju (Lenci i Benotto, 2012) definirana je kao varijacija u radu prethodno korištene mjere $WeedsPrec$. Mjera $WeedsPrec$ određuje težinsku razinu inkluzije vrijednosti vektora u u težinskim vrijednostima vektora v:

$$WeedsPrec(u, v) = \frac{\sum_{f \in F_u \cap F_v} w_u(f)}{\sum_{f \in F_u} w_u(f)}$$

gdje je F_x skup težinskih vrijednosti vektora x, a $w_x(f)$ težinska vrijednost dimenzije f vektora x. Budući da su vektori korišteni pri izračunu vrijednosti ove mjere u navedenom radu distribucijski vektori dobiveni iz distribucijske memorije za engleski jezik, pri čemu su dimenzije pojedinog vektora zapravo stupanj zajedničkog korištenja pojedine riječi uz riječ trenutnog vektora, nisu svi sastavljeni na isti način, odnosno koristeći isti skup riječi za izgradnju vektora značenja pojedine riječi.[3] Samim time potrebno je izdvojiti one dimenzije koje su sumjerljive u postupku izračuna razine inkluzije među vektorima, što je i učinjeno promatranjem samo dimenzija koje su zajedničke pojedinim vektorima u navednom izrazu ($f \in F_u \cap F_v$). Varijacija mjere

[3]Distribucijska memorija je prikaz distribucije riječi u korpusu u obliku trojki (riječ, težina, riječ) kojima je opisana učestalost zajedničke uporabe dviju riječi u korpusu (Baroni i Lenci, 2010).

$WeedsPrec$, nazvana $ClarkeDE$ mjerom definirana je kao:

$$ClarkeDE(u,v) = \frac{\sum_{f \in F_u \cap F_v} min(w_u(f), w_v(f))}{\sum_{f \in F_u} w_u(f)}$$

Razlika u odnosu na $WeedsPrec$ mjeru očituje se u odabiru minimalne vrijednosti između zajedničkih dimenzija vektora u i v, što se pri vrednovanju mjera pokazalo kao efikasniji način identifikacije parova hiperonima i hiponima riječi.

Nova mjera predložena u radu (Lenci i Benotto, 2012) nazvana je invCL mjerom, temelji se na $ClarkeDE$ mjeri, ali ima malo drugačiji pristup u odnosu na nju. Naime, $invCL$ mjera ne uzima u obzir samo razinu inkluzije pojedinog hiponima u njemu odgovarajućem hiperonimu, već pokušava odrediti i obrnuti odnos – u kojoj je mjeri hiperonim sadržan u njemu odgovarajućem hiponimu:

$$invCL(u,v) = \sqrt{ClarkeDE(u,v) \cdot (1 - ClarkeDE(v,u)}$$

Motivacija za ovakav pristup leži u činjenici da će vektor hiperonima imati znatno bogatije definiran vektor u odnosu na njemu odgovarajući hiponim te će samim time razina inkluzije hiponima u takvom vektoru hiperonima biti veća, dok će hiponim imati rijeđe definiran vektor te njemu odgovarajući hiperonim s bogato definiranim vektorom neće imati visoku razinu inkluzije u hiponimu, prema mjeri $ClarkeDE$.

Bitno je naglasiti da su navedene mjere za inkluziju asimetrične, odnosno da se rezultati dobiveni navedenim mjerama odnose na mjeru inkluzije vektora u u vektoru v, a ne obrnuto.

5.3.2. Analiza prisutnosti inkluzije u PLF-modelu

Navedene mjere ispitane su na spomenutom skupu sastavljenom od parova hiperonima i hiponima, ali i parova sinonima riječi, parova meronima riječi (dvije riječi od kojih je prva dio neke cjeline, a druga ta cjelina, npr. *nos – lice*), te parova nasumično izabranih riječi koje nisu povezane semantičkim odnosima. Mjere su vrednovane koristeći prosječnu preciznost (engl. *average precision*) opisanu u radu (Kotlerman et al., 2010). Prosječna preciznost je metoda vrednovanja korištena u postupcima ekstrakcije informacija koja kombinira preciznost, relevantnost i odziv određenog modela. Mjera prosječne preciznosti izračunata je za svaki par riječi iz navedenih skupova parova riječi, pri čemu najbolji mogući iznos prosječne preciznosti (AP = 1) predstavlja idealan slučaj u kojem su, pomoću trenutačno korištene mjere, za sve parove riječi iz promatranog skupa dobivene veće vrijednosti nego za bilo koji par iz preostalih skupova. Dakle, idealan slučaj u smislu inkluzije hiponima i hiperonima bio bi onaj u kojem bi

mjera	Sinonimi	Hiperonimi	Meronimi	Nasumični
$WeedsPrec$	0.45	0.40	0.31	0.32
$ClarkeDE$	0.45	0.39	0.28	0.33
$invCL$	**0.38**	**0.40**	0.31	0.34

Tablica 5.5: Srednja vrijednost prosječne preciznosti pojedine mjere po skupovima riječi

za skup parova hiperonima i hiponima prosječna preciznost bila jednaka 1. Rezultati vrednovanja mjera u vidu srednje vrijednosti prosječne preciznosti prema skupu riječi prikazani su u tablici 5.5.

Rezultati vrednovanja mjera pokazali su da su sve navedene mjere uspješne u identifikaciji parova hiperonima i hiponima prikazanim ovako definiranim distribucijskim vektorima u slučaju usporedbe s vektorima meronima i nasumično izabranih parova riječi. S druge strane, mjere $WeedsPrec$ i $ClarkeDE$ nisu se pokazale uspješne u zadataku prepoznavanja parova hiponima i hiperonima u skupu s parovima sinonima riječi – prosječna preciznost tih mjera veća je u slučaju skupa sastavljenog od sinonima. Za razliku od njih, mjera $invCL$ pokazala se uspješnom i u tom slučaju – najveća prosječna preciznost postignuta je upravo na skupu sastavljenom od parova hiponima i hiperonima riječi, veća i od prosječne preciznosti na skupu sinonima (odnos tih vrijednosti podebljan je i u tablici 5.5. Ovi rezultati opravdavaju i samu definiciju distribucijske hipoteze inkluzije, budući da je vidljivo da se korištenjem mjera definiranih prema toj hipotezi, vrlo uspješno mogu identificirati pravi parovi hiperonima i hiponima riječi.

U samom postupku nadogradnje PLF-modela u smislu maksimizacije razine semantičke inkluzije, potrebno je u početku odrediti u kojoj je mjeri razina semantičke inkluzije prisutna u već postojećoj verziji modela. Kako bi se odredila postojeća razina inkluzije, primijenjene su navedene mjere za određivanje razine inkluzije na vektore imenica, te vektore izraza pridjev–imenica, subjekt–glagol i glagol–objekt ekstrahirane iz korpusa, kao i na vektore istih izraza dobivene PLF-modelom. Parovi hiperonima i hiponima određeni su koristeći alat WordNet, budući da se njegovim korištenjem lako mogu dohvatiti hiperonimi određenih riječi, te se prema tome mogu povezati i parovi određenih izraza koji sadrže te hiperonime, odnosno njihove hiponime.

Analiza prisutnosti razinu inkluzije provedena je na dva načina. U prvom postupku analize određena je prosječna razina inkluzije prema pojedinoj mjeri između parova vektora odgovarajućih hiperonima i hiponima, i to između vektora imenica ekstrahiranih iz korpusa, između vektora izraza koji sadržavaju hiperonime i hiponime

ekstrahiranih iz korpusa, te između vektora istih izraza, ali dobivenih primjenom kompozicijskih pravila PLF-modela. Također, analiza je provedena i na vektorima iz skupa word2vec vektora značenja riječi. Word2vec vektori su vektori značenja riječi dobiveni korištenjem neuronskih mreža za modeliranje značenja pojedine riječi, pri čemu se za modeliranje konteksta u kojem se riječ nalazi koristi pristup vreće riječi (engl. *bag-of-words*) ili skip-gram postupak u kojem se svaka riječ iz konteksta ne vrednuje jednako, već ovisno o udaljenosti riječi iz konteksta od trenutačno promatrane riječi (Mikolov et al., 2013). U okviru provedenih eksperimenata s inkluzijom korišten je javno dostupan prethodno trenirani skup vektora.[4] Važno je napomenuti da word2vec skup vektora riječi sadrži poseban vektor za svaki oblik riječi, dakle svakom nelematiziranom obliku riječi pripada jedan vektor koji opisuje značenje upravo tog oblika riječi, za razliku od vektora korišenih u PLF-modelu, gdje jedan vektor odgovara isključivo lematiziranom obliku riječi. U provedenoj analizi kao reprezentativan vektor određene riječi (imenice), korišten je vektor koji se nalazi uz lematizirani oblik riječi.

Skup sastavljen od parova vektora odgovarajućih hiperonima i hiponima imenica sadrži ukupno 36.553 para imenica u takvom međusobnom odnosu, dok skup parova izraza oblika pridjev–imenica, glagol–objekt i subjekt–glagol, koji sadržavaju hiperonime i hiponime kao imenice, sadrži ukupno 400.515 parova izraza s odgovarajućim hiperonima i hiponimima kao imenicama. Kako bi se analiziralo je li predložene mjere uistinu uspješno detektiraju višu razinu inkluzije između parova hiperonima i hiponima, za svaki analizirani par vektora, izračunata je i mjera inkluzije između tri nasumično izabrana vektora imenice, odnosno izraza, i vektora trenutačno promatranog hiponima. Za mjere razine inkluzije s nasumično izabranim vektorima također je izračunata prosječna vrijednost, koja bi u slučaju idealno definiranih vektora trebala biti što manja. Analiza je obavljena koristeći prethodno predstavljene mjere ($invCL$ i $ClarkeDE$), te su rezultati dani u tablici 5.6.

Rezultati analize iz tablice 5.6 pokazuju da je, sudeći po $ClarkeDE$ mjeri, razina inkluzije među parovima hiperonima i hiponima u vektorima dobivenim PLF-modelom već prisutna u znatnoj mjeri. Naime, upravo u tim vektorima prisutna je najveća razina inkluzije, prema $ClarkeDE$ mjeri, u usporedbi s razinama u vektorima imenica i izraza ekstrahiranim iz korpusa (0.76 kod PLF-modela, 0.63 u vektorima imenica, te 0.58 u vektorima izraza iz korpusa). Također, u PLF-modelu prisutna je najveća razlika između razine inkluzije među stvarnim parovima vektora hiperonima i hiponima i razine inkluzije među vektorima hiponima i nasumično izabranih vektora

[4] Skup vektora korišten u analizi izgradila je tvrtka Google, pa je i javni skup dostupan preko njihovih službenih stranica: `www.google.com`

Parovi vektora	$invCL$		$ClarkeDE$	
	Pravi	Nasumični	Pravi	Nasumični
Imenice (word2vec)	0.4435	0.4720	0.5532	0.5625
Imenice	0.4821	0.4884	0.6341	0.6017
Vektori izraza (korpus)	0.4779	0.4949	0.5798	0.5789
Vektori izraza (PLF)	0.4086	0.4890	0.7627	0.6095

Tablica 5.6: Rezultati analize razine inkluzije prisutne među vektorima iz word2vec skupa, vektorima ekstrahiranim iz korpusa i vektorima dobivenim PLF-modelom

imenica ili izraza (razlika od 0.16 u PLF-modelu, te 0.03 u vektorima imenica). Što se tiče razine inkluzije u vektorima ekstrahiranim iz korpusa, u njima je manje primjetna prisutnost inkluzije u odnosu na vektore izraza. Inkluzija je u najmanjoj mjeri prisutna u slučaju vektora iz word2vec skupa, gdje je čak i razina inkluzije između nasumičnih parova imenica veća od razine inkluzije među stvarnim parovima hiponima i hiperonima. Takvo stanje u vektorima word2vec skupa vektora može se objasniti činjenicom da je u tom skupu svaki oblik riječi prikazan s njemu odgovarajućim vektorom značenja – dakle, ne postoji zajednički vektor koji bi obuhvatio ukupno značenje svih oblika riječi, pa je samim time i značenje riječi koja jest u lematiziranom obliku nedovoljno dobro definirano.

Mogući razlog loših rezultata za izraze ekstrahirane iz korpusa je činjenica da parovi odgovarajućih izraza hiperonima i hiponima ne moraju imati iste frekvencije pojavljivanja u korpusu, pa je samim time i različito definiran njihov vektor značenja, ovisno o učestalosti pojavljivanja određenog izraza u korpusu. Zbog toga je i usporedba tih vektora, u smislu prisutnosti inkluzije među njima, dala lošije rezultate od slučaja u kojem se uspoređuju vektori izraza dobiveni PLF-modelom, u kojem se pomoću vektora i matrice glagola detaljnije određuje semantičko značenje izraza, odnosno sami izgled vektora izraza. Zaključak koji slijedi iz tih rezultata jest da je PLF-model već u ovakvom obliku prilično sposoban razlikovati stvarne parove hiperonima i hiponima od nasumično izabranih parova.

Rezultati dobiveni mjerom $invCL$ nisu se pokazali pretjerano dobrima u smislu identifikacije odnosa inkluzije za dane parove vektora. Posebno loši rezultati primjetni su u slučaju PLF-modela (0.4), gdje je čak i prosječna vrijednost inkluzije među nasumičnim parovima veća od inkluzije među stvarnim parovima vektora hiperonima i hiponima. Razlog takvim rezultatima može se pronaći u činjenici da $invCL$ mjera za inkluziju definirana u radu Lenci i Benotto (2012) namijenjena različitom tipu vek-

tora od onog korištenih u PLF-modelu. Naime, vektori za koje je ta mjera definirana sastoje se od isključivo pozitivnih vrijednosti u svim svojim dimenzijama, dok su vektori korišteni pri izgradnji PLF-modela sastavljeni od realnih brojeva iz raspona $(-1, 1)$. Samim time, za vektore korištene u PLF-modelu ne mogu se očekivati isti rezultati pri računanju invCL mjere inkluzije, prvenstveno zbog različitih rezultata u izrazu $min(w_u(f), w_v(f))$ uzrokovanih negativnim vrijednostima težina w_x.

Drugi oblik analize proveden je također koristeći parove vektora hiperonima i hiponima imenica iz word2vec skupa vektora, vektora imenica i izraza iz korpusa, te vektora izraza dobivenih PLF-modelom, kao i koristeći nasumično izabrane vektore imenice i izraza. Ovaj put, izračunata je mjera inkluzije za svaki od parova vektora hiperonima i hiponima, te za par vektora hiponima i jednog od tri nasumično odabrana vektora imenice ili izraza. Cilj je bio odrediti u koliko slučajeva će mjera inkluzije biti najveća za stvarni par vektora hiperonima i hiponima u odnosu na inkluziju između hiponima i nasumično izabranog vektora. Rezultati ovakve analize prikazani su u tablici 5.7.

Parovi vektora	$invCL$	$ClarkeDE$
Imenice (word2vec)	0.0932	0.1978
Imenice	0.2016	0.5186
Vektori izraza (korpus)	0.1099	0.3278
Vektori izraza (PLF)	0.0042	0.9835

Tablica 5.7: Rezultati uspješnosti identifikacije pravih parova hiperonima i hiponima u kombinaciji s nasumičnim parovima vektora

Rezultati ovakve analize inkluzije pokazali su se iznimno dobrim za PLF-model. Postotak od čak 98% (korištenjem $ClarkeDE$ mjere) uspješno detektiranih ispravnih parova hiperonima i hiponima u kombinaciji s nasumičnim parovima vektora, uvjerljivo je najbolji rezultat u usporedbi s mjerama izračunatim nad ostalim vektorima. Kao i u prethodnoj analizi i u ovoj se pokazalo da mjera $invCL$ nije prikladna za ovako definirane distribucijske vektore. Najbolji rezultat korištenjem te mjere postignut je u vektorima imenica iz korpusa (tek 20%), premda su i oni definirani na neprikladan način za uspješno korištenje te mjere.

Dobiveni rezultati analiza pokazali su da je određena razina inkluzije već i u ovako definiranom PLF-modelu sačuvana u značajnoj mjeri. Ipak, postoji pretpostavka da je i povećanje trenutno prisutne razine inkluzije moguće uz modifikaciju postupka učenja pri treniranju modela. Očito je mjera $ClarkeDE$ prikladna za identifikaciju parova

izraza između kojih je opravdano maksimizirati inkluziju, pa bi se upravo ta mjera mogla uključiti u proces treniranja modela, i to u samu funkciju gubitka, pri čemu bi se maksimizirao iznos $ClarkeDE$ mjere za prethodno definirane parove hiperonima i hiponima. Međutim, budući da $ClarkeDE$ funkcija u svojoj definiciji sadrži izraz $min(w_u(f), w_v(f))$ koji nije derivabilan, funkciju gubitka koja bi u sebi sadržavala mjeru $ClarkeDE$ nije moguće derivirati, jer nije derivabilna. Iz tog razloga postupak treniranja PLF-modela više ne bi bilo moguće provesti korištenjem L2-regularizirane regresije, ali ni standardnim postupkom određivanja težina modela gradijentnim spustom, budući da gradijent tako definirane funkcije nije moguće odrediti. Preostaje razmotriti proširenje modela s takvom funkcijom gubitka na postupke nekonveksne optimizacije, uključivo postupke temeljene na genetskim algoritmima ili postupke optimizacije s ograničenjima. U okviru diplomskog rada takva proširenja nisu implementirana, ali svakako predstavljaju zanimljivo područje istraživanja za daljnji rad na modelu.

6. PLF-model za hrvatski jezik

Izgradnja PLF-modela za hrvatski jezik ostvarena je na isti način kao i implementacija istog modela za engleski jezik, prema postupku opisanom u Paperno et al. (2014). Vrednovanje modela provedeno je također na dužim frazama oblika pridjev–imenica–glagol–pridjev–imenica, ali na drugačiji način od vrednovanja modela u navedenom radu. Detalji izgradnje i vrednovanja modela opisani su u nastavku.

6.1. Izgradnja modela

Kao korpus za izgradnju modela korišten je fHrWaC (Šnajder et al., 2013), derivat korpusa hrWaC (Ljubešić i Erjavec, 2011). Korpus hrWaC je web-korpus stranica domene .hr, iz kojih su filtrirani svi dijelovi koji ne sadržavaju tekst prirodnog jezika (primjerice dijelovi koda i strukture stranice). U fHrWaC korpusu sav sadržaj prikupljenih web-stranica predobrađen je alatima za rastavljanje rečenica, lematizaciju i označavanje vrste riječi, kao i alatom za izgradnju stabla povezanosti riječi u rečenici prema njihovoj ulozi (engl. *dependency tree parsing*). Tako dobiven korpus sastoji se od nešto više od 50 milijuna rečenica, odnosno ukupno 1.232.632.208 obrađenih riječi.

U samom korpusu prebrojana su pojavljivanja pojedinih imenica, glagola i pridjeva, te je tako izdvojeno 30,000 najčešćih riječi za koje je izgrađena osnovna matrica supojavljivanja (engl. *co-occurrence matrix*), koristeći prozor veličine tri pri prolasku kroz korpus. Uz imenice, pridjeve i glagola iz skupa 30,000 najčešćih riječi, vektori supojavljivanja izgrađeni su i za izraze pridjev–imenica, subjekt–glagol i glagol–objekt, i to za one pridjeve i glagole koji su se koristili pri označavanju i ispitivanju skupa podataka za vrednovanje (opisanog u sljedećem odlomku). Na sve tako dobivene vektora supojavljivanja primijenjeni su isti postupci izmjene matrice kao i u slučaju PLF-modela za engleski jezik: izračun pozitivne uzajamne zajedničke informacije (engl. *Positive Pointwise Mutual Information*), smanjenje dimenzionalnosti matrice sa dimenzija $30,000 \times 30,000$ na dimenzije 300×300 korištenjem postupka singularne dekompozicije (engl. *Singular Value Decomposition*), te konačna normalizacija tako

dobivenih vektora na duljinu 1.

Matrice PLF-modela trenirane su korištenjem dobivenih vektora postupkom Ridge regresije uz generaliziranu unakrsnu provjeru, kao i u modelu za engleski jezik. Također, osim matrica za originalni PLF-model, izgrađene su i matrice za predloženu prilagodbu faze treniranja modela (Gupta et al., 2015), te je skup za vrednovanje modela vrednovan i za tu prilagodbu, kao i za prilagodbu faze ispitivanja, za koju se koriste matrice originalnog PLF-modela.

6.2. Vrednovanje modela

Vrednovanje modela za hrvatski jezik obavljeno je na drugačiji način u odnosu na vrednovanje modela za engleski jezik. Model je i u ovom slučaju vrednovan na izrazima oblika pridjev–imenica–glagol–pridjev–imenica (*anvan* oblik), međutim postupak označavanja tih izraza razlikuje se od onog obavljenog na skupovima korištenim pri vrednovanju modela za engleski jezik.

Naime, u postupku označavanja *anvan* izraza za engleski jezik, označivačima su ponuđene po dva izraza koja se razlikuju samo u glagolu, te su oni morali označiti u kolikoj su mjeri ta dva izraza međusobno semantički slična, i to na ljestvici od 1 do 7. Takav postupak može se smatrati prikladnim ako se radi o izrazima koji jesu međusobno slični, no ne i u slučaju međusobno semantički nepovezanih izraza. U tim slučajevima nije jasno kako odrediti u kojoj su mjeri neki izrazi međusobno nepovezani, budući da dva međusobno semantički različita izraza mogu biti različita u više načina. Samim time postupak označavanja sličnosti, odnosno različitosti pojedinih izraza na istoj ljestvici ne predstavlja odgovarajući pristup tom problemu.

Sami skup *anvan* izraza za vrednovanje modela konstruiran je na sličan način kao i *anvan* skupovi za engleski jezik (Kartsaklis et al., 2013; Grefenstette, 2013). Skup je sastavljen od 18 *anvan* izraza s ukupno 6 različitih glagola (3 izraza po glagolu), odabranih iz skupa višeznačnih glagola hrvatskog jezika, prema načelu što višeg stupnja višeznačnosti uz kriterij odabira isključivo tranzitivnih glagola. Skup višeznačnih glagola dobiven je koristeći definicije riječi navedene u Hrvatskom jezičnom portalu,[1] gdje su za svaku riječ navedena sva značenja koja ona može imati. Subjekti i objekti pojedinih glagola, kao i njima odgovarajući pridjevi, odabrani su koristeći distribucijsku memoriju hrvatskog jezika (Šnajder et al., 2013) koja između ostalih informacija, sadrži i one o najčešćim subjektima i objektima pojedinih glagola, kao i o najčešćim

[1] Hrvatski jezični portal: `hjp.znanje.hr`

atributima pojedinih imenica.

Tako izgrađenih 18 izraza označilo je ukupno troje označivača, ali na drugačiji način negoli u slučaju skupa izraza za engleski jezik. Naime, u postupku označavanja svaki od označivača za svaku je od riječi u izrazu naveo do tri sinonima koji bi tu riječ mogli zamijeniti u danom izrazu, ali tako da se semantičko značenje izraza ne mijenja. Takvim postupkom označavanja dobiva se skup izraza koji u određenoj mjeri imaju dosta slično semantičko značenje kao i početni izraz. Nekoliko primjera konstruiranih izraza zajedno sa zamjenskim riječima neke od riječi u izrazu dani su u tablici 6.1 (riječ čije zamjenske riječi su u tablici je podebljana), dok je cjelokupan skup izraza korištenih u postupku označavanja dan u dodatku A, zajedno sa zamjenskim riječima koje su označivači ponudili za riječi u pojedinom izrazu.

Izraz	Zamjenske riječi
sportski *automobil prijeći veliku udaljenost*	*brz, luksuzan, opasan, trkaći*
nezavisna ***država*** *voditi žestoku borbu*	*carstvo, kraljevina, nacija, pokrajina, republika, zemlja*
dobar igrač ***dati*** *pobjednički gol*	*pogoditi, postići, zabiti, zadati*
gradsko vijeće dati ***pozitivno*** *mišljenje*	*afirmativno, dobro, odobravajuće, potvrdno, povoljno, sjajno*
popularan pjevač izdati posljednji ***album***	*CD, nosač, pjesma, ploča*

Tablica 6.1: Primjeri izraza iz skupa *anvan* izraza za označavanje sa zamjenskim riječima koje su ponudili označivači

Očekivano je da će semantička sličnost početnog izraza s izrazom u kojem je jedna od riječi zamijenjena nasumično izabranom biti manja od sličnosti s izrazom u kojem je ta ista riječ zamijenjena s nekom od riječi koju su ponudili označivači. Upravo na taj način pristupljeno je i vrednovanju modela, usporedbom sličnosti početnog izraza s izrazom koji sadrži ponuđenu riječ sinonim, sa sličnošću početnog izraza s izrazom koji sadrži nasumično odabranu riječ.

Iz tih razloga, vrednovanju modela pristupljeno je kao vrednovanju TOEFL-zadataka, pri čemu je za svaki par početnog izraza i izraza sa zamijenjenom riječju, konstruirano još tri para početnog izraza i izraza sa nasumično odabranom trenutno obrađivanom riječi. U tako definiranim zadacima, izračunata je sličnost između parova izraza, te je za svaki model izračunata točnost modela, definirana kao postotak zadataka u kojima je najveća sličnost s početnim izrazom dobivena u slučaju izraza s riječju koja dolazi iz skupa sinonima. Kao modeli za izgradnju vektora izraza korišteni su jednostavan aditivni model, jednostavan multiplikativni model, osnovni PLF-model te prilagodbe PLF-modela: model temeljen na prilagodbi faze treniranja, te model temeljen na prilagodbi faze ispitivanja. Rezultati vrednovanja modela, u vidu ukupne točnosti

Model	Točnost
Aditivan model	73.96%
Multiplikativan model	47.68%
PLF-model	**75.43%**
PLF-model prilagodba treniranja	67.00%
PLF-model prilagodba ispitivanja	73.48%

Tablica 6.2: Ukupne točnosti modela u vrednovanju PLF-modela na TOEFL-zadacima za hrvatski jezik

Model	A1	N1	V	A2	N2
Aditivan model	73.75	**92.04**	45.98	**69.23**	**89.74**
Multiplikativan model	40.00	61.36	34.48	39.74	62.82
PLF-model	**75.00**	85.23	**65.51**	65.38	85.89
PLF-model prilagodba treniranja	58.75	89.77	50.57	51.28	83.33
PLF-model prilagodba ispitivanja	72.50	85.23	59.77	65.38	84.61

Tablica 6.3: Točnosti modela (u postotcima) u vrednovanju PLF-modela na TOEFL-zadacima za hrvatski jezik, prema grupama izraza.

modela na TOEFL-zadacima, dani su u tablici 6.2.

Rezultati pokazuju da je, gledajući ukupne točnosti pojedinog modela (dakle neovisno o vrsti riječi koja je mijenjana – pridjev, imenica ili glagol), PLF-model postigao najveću točnost u iznosu od 75.43%, iako se ta točnost nije pokazala statistički značajnom (prema McNemarovom testu usporedbe značajnosti razlike između binarnih predikcija modela).[2] Nakon njega, najboljim se pokazao jednostavan aditivni model, sa točnošću od 73.96%. Međutim, ako se definirani zadaci za vrednovanje modela podijele u grupe prema vrsti riječi koja je zamijenjena u izrazu, odnosno prema tome je li u izrazu zamijenjen pridjev subjekta, subjekt, glagol, pridjev objekta ili sami objekt, dobivamo 5 različitih grupa izraza za vrednovanje, a s njima i različite točnosti modela ovisno o kojoj se grupi izraza radi. Detaljnija analiza točnosti modela ovisno o grupama izraza dana je u tablici 6.3.

Analiza obavljena na grupama izraza pokazala je zanimljive rezultate. PLF-model pokazao se najboljim u dvije od pet grupa – u onoj u kojoj je zamijenjen pridjev su-

[2]McNemarov test opisan je detaljnije na: `en.wikipedia.org/wiki/McNemar\%27s_test`

bjekta (A1) i u grupi u kojoj je zamijenjen glagol (V). Jednostavan aditivni model postigao je najveću točnost u tri od pet grupa – u grupi u kojoj je zamijenjen subjekt (N1), pridjev objekta (A2) i sam objekt (N2). Gledajući samo grupu u kojoj je u izrazima zamijenjen glagol, rezultati su pokazali da je PLF-model, zajedno sa svojim prilagodbama znatno uspješniji od aditivnog i multiplikativnog modela (razlika od 20% u točnosti između PLF-modela i aditivnog modela), što je usporedivo s rezultatima vrednovanja PLF-modela za engleski jezik. Naime, u *anvan* izrazima korištenim pri vrednovanju modela za engleski jezik, izrazi se također razlikuju samo u glagolu, pa je ovakav rezultat PLF-modela u grupi izraza koji se razlikuju samo u glagolu potvrda uspješnosti modela, te njegove primijenjivosti na hrvatski jezik. Nad ovako dobivenim rezultatima izračunata je razina statističke značajnosti korištenjem McNemarovog testa za po dva modela s najboljim rezultatima u svakoj grupi izraza, pri čemu je testirana razlika u različito klasificiranim primjerima pojedinih modela. Ipak, u niti jednoj grupi najbolji model nije postigao statistički značajnu razliku u odnosu na drugi najbolji model u pojedinoj grupi izraza.

U ostalim grupama izraza, PLF-model nije postigao bolje rezultate od jednostavnog aditivnog modela, kao ni od multiplikativnog (koji zaostaje za ostalim modelima u svim grupama). Mogući uzrok lošijih performansi PLF-modela u ostalim grupama može se pronaći u samoj definiciji primjene modela. Naime, u pravilima kompozicije izraza u PLF-modelu, glagol u *anvan* frazi ima najveći utjecaj na izgled konačnog vektora izraza. Vektori izraza pridjev–subjekt i pridjev–objekt množe se matricama subjekta i objekta glagola, te se zbroju ta dva vektora dodaje i sami vektor glagola. Iz toga je očito da je u kompoziciji vektora cjelokupnog izraza najviše parametara koji opisuju značenje glagola u izrazu. Samim time, promjenom imenice ili pridjeva u izrazu, promijeniti će se manji broj parametara modela negoli u slučaju promjene glagola u izrazu. S druge strane, u jednostavnom aditivnom modelu svaka od riječi ima jednak utjecaj na izgled konačnog vektora, budući da se konačni vektor gradi kao zbroj vrijednosti vektora svih riječi u izrazu, dakle svaka riječ predstavljena je s jednakim brojem parametara određenim duljinom vektora riječi. Samim time, promjena bilo koje riječi u izrazu imati će jednak utjecaj na izgradnju konačnog vektora izraza, što nije slučaj kod PLF-modela. Upitno je, dakle, koliko je opravdan toliki broj parametara koji opisuju značenje glagola s obzirom na samu semantiku prirodnog jezika, posebno gledajući opisani način vrednovanja uspješnosti modela (kao TOEFL–zadataka). Naime, zamjenom imenice u jednom *anvan* izrazu s potpuno semantički različitom imenicom, neće se promijeniti dovoljan broj parametara u kompoziciji PLF-modela kako bi takva zamjena bila primjetna u konačnom vektoru izraza, dok će zamjena glagola izmijeniti

vrlo velik broj parametara i samim time dovesti do znatno različitog vektora značenja izmijenjenog izraza u odnosu na vektor značenja početnog izraza.

Uzevši u obzir motivaciju iza samog modela leksičke funkcije – formalnu semantiku prirodnog jezika, prema kojoj se jedna riječ u izrazu ponaša kao funkcija koja djeluje na značenje druge riječi, PLF-model uspješno je primijenio ideju takvog utjecaja riječi na semantiku izraza u vektorski prostor. Provedena analiza po grupama izraza pokazala je da je PLF-model zaista uspješan model kompozicijske distribucijske semantike u pogledu identifikacije semantički sličnih izraza. Međutim, postoji još prostora za poboljšanja i ispitivanja modela, budući da u prirodnom jeziku samo semantičko značenje pojedinog izraza ne mora nužno biti u najvećoj mjeri određeno značenjem glagola, kao što je slučaj u kompoziciji PLF-modela, u kojoj najveći broj parametara pri izgradnji vektora izraza pripada upravo glagolu. Ako bi se u modelu uspješno implementirao takav pristup, u kojem bi i promjena imenice ili pridjeva promijenila značenje izraza u dovoljnoj mjeri, model bi potencijalno postizao bolje rezultate i u slučajevima zamjene imenice ili pridjeva u izrazu, iako bi u tom slučaju bilo potrebno

7. Mogućnosti primjene PLF-modela

Osim primjene PLF-modela na spomenute probleme semantičke sličnosti između višerječnih izraza, model je moguće primijeniti i na druge semantičke probleme. U okviru diplomskog rada analizirana je primjena PLF-modela za hrvatski jezik u zadatku semantičke kompozitnosti višerječnih izraza te u zadatku semantičke devijantnosti višerječnih izraza. U nastavku su opisani navedeni problemi te predstavljeni rezultati dobiveni primjenom PLF-modela na njih.

7.1. Semantička kompozitnost

Semantička kompozitnost izraza predstavlja razinu kompozicije izraza definiranu utjecajem značenja pojedinog dijela izraza na semantičko značenje cjelokupnog izraza (Baldwin, 2006). Drugim riječima, ona opisuje u kojoj je mjeri semantičko značenje izraza doista sastavljeno od semantičkih značenja njegovih dijelova kada se oni koriste samostalno. Primjer jednog potpuno kompozitnog semantičkog izraza je izraz "*maslinovo ulje*" (ulje koje je zaista nastalo od maslina), dok je primjer jednog nekompozitnog izraza izraz "*hladni rat*" (rat koji je "hladan" u prenesenom značenju). Dakle, glavna karakteristika nekompozitnih izraza je to što se značenje izraza ne može odrediti iz doslovnih značenja njegovih dijelova.

Eksperimenti u kojima se analizirala uspješnost pojedinih modela kompozicijske distribucijske semantike za hrvatski jezik opisani su u radu (Šnajder i Almić, 2015). Za potrebe analize kompozitnosti višerječnih izraza sastavljen je prigodni skup kompozitnih i nekompozitnih izraza od ukupno 200 izraza (100 kompozitnih i 100 nekompozitnih). Stupanj kompozitnosti pojedinog izraza označilo je 24 označivača, pri čemu su za svaki ponuđeni izraz označili u kojoj je mjeri taj izraz za njih kompozitan, i to na ljestvici od 1 do 5 (1 označava nekompozitan izraz, 5 kompozitan).

Otkrivanje semantički nekompozitnih izraza obavljeno je koristeći modele kompozicijske distribucijske semantike za izgradnju vektora značenja izraza pomoću algebarskih zapisa značenja njegovih dijelova. Tako izgrađen vektor, za čiju izgradnju su se

koristili zapisi koji sadrže značenja njegovih dijelova, usporedio se s vektorom izraza ekstrahiranim iz korpusa koji opisuje njegovo stvarno značenje. Očekivano ponašanje uspješnih modela kompozicijske distribucijske semantike je postizanje velike razlike u sličnosti između vektora dobivenih kompozicijom algebarskih zapisa dijelova izraza i vektora izraza ekstrahiranih iz korpusa u slučaju obiju vektora.

Vrednovanje u radu (Šnajder i Almić, 2015) provedeno je računajući Spearmanov koeficijent korelacije između kosinusne sličnosti između vektora izraza dobivenih modelom i vektora izraza ekstrahiranih iz korpusa i medijana oznaka kompozitnosti pojedinog izraza od strane označivača. Između svih vrednovanih modela opisanih u navedenom radu, kao najbolji model za otkrivanje kompozitnosti pojedinog izraza istaknuo se model linearne kombinacije sastavljen od kompozicije jednostavnog aditivnog modela, multiplikativnog modela, te dvaju težinskih modela u kojima je veća težina pridana vektoru prve riječi u izrazu, odnosno vektoru druge riječi u izrazu (Spearmanov koeficijent za ovaj model iznosi 0.48). Na drugom mjestu po uspješnosti našao se jednostavan aditivni model (0.46).

Na istom skupu izraza vrednovan je i PLF-model, i to na dva načina. Prvi način odnosi se na standardan postupak izgradnje PLF-modela, dakle treniranje matrica pojedinih pridjeva i glagola sa svim vektorima izraza u kojima se pojedini pridjev ili glagol pojavljuje. Drugi način podrazumijeva malo izmijenjen način treniranja matrica u odnosu na originalan. Naime, u problemu semantičke kompozitnosti zapravo se određuje u kojoj mjeri se značenja pojedinih dijelova izraza prenose u izgradnju značenja cjelokupnog izraza. Zbog toga bi bilo primjereno vrednovati i verziju PLF-modela u kojem se matrice pridjeva i glagola ne grade iz apsolutno svih izraza u kojima se taj pridjev ili glagol koristi, već bi bilo smisleno izostaviti one izraze čija se kompozitnost ispituje. Na taj način izuzeli bi se ti primjeri iz skupa za treniranje matrica, pa bi samim time matrica pojedinog pridjeva ili glagola imala drugačiji pristup izgradnji vektora značenja za izraz koji nije viđen u postupku treniranja. Pokazalo se da je takav pristup prikladniji za određivanje kompozitnosti pojedinih izraza.

Rezultati vrednovanja dviju navedenih verzija PLF-modela u usporedbi s aditivnim i multiplikativnim modelom dani su u tablici 7.1. Jednostavan aditivni model postiže bolje rezultate od PLF-modela i predložene izmijenjene verzije PLF-modela, iako nedovoljno statistički značajne prema rezultatu uparenog t-testa u kojem je dobivena p-vrijednost iznosa 0.284. Uzrok boljih rezultata aditivnog modela u odnosu na ostale može se pronaći u razlikama u treniranju zapisa značenja riječi u pojedinim modelima i u samim zapisima značenja korištenim u modelima. Naime, u PLF-modelu za prikaz zapisa značenja pojedine riječi koriste se vektori i matrice, pri čemu se matrice izravno

Model	Spearmanov ρ
Aditivan model	0.44
Multiplikativan model	-0.19
PLF-model	0.40
PLF-model (izmijenjeni)	0.41

Tablica 7.1: Spearmanov koeficijent korelacije za različite modele

treniraju s vektorima izraza ekstrahiranim iz korpusa (vektorima kompozitnih i nekompozitnih izraza), dakle postupkom nadziranog strojnog učenja pri čemu se uče predvidjeti vektori izraza iz vektora imenica. S druge strane, u jednostavnom aditivnom modelu za zapis značenja riječi koriste se samo vektori supojavljivanja pojedine riječi u korpusu, koji za svoju izgradnju ne koriste vektore značenja izraza ekstrahirane iz korpusa. Zbog postupka nadziranog strojnog učenja za izgradnju matrica, PLF-model u sebi sadrži veći broj parametara (u vektorima i matricama), kojima se uspješnije prikazuju vektori izraza ekstrahirani iz korpusa, negoli aditivan model u kojem se vektori izraza procjenjuju samo s vektorima značenja pojedinih riječi. Prisutnost vektora kompozitnih i nekompozitnih izraza u postupku treniranja PLF-modela, kao i povećani broj parametara modela koji u sebi sadrže informacije o značenju pojedine riječi dovode do veće sličnosti izgrađenih vektora izraza vektorima ekstrahiranim iz korpusa, čime se ne može dovoljno uspješno primijetiti razlika između kompozitnih i nekompozitnih vektora izraza ekstrahiranih iz korpusa i vektora dobivenih PLF-modelom. Zbog toga je jasno da će aditivan model s manjim brojem parametara i izostankom vektora izraza u postupku izgradnje modela, uspješnije modelirati razliku između vektora izraza ekstrahiranih iz korpusa i onih dobivenih modelom te tako uspješnije modelirati i semantičku kompozitnost višerječnih izraza.

7.2. Semantička devijantnost

Semantička devijantnost značajka je određenih jezičnih izraza koji su po svojoj strukturi gramatički nepravilni, predstavljaju lažne činjenice ili su jednostavno besmisleni. Kada je riječ u izrazima sastavljenima od dvije riječi semantička devijantnost određena je samo međusobnim odnosom te dvije riječi koje čine izraz. Samim time, izraz može biti ili semantički devijantan ili semantički smislen. S druge strane, ukoliko se promatra semantička devijantnost višerječnog izraza sastavljenog od tri ili više riječi, nije u

potpunosti jednostavno odrediti je li izraz potpuno devijantan ili ne. Moguće je, naime, da je u danom izrazu samo jedna riječ devijantna u odnosu na druge, dok je ostatk izraza potpuno semantički smislen. Tada govorimo o djelomičnoj semantičkoj devijantnosti izraza. Radikalniji primjer je onaj u kojem su sve riječi u izrazu devijantne, pa je tako i sam izraz u potpunosti semantički devijantan.

7.2.1. Semantička devijantnost kraćih izraza

Sam problem semantičke devijantnosti nije dosad u dovoljnoj mjeri istražen i definiran, vjerojatno zbog činjenice da je devijantnost sama po sebi jako širok pojam – izraz može biti devijantan na više različitih, prethodno spomenutih načina. U radu (Vecchi et al., 2011) opisan je postupak identifikacije semantički devijantnih izraza oblika pridjev–imenica na engleskom jeziku, koristeći modele kompozicijske distribucijske semantike. Izrazi korišteni pri vrednovanju uspješnosti modela u identificiranju semantički devijantnih izraza, konstruirani su iz korpusa engleskog jezika, pri čemu su za pridjeve korištene u izrazima izabrani pridjevi iz skupa 200 najčešćih pridjeva u korpusu, te su kombinirani s imenicama iz skupa 8,000 najčešćih imenica iz korpusa. Iz tako dobivenog skupa kombinacija pridjeva i imenica izdvojeni su izrazi s 30 nasumično izabranih pridjeva, iz kojih je dodatno izdvojeno po 100 nasumično izabranih izraza. Svaki od tako dobivenih izraza, dvojica autora označila su kao devijantan, srednje devijantan ili nedevijantan izraz, te su u konačnici u skupove devijantnih, odnosno nedevijantnih izraza, dodali samo one izraze oko kojih su se oboje složili u postupku označavanja. Tako su nastali skupovi od 413 devijantnih (primjerice, *parliamentary potato – parlamentarna rajčica* i *blind pronounciation – slijepi izgovor*) i 280 nedevijantnih izraza (izrazi kao što su *blind cook – slijepi kuhar* ili *vulnerable gunman – ranjiv streljač*).

Kako bi vrednovali sposobnost modela kompozicijske distribucijske semantike u prepoznavanju devijantnosti takvih izraza, predložili su tri mjere za identifikaciju devijantnosti izraza:

1. duljina vektora izraza dobivenih modelom;
2. kosinusna sličnost vektora izraza s vektorom odgovarajuće imenice koja je u izrazu;
3. gustoća susjedstva vektora izraza u vektorskom prostoru imenica, pridjeva i izraza oblika pridjev–imenica.

Vektorski prostor u kojem je mjerena posljednje predložena mjera (gustoća susjedstva vektora) izgrađen je od 45.000 izraza – 8.000 najčešćih imenica, 4.000 najčešćih pridjeva i 33.000 najčešćih izraza pridjev–imenica. Za izgradnju vektora supojavljivanja korišten je skup od 10.000 najčešćih imenica, pridjeva i glagola, te je tako dobivena matrica dimenzija 45.000×10.000 reducirana SVD-om na dimenzije 45.000×300. Tako dobiven vektorski prostor korišten je za računanje posljednje dvije predložene mjere.

U postupku vrednovanja modela u zadatku semantičke devijantnosti izraza, predložene mjere izračunate su za sve izraze iz skupa devijantnih i nedevijantnih izraza, te je određena prosječna vrijednost tih mjera za svaki skup izraza. Tako dobivene mjere uspoređene su dvostranim Welchovim *t* testom s izračunatom mjerom značaja, pri čemu su hipoteze za pojedine mjere definirane na sljedeći način:

1. vektori nedevijantnih izraza trebali bi biti veće duljine od vektora devijantnih fraza,

2. vektori nedevijantnih izraza trebali bi imati veću kosinusnu sličnost s vektorima imenica koje se u izrazu pojavljuju od vektora devijantnih izraza i njima odgovarajućih imenica,

3. vektori nedevijantnih izraza trebali bi imati veću prosječnu kosinusnu sličnost sa svojim najbližim susjedima u spomenutom vektorskom prostoru od vektora devijantnih izraza i njihovih susjeda u istom prostoru.

Prva hipoteza motivirana je činjenicom da dimenzije distribucijskog vektora pojedinog izraza zapravo govore u kojoj se mjeri taj izraz koristi u nekom značenju. Samim time, male vrijednosti u većini dimenzija vektora izraza značile bi da je izraz besmislen – ne mogu se odrediti konteksti u kojima se izraz koristi. Druga hipoteza temelji se na ideji da bi svaka smislena kombinacija pridjeva i imenice, odnosno imenice i glagola, trebala u određenoj mjeri očuvati značenje imenice i u vektoru izraza. Ukoliko je vektor izraza vrlo različit od vektora imenice riječ je o devijantnom izrazu, budući da je u tako dobivenom vektoru izgubljeno osnovno značenje imenice. Posljednja hipoteza temelji se na pretpostavci da se u vektorskom prostoru, u kojem se nalaze vektori izraza i pojedinih riječi, doista nalaze izrazi koji obuhvaćaju širok raspon često korištenih konteksta u kojima se riječi u korpusu nalaze. Samim time, ako izgrađeni vektor izraza nema mnogo susjeda u okolini tako definiranog vektorskog prostora, očito je riječ o vektoru izraza koji ne pripada pronađenim kontekstima u korpusu.

Model	duljina	kosinus	gustoća
add	**7.89**	0.31	2.63
mult	3.16	-0.56	2.68
lm	0.16	0.55	-0.23
alm	0.48	**1.37**	**3.12**

Tablica 7.2: Rezultati *t* testa za razlike prosječnih vrijednosti pojedinih mjera između skupova devijantnih i nedevijantnih izraza

Modeli kojima su izrazi vrednovani su: jednostavan aditivni model (add), jednostavan multiplikativni model (mult), modelom linearnog mapiranja (lm – model temeljen na težinskom zbroju dvaju vektora za izgradnju konačnog vektora značenja izraza (Guevara, 2010)), te modelom linearnog mapiranja prema pridjevu (alm) što je zapravo matrica PLF-modela za pojedini pridjev. Modeli su vrednovani računajući razliku u prosječnim vrijednostima pojedine mjere za svaki od skupova izraza – skup devijantnih i skup nedevijantnih izraza. Rezultati vrednovanja modela na ovako definiran način, u vidu *t* vrijednosti za pojedini model i mjeru, prikazani su u tablici 7.2.

Vrednovanje je pokazalo da jednostavan aditivni model, kao i jednostavan multiplikativni model, uspješno identificiraju devijantne izraze koristeći mjeru duljine vektora kao način pronalaženja devijantnih izraza. Matrice PLF-modela pokazale su se najprikladnijima za identifikaciju devijantnih izraza pomoću gustoće susjedstva vektora u vektorskom prostoru, kao i pri korištenju mjere kosinusne sličnosti između vektora izraza i vektora imenica, premda je u tom slučaju postignuta najmanja razlika u skupovima devijantnih i nedevijantnih izraza.

Rezultati u cjelini pokazali su da su modeli kompozicijske distribucijske semantike sposobni na ovako definirane načine identificirati devijantne izraze, te ih u određenoj mjeri i razlikovati od nedevijantnih. Na tragu tih otkrića, provedeno je i vrednovanje uspješnosti PLF-modela u zadatku identifikacije devijantnih izraza na hrvatskome jeziku sastavljenih od više od dvije riječi, koje je opisano u sljedećem odlomku.

7.2.2. Semantička devijantnost duljih izraza

Semantička devijantnost duljih izraza, odnosno izraza koji sadrže više od dvije riječi, složeniji je problem od semantičke devijantnosti kraćih izraza. Naime, u slučaju duljih izraza postoji i veći broj riječi koje mogu biti devijantne u odnosu na ostatak izraza. Primjerice, kod prethodno spomenutih *anvan* izraza, moguće je da je samo pridjev koji

opisuje subjekt devijantan, ili pak da je u cijelom izrazu samo glagol devijantan, dok su ostale riječi u izrazu smislene i u takvom međusobnom odnosu upotrebljive. Takav izraz, u kojemu je jedna riječ devijantna, može se smatrati devijantnim u cjelini, ali ne u istoj mjeri u kojoj je izraz sastavljen od dvije riječi devijantan. Naime, u duljem izrazu s jednom devijantnom riječi postoji i veći broj nedevijantnih riječi, koje utječu na devijantnost, odnosno nedevijantnost izraza u cjelini.

Ukoliko se u izrazu nalazi više od jedne devijantne riječi, primjerice ukoliko je u *anvan* izrazu uz devijantni pridjev koji opisuje subjekt, devijantan i pridjev koji opisuje objekt, devijantnost izraza u cjelini veća je od one izraza sa samo jednom devijantnom riječi. Međutim, razina te devijantnosti nije jasno definirana, već bi se moglo reći da ovisi o riječi koja je devijantna i njenom utjecaju na značenje izraza u cjelini, pa time i njegovu devijantnost. Moguće je samim time da je i izraz sa samo jednom devijantnom riječi više devijantan od onog s dvije ili više devijantnih riječi, ukoliko je utjecaj jedne devijantne riječi na devijantnost izraza veći od utjecaja dviju ili više devijantnih riječi u izrazu.

U cilju vrednovanja uspješnosti identifikacije devijantnosti duljih izraza na hrvatskome jeziku pomoću modela kompozicijske distribucijske semantike, sastavljen je skup od 150 potencijalno devijantnih *anvan* izraza. Svaki od izraza označen je od strane 5 označivača, pri čemu je svaki označivač za svaki ponuđeni izraz označio svoj subjektivan dojam devijantnosti danog izraza, i to na ljestvici od 1 do 5 (gdje 1 označava potpuno devijantan izraz, a 5 potpuno nedevijantan izraz). Kako bi se postigao podjednak omjer izraza sa svim mogućim vrijednostima takve ljestvice, skup podataka konstruiran je koristeći skup smislenih *anvan* izraza u kojima je nasumično izabrano od jedne do 4 riječi, koje su zamijenjene nasumično izabranim riječima. Na taj način stvoren je skup izraza koji bi, ovisno o devijantnosti pojedine riječi u njima, mogli biti u različitoj mjeri devijantni. Primjeri nekih tako konstruiranih izraza, zajedno s njihovim prosječnim oznakama, te medijanom oznaka označivača, dani su u tablici 7.3, dok je cjelokupan skup izraza korištenih pri označavanju, zajedno s izračunatim medijanom i prosječnom ocjenom devijantnosti, dan u dodatku B.

Za potrebe vrednovanja modela pomoću prethodno definiranih mjera za devijantnost kraćih izraza, izgrađen je vektorski prostor sastavljen od 30.000 najčešćih riječi (imenica, pridjeva i glagola) iz korpusa fHrWaC, te određen broj vektora izraza oblika pridjev–imenica, subjekt–glagol, te glagol–objekt. Za ekstrakciju vektora kraćih izraza iz korpusa korišteni su pridjevi i glagoli iz skupa 100 najčešćih pridjeva, odnosno glagola u korpusu fHrWaC, čime je u vektorski prostor od 30.000 najčešćih riječi dodano i 281.530 vektora oblika pridjev–imenica (gdje je pridjev iz skupa 100 najčešćih

Izraz	Medijan	Prosječna oznaka
oteta sloboda sadržavati poljoprivrednu ostavku	1	1.0
nepotreban minus dati tranzicijski gol	1	2.0
članska iskaznica priznati ekološku pogodnost	2	2.6
zaljubljen par osjetiti liberalnu pomoć	2	2.8
riješena klima voditi znanstveno istraživanje	3	2.6
uspješan tenisač igrati težak let	3	2.8
suvremena znanost ponuditi ispravno piće	4	3.2
oštar organizator prodati besplatnu ulaznicu	4	3.6
dobar igrač dati pobjednički gol	5	5.0

Tablica 7.3: Primjeri izraza iz skupa potencijalno devijantnih izraza, zajedno s medijanima i prosječnim oznakama devijantnosti

pridjeva u korpusu), te 352.753 vektora oblika subjekt–glagol i 203.082 vektora oblika glagol–objekt (gdje su glagoli iz skupa 100 najčešćih glagola u korpusu). Tako dobiven vektorski prostor, s preko 830.000 vektora, korišten je za usporedbu sličnosti vektora izraza s vektorima riječi i vektorima izraza ekstrahiranim iz korpusa.

Postupak vrednovanja uspješnosti identifikacije devijantnih izraza pomoću modela kompozicijske distribucijske semantike proveden je na sličan način kao i kod kraćih izraza. Međutim, u slučaju duljih izraza kao mjera za devijantnost izraza korištena je samo mjera kosinusne sličnosti između prvih 20 susjeda vektora u izgrađenom vektorskom prostoru. Mjera duljine vektora nije korištena zbog činjenice da se u PLF-modelu konačni vektori izraza normaliziraju na duljinu 1, a osim toga normaliziraju se i određeni vektori dobiveni u prethodnim koracima računanja konačnog vektora izraza, primjerice vektor dobiven množenjem vektora imenice subjekta s matricom pridjeva. Samim time, mjera duljine vektora nije prikladna za izračun devijantnosti duljih izraza, budući da se duljina vektora izraza definira u samom modelu. Mjera kosinusne sličnosti vektora izraza s vektorom njemu odgovarajuće imenice nije prikladna za korištenje u PLF-modelu, budući da u duljim izrazima ne postoji odgovarajuća referentna riječ s kojom bi se vektor izraza mogao usporediti. Jedina u potpunosti prikladna mjera za računanje devijantnosti izraza je mjera prosječne kosinusne sličnosti vektora izraza s njegovih N najbližih susjeda u vektorskom prostoru, jer je taj prostor sastavljen od vektora riječi i kraćih izraza koji su usporedivi s dobivenim vektorom značenja duljeg izraza. Upravo u ovakvoj primjeni PLF-modela do izražaja dolazi njegova sposobnost

Grupa izraza	Broj izraza
Grupa 1	49
Grupa 2	38
Grupa 3	22
Grupa 4	22
Grupa 5	19

Tablica 7.4: Broj izraza po pojedinim grupama izraza, grupirani prema medijanu oznaka devijantnosti izraza

prikaza višerječnog izraza jedinstvenim vektorom koji je usporediv s vektorima izraza različitih duljina. Neovisno o duljini višerječnog izraza izgrađeni vektor će u konačnici biti jednake duljine kao i vektor jedne riječi ili izraza sastavljenog od dvije riječi, a uz to će biti i usporedivi u zajedničkom vektorskom prostoru.

Izrazi iz skupa za označavanje podijeljeni su u pet grupa, prema medijanu oznake devijantnosti koje su dali označivači. Takva podjela rezultirala je grupama izraza podjednakih veličina, ili barem ne pretjerano različitih. Raspodjela izraza po pojedinim grupama medijana nalazi se u tablici 7.4.

Vektori duljih izraza određeni su korištenjem pet različitih modela kompozicijske distribucijske semantike: jednostavnog aditivnog modela, jednostavnog multiplikativnog modela, PLF-modela, PLF-modela s prilagodbom faze treniranja i PLF-modela s prilagodbom faze ispitivanja. Za svaki tako dobiven vektor izraza izračunata je prosječna vrijednost kosinusne sličnosti između njega i njegovih 10 najbližih susjeda u vektorskom prostoru (prema kosinusnoj sličnosti), te je na razini grupe izračunat prosjek tih vrijednosti. Odabir broja 10 kao broja najbližih susjeda vektora u vektorskom prostoru obavljen je po uzoru na eksperimente određivanja semantičke devijantnosti kraćih izraza, opisane u radu (Lenci i Benotto, 2012), u kojima je također kao broj najbližih susjeda trenutačno promatranog vektora izabran broj 10. Prosječne vrijednosti grupa izraza prema korištenom modelu prikazane su u tablici 7.5.

Rezultati vrednovanja modela na ovako definiranom zadataku pokazali su da su zapravo svi modeli, osim jednostavnog multiplikativnog modela, sposobni na određeni način razlikovati potpuno devijantan od nedevijantnog izraza. Kao najbolji u tome, ističu se jednostavan aditivni model i prilagodba PLF-modela u fazi ispitivanja, kod kojih je razlika u prosječnoj kosinusnoj sličnosti između grupa 1 i 5 0.4, odnosno 0.6. Originalno predloženi PLF-model također postiže razliku u sličnosti između grupa 1 i

Model	1	2	3	4	5
Aditivni model	0.7866	0.7886	0.7881	0.8011	0.8265
Multiplikativni model	0.8420	0.8417	0.8403	0.8413	0.8425
PLF-model	0.6693	0.6700	0.6861	0.6781	0.6837
PLF-model prilagodba treniranja	0.2362	0.2361	0.2639	0.2525	0.3033
PLF-model prilagodba ispitivanja	0.5633	0.5479	0.5630	0.5742	0.6235

Tablica 7.5: Prosječne kosinusne sličnosti između prvih 10 susjeda vektora izraza po grupama izraza određenim medijanom oznaka označivača

5, no i dalje manju negoli u slučaju prethodno spomenutih modela (0.14).

Što se tiče porasta prosječne sličnosti od grupe 1 prema grupi 5, određeni porast je primjetan kod jednostavnog aditivnog modela, te originalnog PLF-modela i njegove prilagodbe ispitne faze. Međutim, ti porasti ipak nisu dovoljno veliki, niti u potpunosti prisutni između svih grupa (između nekih grupa čak postoji i pad sličnosti). Ipak, može se reći da postoji razlika između prosječne kosinusne sličnosti sa susjedima potpuno devijantnih i nedevijantnih izraza, koja predstavlja zanimljivo područje istraživanja u budućnosti.

Uz navedeno vrednovanje po grupama devijantnosti, obavljeno je i vrednovanje pomoću binarnog klasifikatora u cilju klasifikacije duljih izraza kao devijantnih odnosno nedevijantnih. Označeni izrazi podijeljeni su u dva skupa – skup devijantnih izraza, čiji je medijan oznaka manji od 2.5 i skup nedevijantnih izraza, s medijanom oznaka većim od 2.5. Skup devijantnih izraza sadrži ukupno 87 izraza, dok skup nedevijantnih izraza sadrži ukupno 63 izraza. Kao binarni klasifikator kojim se određivala devijantnost danog izraza korišten je stroj potpornih vektora (engl. *Support Vector Machine – SVM*) s radijalnim baznim funkcijama. Ulazni vektori pojedinog izraza sastavljeni su od prosječnih kosinusnih sličnosti vektora danog izraza s njegovih 10 najbližih susjeda u izgrađenom vektorskom prostoru, pri čemu je vektor značenja izraza dobiven nekim od 5 modela korištenih pri vrednovanju devijantnosti po grupama. Dakle, ulazni vektor svakog izraza u modelu SVM-a sastoji se od 5 vrijednosti, koje predstavljaju prosječnu kosinus sličnost s 10 susjeda vektora dobivenog aditivnim i multiplikativnim modelom, PLF-modelom, te prilagodbama PLF-modela (ispitne i faze treniranja). Budući da je skup označenih izraza sastavljen od svega 150 izraza, za odabir optimalnih parametara modela korišten je postupak ugniježđene unakrsne provjere, i to s vanjskom petljom u kojoj je obavljena unakrsna provjera izdvajanjem jednog primjera i unutarnjom pet-

Mjera	Rezultat
Točnost	0.5839
Preciznost	0.6091
Odziv	0.7791
F1-mjera	0.6837

Tablica 7.6: Rezultati vrednovanja binarnog klasifikatora za određivanje devijantnosti izraza

ljom u kojoj je obavljena unakrsna provjera podjelom skupa za treniranje na 5 dijelova (engl. *K-fold cross-validation*). Vrednovanje u takvom postupku obavljeno je računanjem točnosti modela na pojedinim testnim primjerima. Konačni rezultati modela, u vidu točnosti, ali i preciznosti, odziva i F1-mjere, dani su u tablici 7.6.

Dobiveni rezultati pokazali su da je klasifikator s točnošću većom od 50% uspio razvrstati izraze u skupove devijantnosti, definirane na spomenut način (s granicom medijana od 2.5). Također, odziv klasifikatora je veći od njegove preciznosti, što zapravo govori da možda sama granica između devijantnih i nedevijantnih izraza nije u potpunosti jasno definirana, pa je zbog toga klasifikator i nedevijantne izraze označio devijantnim (što potvrđuje razlika između preciznosti i odziva).

Iako su rezultati vrednovanja pokazali da su pojedini korišteni modeli sposobni razlikovati devijantne od nedevijantnih izraza, ipak je potrebno sami postupak vrednovanja u određenoj mjeri izmijeniti, čime bi se postigli potencijalno bolji i kvalitetniji rezultati. Prvenstveno se to odnosi na postupak konstruiranja skupa potencijalno devijantnih izraza. Naime, kao što je prethodno naglašeno, izraz sam po sebi može biti devijantan na više različitih načina u različitim mjerama. Kako bi skup potencijalno devijantnih izraza bio što prikladnije sastavljen, u smislu prisutnosti različitih vrsta devijantnosti, potrebno je detaljnije proučiti sami fenomen semantičke devijantnosti izraza, te u skup uključiti izraza koji su devijantni na različite načine. Primjerice, u slučaju devijantnosti kraćih izraza devijantnima se smatraju izrazi koji su po svom značenju besmisleni (nije zamisliv kontekst u kojem bi se mogli upotrijebiti) – *parlamentarna rajčica* ili pak izrazi koji su kontradiktorni sa značenjem riječi koja ga čini – *bezbojna boja*. Osim načina na koje izrazi mogu biti devijantni, potrebno je i detaljnije proučiti u kojoj mjeri devijantnost jednog dijela izraza utječe na devijantnost izraza u cjelinu. Uzevši u obzir *anvan* izraze, takvo razmatranje odnosi se, primjerice, na utjecaj devijantnosti izraza subjekt–glagol na devijantnost cjelokupnog *anvan* izraza. Skup podataka korišten u radu mogao bi poslužiti kao početni korak takve analize, pri

čemu bi za izraze koji su označeni devijantnima trebalo dodatno označiti koji je točno dio izraza devijantan za pojedinog označivača. Time bi se mogla odrediti upravo korelacija između devijantnosti dijela izraza s devijantnošću izraza u cjelini. Također, budući da je određivanje razine semantičke devijantnosti izraza prilično subjektivan problem, skup izraza trebalo bi označiti više od pet označivača (koliko je označilo korišteni skup), čime bi oznake devijantnosti pojedinog izraza bile pouzdanije.

8. Zaključak

Modeli kompozicijske distribucijske semantike pokazali su se izuzetno uspješnima u modeliranju značenja pojedinih riječi ili kraćih izraza pomoću vektora u semantičkom vektorskom prostoru. Unatoč uspjehu u modeliranju značenja tako kratkih jezičnih konstrukcija, i dalje postoje poteškoće u modeliranju značenja dužih jezičnih izraza u jednom takvom vektorskom prostoru. Cilj diplomskog rada bio je proučiti postojeće modele kompozicijske distribucijske semantike i vrednovati njihovu uspješnost na takvim zadacima.

U diplomskom radu opisani su postojeći distribucijsko semantički modeli i modeli kompozicijske distribucijske semantike, s posebnim naglaskom na modele temeljene na tenzorskoj algebri: model leksičke funkcije (Baroni i Zamparelli, 2010), praktični model leksičke funkcije (Paperno et al., 2014) te predložene prilagodbe praktičnog modela leksičke funkcije (Gupta et al., 2015). Razmotrena su i proširenja modela u vidu maksimizacije sličnosti između sinonima i maksimizacije inkluzije između parova hiponima i hiperonima riječi. Izgrađen je praktični model leksičke funkcije, zajedno s predloženim prilagodbama modela, za hrvatski jezik. Za potrebe vrednovanja modela sastavljen je skup izraza oblika pridjev–imenica–glagol–pridjev–imenica, pri čemu su označivači za svaku od riječi u pojedinom izrazu ponudili zamjenske riječi kojima se čuva semantičko značenje početnog izraza. Praktični model leksičke funkcije postigao je najveću točnost u zadatku prepoznavanja izraza semantički najsličnijeg početnom izrazu (75.43%).

Osim u zadatku semantičke sličnosti, izgrađeni model vrednovan je i u zadatku semantičke kompozitnosti kraćih izraza (Šnajder i Almić, 2015), kao i u zadatku semantičke devijantnosti duljih izraza. U problemu semantičke kompozitnosti model nije pokazao bolje rezultate od onih koji su postignuti u spomenutom radu. Za vrednovanje modela u problemu semantičke devijantnosti izgrađen je poseban skup potencijalno devijantnih izraza, čija je razina devijantnosti označena od strane petero označivača. Pri vrednovanju se pokazalo da je model spodoban u određenoj mjeri razlikovati potpuno devijantne od nedevijantnih izraza – prosječna kosinusna sličnost sa susjedima

devijantnih izraza iznosi 0.6693, dok je prosječna kosinusna sličnost sa susjedima nedevijantnih izraza 0.6837.

U daljnjem radu na modelu praktične leksičke funkcije trebalo bi implementirati predloženi postupak maksimizacije inkluzije hiperonima i hiponima, ali i razmisliti o drugačijim načinima definiranja kompozicijskih pravila modela, čime bi se izbjegao veliki utjecaj koji značenje glagola ima na značenje cjelokupnog izraza. U zadatku semantičke devijantnosti duljih izraza potrebno bi bilo konstruirati skup potencijalno devijantnih izraza na prikladniji način, uzimajući u obzir sve načine na koje pojedini izraz može biti devijantan.

Ukupno gledano, praktični model leksičke funkcije pokazao se kao jednostavan i uspješan model u modeliranju vektorskog prikaza semantičkog značenja višerječnih izraza. Ipak, još uvijek ima prostora za napredak i postizanje boljih rezultata u svim navedenim primjenama modela, a koristeći neke od zaključaka koji su navedeni u radu.

LITERATURA

Ethem Alpaydin. *Introduction to Machine Learning*. The MIT Press, 2nd izdanju, 2010. ISBN 026201243X, 9780262012430.

Timothy Baldwin. Compositionality and multiword expressions: Six of one, half a dozen of the other. U *Invited talk given at the COLING/ACL'06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, July*, 2006.

Marco Baroni i Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

Marco Baroni i Roberto Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. U *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, stranice 1183–1193. Association for Computational Linguistics, 2010.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, i Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

Gene H Golub, Michael Heath, i Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Edward Grefenstette. Category-theoretic quantitative compositional distributional models of natural language semantics. *CoRR*, abs/1311.1539, 2013. URL `http://arxiv.org/abs/1311.1539`.

Edward Grefenstette, Georgiana Dinu, Yao-Zhong Zhang, Mehrnoosh Sadrzadeh, i Marco Baroni. Multi-step regression learning for compositional distributional semantics. *arXiv preprint arXiv:1301.6939*, 2013.

Gregory Grefenstette. *Explorations in automatic thesaurus discovery*, svezak 278. Springer Science & Business Media, 1994.

Emiliano Guevara. A regression model of adjective-noun compositionality in distributional semantics. U *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, stranice 33–37. Association for Computational Linguistics, 2010.

Abhijeet Gupta, Jason Utt, i Sebastian Padó. Dissecting the practical lexical function model for compositional distributional semantics. *Lexical and Computational Semantics (* SEM 2015)*, stranica 153, 2015.

Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, i Stephen Pulman. Separating disambiguation from composition in distributional semantics. U *CoNLL*, stranice 114–123, 2013.

Lili Kotlerman, Ido Dagan, Idan Szpektor, i Maayan Zhitomirsky-Geffet. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16 (04):359–389, 2010.

Alessandro Lenci i Giulia Benotto. Identifying hypernyms in distributional semantic spaces. U *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, stranice 75–79. Association for Computational Linguistics, 2012.

Nikola Ljubešić i Tomaž Erjavec. hrwac and slwac: Compiling web corpora for croatian and slovene. U Ivan Habernal i Václav Matousek, urednici, *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, Lecture Notes in Computer Science, stranice 395–402. Springer, 2011.

Kevin Lund i Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208, 1996.

Tomas Mikolov, Kai Chen, Greg Corrado, i Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Jeff Mitchell i Mirella Lapata. Vector-based models of semantic composition. U *ACL*, stranice 236–244, 2008.

Kee Siong Ng. A simple explanation of partial least squares, 2013.

Denis Paperno, Nghia The Pham, i Marco Baroni. A practical and linguistically-motivated approach to compositional distributional semantics. U *ACL (1)*, stranice 90–99, 2014.

Gerard Salton, Anita Wong, i Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

Jan Šnajder i Petra Almić. Modeling semantic compositionality of croatian multiword expressions. *Informatica*, 39(3):301, 2015.

Jan Šnajder, Sebastian Padó, i Željko Agić. Building and evaluating a distributional memory for croatian. U *51st Annual Meeting of the Association for Computational Linguistics*, stranice 784–789, 2013.

Eva Maria Vecchi, Marco Baroni, i Roberto Zamparelli. (linear) maps of the impossible: capturing semantic anomalies in distributional space. U *Proceedings of the Workshop on Distributional Semantics and Compositionality*, stranice 1–9. Association for Computational Linguistics, 2011.

Martin Warin i HM Volk. Using wordnet and semantic similarity to disambiguate an ontology. *Retrieved January*, 25:2008, 2004.

Dodatak A
Skup označenih *anvan* izraza u zadatku semantičke sličnosti

U nastavku se nalaze upute za označavanje *anvan* izraza u zadatku semantičke sličnosti izraza, koje su označivači dobili na uvid prije samog označavanja. Nakon uputa slijedi pregled izraza koji su označeni, zajedno sa zamjenskim riječima koje su označivači ponudili za pojedinu riječ.

A.1. Upute za označavanje

U nastavku se nalazi 90 zadataka s ukupno 18 različitih fraza oblika: pridjev–imenica–glagol–pridjev–imenica. Prva imenica predstavlja subjekt, druga objekt, a povezane su pomoću središnjeg glagola – predikata.

U svakom od zadataka jedna je riječ napisana velikim slovima. Za tu riječ potrebno je napisati do tri riječi koje bi ju mogle zamijeniti, ali tako da se semantičko značenje izraza ne mijenja ili barem ne previše. Jednostavnije rečeno, za riječ pisanu velikim slovima trebate napisati do tri sinonima.

Slijedi nekoliko primjera koji će Vam malo bolje pojasniti o čemu se radi:

1. iskusan natjecatelj OSVOJITI velika nagrada – dobiti, uzeti, odnijeti;
2. iskusan natjecatelj osvojiti VELIKA nagrada – važna, glavna, cijenjena;
3. iskusan natjecatelj osvojiti velika NAGRADA – priznanje, dobitak, zgoditak.

A.2. Skup označenih izraza

Tablica A.1 sadrži sve izraze korišene u označavanju, zajedno s pripadajućim zamjenskim riječima za trenutačno promatranu riječ u izrazu (naznačenu velikim slovima) i ukupnim brojem označivača koji su tu riječ ponudili kao zamjensku. Sve riječi u tablici dane su u svom lematiziranom obliku, pogodnom za obradu u korištenim modelima.

Izraz	**Zamjenske riječi**
ODLIČAN đak prijeći brz cesta	*izvrstan* (2), *dobar* (2), *vrli* (1), *vrhunski* (1), *uzoran* (1), *sjajan* (1), *marljiv* (1)
odličan ĐAK prijeći brz cesta	*učenik* (3), *školarac* (2), *student* (2)
odličan đak PRIJEĆI brz cesta	*putovati* (1), *proći* (1), *pretrčati* (1), *prelaziti* (1), *prekoračiti* (1)
odličan đak prijeći BRZ cesta	*hitar* (2), *prometan* (1), *opasan* (1)
odličan đak prijeći brz CESTA	*put* (2), *ulica* (1), *prometnica* (1), *kolnik* (1), *drum* (1), *autocesta* (1)
OZBILJAN kandidat prijeći bodovan prag	*zainteresiran* (1), *uzoran* (1), *uporan* (1), *spreman* (1), *siguran* (1), *prikladan* (1), *pravi* (1), *marljiv* (1)
ozbiljan KANDIDAT prijeći bodovan prag	*pristupnik* (2), *natjecatelj* (2), *aspirant* (1)
ozbiljan kandidat PRIJEĆI bodovan prag	*proći* (2), *zadovoljiti* (1), *preskočiti* (1), *ostvariti* (1), *nadvisiti* (1), *ispuniti* (1)
ozbiljan kandidat prijeći BODOVAN prag	*zadan* (1), *tražen* (1), *postavljen* (1), *minimalan* (1), *ciljni* (1)
ozbiljan kandidat prijeći bodovan PRAG	*granica* (2), *uvjet* (1), *razina* (1), *linija* (1), *limit* (1), *crta* (1), *cilj* (1)
SPORTSKI automobil prijeći velik udaljenost	*brz* (3), *trkaći* (1), *opasan* (1), *luksuzan* (1)
sportski AUTOMOBIL prijeći velik udaljenost	*auto* (3), *vozilo* (2), *sredstvo* (1), *kabriolet* (1)
sportski automobil PRIJEĆI velik udaljenost	*prevaliti* (2), *voziti* (1), *putovati* (1), *proći* (1), *prelaziti* (1), *pokriti* (1), *odvesti* (1)
sportski automobil prijeći VELIK udaljenost	*ogroman* (3), *dug* (2), *znatan* (1), *obilan* (1)
sportski automobil prijeći velik UDALJENOST	*put* (2), *razdaljina* (1), *dužina* (1), *dionica* (1)
NEPOZNAT muškarac baciti letimičan pogled	*stran* (2), *neznan* (2), *tajan* (1), *sumnjiv* (1)
nepoznat MUŠKARAC baciti letimičan pogled	*čovjek* (3), *tinejdžer* (1), *osoba* (1), *muž* (1), *gospodin* (1), *dječak* (1)
nepoznat muškarac BACITI letimičan pogled	*usmjeriti* (1), *uputiti* (1)
nepoznat muškarac baciti LETIMIČAN pogled	*površan* (2), *brz* (2), *nehajan* (1), *kratkotrajan* (1), *kratak* (1)
nepoznat muškarac baciti letimičan POGLED	*oko* (1)
EKOLOŠKI incident baciti težak ljaga	*prirodan* (2)
ekološki INCIDENT baciti težak ljaga	*nesreća* (2), *problem* (1), *nezgoda* (1), *katastrofa* (1), *izgred* (1), *havarija* (1)

ekološki incident BACITI težak ljaga	*staviti* (2), *stvoriti* (1), *pridodati* (1), *označiti* (1), *napraviti* (1)
ekološki incident baciti TEŽAK ljaga	*velik* (3), *ozbiljan* (2), *strašan* (1), *ružan* (1), *ogroman* (1)
ekološki incident baciti težak LJAGA	*sramota* (3), *mrlja* (2), *stigma* (1)
MASKIRAN napadač baciti atomski bomba	*neprepoznatljiv* (2), *nepoznat* (2), *sumnjiv* (1), *prerušen* (1)
maskiran NAPADAČ baciti atomski bomba	*terorist* (2), *zločinac* (1), *razbojnik* (1), *muškarac* (1), *lopov* (1), *agresor* (1)
maskiran napadač BACITI atomski bomba	*staviti* (1), *raznijeti* (1), *koristiti* (1), *izbaciti* (1), *detonirati* (1), *aktivirati* (1)
maskiran napadač baciti ATOMSKI bomba	*nuklearan* (3), *smrtonosan* (1), *opasan* (1), *fisijski* (1)
maskiran napadač baciti atomski BOMBA	*eksploziv* (2), *prasak* (1), *oružje* (1), *eksplozija* (1)
NEZAVISAN država voditi žestok borba	*samostalan* (2), *neovisan* (2), *suveren* (1), *slobodan* (1)
nezavisan DRŽAVA voditi žestok borba	*zemlja* (2), *republika* (1), *pokrajina* (1), *nacija* (1), *kraljevina* (1), *carstvo* (1)
nezavisan država VODITI žestok borba	*sudjelovati* (1), *imati* (1), *biti* (1)
nezavisan država voditi ŽESTOK borba	*smrtonosan* (1), *oštar* (1), *opasan* (1), *napet* (1), *krvav* (1), *intenzivan* (1), *buran* (1)
nezavisan država voditi žestok BORBA	*rat* (2), *bitka* (2), *tučnjava* (1), *meč* (1), *boj* (1)
LEGENDARAN trener voditi suparnički momčad	*uspješan* (2), *poznat* (2), *znamenit* (1), *izvanredan* (1), *cijenjen* (1)
legendaran TRENER voditi suparnički momčad	*voditelj* (2), *mentor* (1), *menadžer* (1), *izbornik* (1), *instruktor* (1)
legendaran trener VODITI suparnički momčad	*trenirati* (3), *učiti* (2), *savjetovati* (1), *predvoditi* (1), *obučavati* (1)
legendaran trener voditi SUPARNIČKI momčad	*protivnički* (3), *neprijateljski* (2), *susjedan* (1), *rivalski* (1)
legendaran trener voditi suparnički MOMČAD	*tim* (3), *ekipa* (3), *selekcija* (1), *društvo* (1)
MEĐUNARODAN udruga voditi znanstven istraživanje	*internacionalan* (3), *ujedinjen* (1)
međunarodan UDRUGA voditi znanstven istraživanje	*udruženje* (2), *zajednica* (1), *tim* (1), *organizacija* (1), *grupa* (1), *društvo* (1), *agencija* (1)
međunarodan udruga VODITI znanstven istraživanje	*raditi* (2), *zapovijedati* (1), *provoditi* (1), *predvoditi* (1), *pokretati* (1), *organizirati* (1), *nadgledati* (1)

međunarodan udruga voditi ZNANSTVEN istraživanje	*stručan* (1), *napredan* (1), *istraživački* (1), *akademski* (1)
međunarodan udruga voditi znanstven ISTRAŽIVANJE	*ispitivanje* (2), *traženje* (1), *rad* (1), *proučavanje* (1)
NOV predsjednik dati neopoziv ostavka	*novoizabran* (2), *svjež* (1), *sadašnji* (1), *nov* (1), *mlad* (1), *aktualan* (1)
nov PREDSJEDNIK dati neopoziv ostavka	*predstavnik* (2), *šef* (1), *vođa* (1), *voditelj* (1), *predstojnik* (1), *poglavar* (1), *direktor* (1)
nov predsjednik DATI neopoziv ostavka	*podnijeti* (2), *stavljati* (1), *priložiti* (1), *dati* (1)
nov predsjednik dati NEOPOZIV ostavka	*trajan* (1), *neodgodiv* (1), *konačan* (1)
nov predsjednik dati neopoziv OSTAVKA	*otkaz* (2), *rezignacija* (1), *odlazak* (1)
GRADSKI vijeće dati pozitivan mišljenje	*općinski* (2), *provincijski* (1), *mjesni* (1), *lokalan* (1)
gradski VIJEĆE dati pozitivan mišljenje	*skupština* (2), *zajednica* (1), *sabor* (1), *odbor* (1)
gradski vijeće DATI pozitivan mišljenje	*vraćati* (1), *stavljati* (1), *izraziti* (1), *iznijeti* (1), *dati* (1)
gradski vijeće dati POZITIVAN mišljenje	*dobar* (2), *sjajan* (1), *povoljan* (1), *potvrdan* (1), *odobravajući* (1), *afirmativan* (1)
gradski vijeće dati pozitivan MIŠLJENJE	*stav* (1), *stajalište* (1), *pogled* (1), *ocjena* (1), *gledište* (1)
DOBAR igrač dati pobjednički gol	*talentiran* (2), *odličan* (2), *uspješan* (1), *super* (1), *sjajan* (1), *kvalitetan* (1)
dobar IGRAČ dati pobjednički gol	*sportaš* (2), *čovjek* (1), *nogometaš* (1), *napadač* (1)
dobar igrač DATI pobjednički gol	*zabiti* (2), *pogoditi* (2), *zadati* (1), *postići* (1), *dati* (1)
dobar igrač dati POBJEDNIČKI gol	*slavljenički* (1), *pobjedonosan* (1), *odlučujući* (1), *bitan* (1)
dobar igrač dati pobjednički GOL	*zgoditak* (2), *pogodak* (2), *cilj* (1)
OBJEKTIVAN promatrač vidjeti bitan razlika	*nepristran* (2), *razuman* (1), *pošten* (1), *ozbiljan* (1), *nezavisan* (1), *neutralan* (1), *neovisan* (1)
objektivan PROMATRAČ vidjeti bitan razlika	*gledatelj* (2), *čovjek* (1), *posmatrač* (1), *osoba* (1), *gledaoc* (1)
objektivan promatrač VIDJETI bitan razlika	*uočiti* (3), *primijetiti* (3), *zamijetiti* (1), *uvidjeti* (1), *razumijeti* (1)
objektivan promatrač vidjeti BITAN razlika	*znatan* (2), *značajan* (1), *velik* (1), *važan* (1), *temeljan* (1), *suštinski* (1), *stvaran* (1)
objektivan promatrač vidjeti bitan RAZLIKA	*različitost* (1), *promjena* (1), *nesličnost* (1), *neslaganje* (1)

EKONOMSKI analitičar vidjeti alternativan rješenje	*financijski* (2), *gospodarski* (1)
ekonomski ANALITIČAR vidjeti alternativan rješenje	*stručnjak* (2), *znalac* (1), *savjetnik* (1), *poznavatelj* (1), *konzultant* (1)
ekonomski analitičar VIDJETI alternativan rješenje	*pronaći* (2), *primjetiti* (2), *znati* (1), *uvidjeti* (1), *uočiti* (1), *saznati* (1), *otkrivati* (1)
ekonomski analitičar vidjeti ALTERNATIVAN rješenje	*drugi* (2), *siguran* (1), *pomoćan* (1), *nov* (1), *moguć* (1), *drugačiji* (1), *dodatan* (1)
ekonomski analitičar vidjeti alternativan RJEŠENJE	*sredstvo* (1), *pristup* (1), *postupak* (1), *način* (1), *cilj* (1)
ZAINTERESIRAN posjetitelj vidjeti animiran film	*znatiželjan* (1), *zaintrigiran* (1), *uzbuđen* (1)
zainteresiran POSJETITELJ vidjeti animiran film	*čovjek* (1), *posjetilac* (1), *osoba* (1), *klijent* (1), *gost* (1), *gledatelj* (1)
zainteresiran posjetitelj VIDJETI animiran film	*pogledati* (2), *gledati* (2)
zainteresiran posjetitelj vidjeti ANIMIRAN film	*crtan* (3)
zainteresiran posjetitelj vidjeti animiran FILM	*serija* (1), *isječak* (1), *crtić* (1)
SADAŠNJI vlada izdati služben priopćenje	*trenutan* (3), *postojeći* (2), *vladajući* (1), *izabran* (1), *današnji* (1), *aktualan* (1)
sadašnji VLADA izdati služben priopćenje	*vlast* (2), *zajednica* (1), *skupština* (1)
sadašnji vlada IZDATI služben priopćenje	*objaviti* (2), *priopćiti* (1), *objelodaniti* (1), *izdati* (1), *donijeti* (1), *dati* (1)
sadašnji vlada izdati SLUŽBEN priopćenje	*javan* (1), *formalan* (1)
sadašnji vlada izdati služben PRIOPĆENJE	*odluka* (2), *izjava* (2), *vijest* (1), *rješenje* (1), *objava* (1), *obavijest* (1), *mišljenje* (1)
POPULARAN pjevač izdati posljednji album	*poznat* (2), *omiljen* (2), *slušan* (1), *cijenjen* (1)
popularan PJEVAČ izdati posljednji album	*zvijezda* (1), *svirač* (1), *izvođač* (1), *grupa* (1), *bend* (1)
popularan pjevač IZDATI posljednji album	*objaviti* (3), *staviti* (1), *snimiti* (1)
popularan pjevač izdati POSLJEDNJI album	*zadnji* (3), *najnoviji* (1)
popularan pjevač izdati posljednji ALBUM	*ploča* (2), *pjesma* (1), *nosač* (1), *CD* (1)
DRŽAVAN agencija izdati potreban dozvola	*županijski* (1), *nacionalan* (1), *gradski* (1)
državan AGENCIJA izdati potreban dozvola	*institucija* (2), *ured* (1), *organizacija* (1)
državan agencija IZDATI potreban dozvola	*dati* (3), *prodati* (1), *odobriti* (1), *dodijeliti* (1)
državan agencija izdati POTREBAN dozvola	*nužan* (2), *tražen* (1), *obavezan* (1), *neophodan* (1), *bitan* (1)
državan agencija izdati potreban DOZVOLA	*odobrenje* (2), *rješenje* (1), *pristanak* (1), *papir* (1), *obrazac* (1), *licenca* (1)

Tablica A.1: Izrazi iz skupa *anvan* izraza za označavanje sa zamjenskim riječima koje su ponudili označivači i ukupnim brojem označivača koji su ponudili pojedinu zamjensku riječ

Dodatak B
Skup označenih *anvan* izraza u zadatku semantičke devijantnosti

U nastavku se nalaze upute za označavanje *anvan* izraza u zadatku semantičke devijantnosti izraza, koje su označivači dobili na uvid prije samog označavanja. Nakon uputa slijedi pregled izraza koji su označeni, zajedno s medijanom i prosječnom oznakom devijantnosti za svaki od izraza (devijantnost je označena u razinama od 1 – potpuno devijantan izraz, do 5 – potpuno nedevijantan izraz).

B.1. Upute za označavanje

U nastavku se nalaze izrazi za koje trebate označiti razinu semantičke devijantnosti.

Semantička devijantnost može se definirati kao svaki međusobni odnos dviju ili više riječi koji je sam po sebi besmislen ili kontradiktoran. Primjerice, fraze poput "*bezbojna zelena boja*", "*burno opuštanje*" ili "*parlamentarna rajčica*" smatraju se semantički devijantnima – riječi koje ih čine ne mogu se dovesti u takav međusoban odnos, odnosno nalaze se u pomalo neobičnom (neprirodnom) međusobnom odnosu.

U slučaju višerječnih izraza (sastavljenih od više od dvije riječi) problem je malo složeniji. Devijantnost cjelokupne fraze ovisi u utjecajima devijantnosti njenih dijelova. Primjerice, moguće je da se u izraz ne uklapa samo glagol ili samo imenica, ali je moguće i da se u izraz ne uklapaju ni glagol ni imenica (kao ni ostatak riječi u izrazu). Zbog toga u sljedećim frazama, prema vlastitom subjektivnom dojmu, trebate označiti u kojoj mjeri je pojedini izraz za Vas devijantan: 1 označava potpuno devijantnu frazu, dok 5 označava potpuno smislenu (nedevijantnu) frazu.

B.2. Skup označenih izraza

Tablica B.1 sadrži sve izraze korišene u označavanju devijantnosti, zajedno s pripadajućim medijanom ocjena označivača i prosječnom vrijednosti tih oznaka za pojedini izraz. Devijantnost svakog izraza označilo je ukupno petoro označivača i to na cjelobrojnoj ljestvici od 1 do 5 (1 predstavlja potpuno devijantan, a 5 potpuno nedevijantan izraz).

Izraz	Medijan	Prosjek
nepokretan stol prijeći velik udaljenost	3	2.6
inovativan incident baciti težak ljaga	3	2.6
mjesečni automobil prijeći velik udaljenost	3	3.4
materijalan kandidat prijeći bodovan prag	3	3.2
otpušten predsjednik dati hranjiv ostavka	2	2.4
legendaran trener voditi suparnički stranka	4	4.2
otet sloboda sadržavati poljoprivredan ostavka	1	1.0
spontan predsjednik prijeći neopoziv prag	3	2.2
sportski automobil organizirati velik vjenčanje	3	2.6
zaštićen sitnica ugroziti javan mišljenje	4	3.0
talentiran vođa prepoznati nepoznat kolega	4	4.0
slučajan promatrač pronaći racionalan razlika	4	4.4
državan agencija izdati potreban dozvola	5	5.0
zračan agencija postaviti potreban brod	4	3.0
neiskusan sposobnost voditi nov odjel	3	2.4
zainteresiran zvučnik prikazati skraćen provizija	1	1.0
pripremljen natjecatelj prijeći velik udaljenost	5	5.0
zračan ribarstvo poduzeti hranjiv mjera	1	1.6
aktivan liječenje posjetiti prikladan lijek	3	3.0
ozbiljan konkurencija baciti dobrovoljan kazna	2	3.0
odvojen korekcija oboriti drven gol	1	1.2
atraktivan ravnoteža koristiti obnovljiv kultura	1	1.4
pristupačan magazin preporučiti posljednji album	3	2.8
biološki država voditi prijenosan borba	2	1.8
maskiran osvajač baciti pečen bomba	2	2.4
akademski sunce obasjati biološki utvrda	2	2.0
ekonomski opozicija predložiti alternativan dogovor	4	4.0
sadašnji vlada izdati službben priopćenje	5	5.0
industrijski odlagalište istraživati bijel tržište	3	2.8
odsutan posjetitelj vidjeti zanimljiv crtež	4	3.8
daljinski pobjednik sakriti prirodan plastika	2	1.8
sportski automobil prokomentirati velik udaljenost	2	2.2

ozbiljan bol upoznati diplomiran stan	1	1.2
objektivan promatrač uočiti bitan razlika	5	5.0
kazališni automobil prijeći velik postrojba	2	1.6
odličan đak prijeći vraćen cesta	2	2.2
popularan udruga primiti kraljevski posjet	4	4.0
sezonski kut pokrenuti stabilan dozvola	1	1.0
nezavisan izvedba voditi plodan borba	2	2.0
nezavisan država oštetiti emotivan borba	2	2.4
lovački akademija dati neopoziv stijena	2	1.8
rukometan igrač postići pobjednički koš	4	3.6
veseo bakterija obići pozitivan otok	1	1.4
anoniman igrač uvesti približan okolnost	2	2.0
sportski žlica upozoravati čist udaljenost	1	1.2
nov automobil posjetiti kišan demonstracija	2	2.4
pripremljen kandidat proći misaon intervju	4	4.4
brz automobil prijeći velik udaljenost	5	5.0
popularan pjevač izdati posljednji član	4	4.0
lokalan strah upaliti mračan svjetlo	1	1.2
budući općina vidjeti dinamičan sklonište	1	1.4
maskiran napadač baciti atomski bomba	5	4.8
opravdan olovka nacrtati morski ptica	2	2.2
nov izbor dati neopoziv manekenka	2	1.6
jeftin lijek pobijediti veseo bolest	3	3.4
pješački duhan dobiti pregovarački pozicija	1	1.4
staklen pismo naručiti prolazan udarac	1	1.0
ozbiljan korak prijeći obalan naslov	1	1.2
popularan pjevač izdati zadnji album	5	5.0
nepoznat jelo predstaviti poznat kuhinja	5	4.4
ključan vozilo prevariti naivan vratar	2	2.2
vrijedan radnik uzeti digitalan odmor	3	2.8
nov dijete dati neopoziv ostavka	2	1.6
ozbiljan pljačka prijeći treći prag	2	1.8
maskiran napadač shvatiti atomski bomba	2	2.6
težak fakultet prodati kratak dar	1	1.6
pijan dizel doživjeti velik poraz	1	1.4
kopnen umijeće izdati motiviran album	1	1.0
uspješan plivač preplivati valovit planina	3	2.8
sunčan igrač pretrčati cijeli teren	4	3.4
sportski automobil prijeći ogroman udaljenost	5	5.0
policijski prozor ubosti industrijski osmijeh	1	1.0
popularan pjevač izdati navijački album	5	5.0

talentiran glazbenik svirati usni gitara	3	3.2
snažan kava kupiti biološki narav	1	1.4
kraljevski vlada izdati službben posjedovanje	3	2.8
odličan đak prijeći proizveden snaga	2	1.8
riješen klima voditi znanstven istraživanje	2	1.6
eventualan muškarac baciti letimičan završnica	1	1.0
talentiran predsjednik garantirati neopoziv ostavka	5	4.6
dnevni igračka spasiti propao zabava	3	3.4
gradski vijeće dati pozitivan mišljenje	5	5.0
glazben ekran ukrasti nizak pozornost	1	1.2
uspješan klapa pjevati visok boja	3	2.8
internacionalan udruga voditi znanstven istraživanje	5	5.0
papirnat svečanost pojesti razuman biskvit	1	1.0
otpadan posjetitelj ugledati prehramben tenisač	1	1.2
romantičan semafor otežavati divlji tisak	1	1.2
opsežan rajčica prethoditi sposoban bomba	1	1.2
časan dužnosnik najaviti pozitivan eksplozija	4	3.6
otet incident darovati težak kartica	1	1.2
nov ušteda zadržavati promotivan ostavka	1	1.6
pješački sudnica izdati posljednji album	1	1.6
sportski automobil prijeći velik udaljenost	5	5.0
žedan igrač piti elektronski cesta	2	2.2
ozbiljan ponašanje zadovoljavati atomski sila	2	1.6
ekološki incident baciti težak sramota	4	3.6
riješen klima voditi znanstven istraživanje	3	2.6
suočen zgrada prijeći pravan izričaj	1	1.0
popularan pjevač izdati posljednji album	5	5.0
zainteresiran tehnologija predstaviti kvalifikacijski knjiga	1	1.2
ekonomski kemija pročitati ustanovljen rješenje	1	1.4
sretan udruga voditi znanstven istraživanje	4	4.0
odličan đak prijeći brz cesta	4	4.2
lokalan dizajn baciti bitan željeznica	1	1.4
pažljiv gledatelj vidjeti umirovljen znak	2	2.8
dobar igrač dati pobjednički gol	5	5.0
izvrstan natjecatelj opravdati visok stadion	3	3.0
kapitalan svjetlo ugledati suparnički momčad	1	1.4
vrijedan student prijeći opasan cesta	5	4.6
doživotan priključak provesti blagdanski vrijeme	1	1.6
praktičan novac priznati biološki korupcija	1	1.2
legendaran trener voditi suparnički tim	5	5.0
preporučen terapija uzrokovati internacionalan ozdravljenje	4	4.0

agresivan tlak sastaviti blokiran porast	1	1.2
dizajnerski vojnik emitirati plastični mjesec	1	1.2
dobar plivanje postići zapažen gol	1	1.4
suvremen znanost ponuditi ispravan piće	4	3.2
ukusan stadion privući mlad navijač	2	2.8
članski iskaznica priznati ekološki pogodnost	2	2.6
ozbiljan dogovor uključiti redovit ambicija	2	1.6
nepotreban minus dati tranzicijski gol	1	2.0
oštar organizator prodati besplatan ulaznica	4	3.6
uspješan tenisač igrati težak let	3	2.8
pametan stol učiti težak gradivo	2	2.4
iznenađen tunel servirati ukusan desert	1	1.4
star turist posjetiti poznat smijeh	2	2.2
glumački letjelica izvesti uzbudljiv mišljenje	1	1.4
odličan đak stećи slab snaga	2	2.4
mlad učenik raspolagati zračni cesta	2	1.8
međunarodan uzvrat voditi znanstven istraživanje	2	1.6
hladan grijalica zagrijati vruć prostor	2	2.0
uporan muškarac zapaliti kristalan vatra	3	3.2
veseo prilika vagati kamen povrće	1	1.2
otrovan sjedište upotrijebiti skriven nalaz	1	1.4
nogometan trener uloviti idealan formacija	4	4.4
iskusan natjecatelj opravdati spontan očekivanje	4	3.6
zaljubljen par osjetiti liberalan pomoć	2	2.8
apsolutan pas gristi virtualan namještaj	3	2.4
prihvaćen predmet motivirati izgubljen napad	1	1.6
ugrađen stvarnost poklopiti novinarski članak	1	1.4
državan izaslanstvo položiti optužen vijenac	2	2.8
star ormar čuvati iznenadan uspomena	4	3.8
instaliran grana trebati nov tipkovnica	1	1.6
iskusan lopta peći ukusan kolač	1	1.2
morski prostor pogoditi konačan rezultat	2	1.8
marljiv radnik pakirati gotov proizvod	5	4.4
kreativan udaljenost otkriti nov materijal	2	1.6
moderan pribor ugostiti ugledan osjećaj	1	1.0
uzrokovan promatrač baciti letimičan ugođaj	1	1.4

Tablica B.1: Izrazi iz skupa *anvan* izraza za označavanje devijantnosti s medijanom i prosječnom ocjenom devijantnosti dobivenim označavanjem devijantnosti za svaki izraz od strane petoro označivača.

Kompozicijska distribucijska semantika temeljena na modelu leksičke funkcije

Sažetak

Kompozicijska distribucijska semantika bavi se izgradnjom prikaza značenja višerječnih fraza u vektorskome prostoru. Rad opisuje u literaturi korištene modele kompozicijske distribucijske semantike, s naglaskom na modele temeljene na modelu leksičke funkcije. Posebno je proučen i opisan praktični model leksičke funkcije, zajedno s predloženim prilagodbama modela. Razmotrena su proširenja modela s obzirom na relaciju semantičke inkluzije među distribucijskim vektorima imenica. Izgrađen je praktični model leksičke funkcije za hrvatski jezik te je za potrebe njegovog vrednovanja sastavljen skup izraza za vrednovanje modela. Izgrađeni model primijenjen je i na probleme semantičke kompozitnosti i semantičke devijantnosti duljih fraza. Rezultati vrednovanja modela potvrdili su da je model, iako jednostavan, sposoban uspješno modelirati semantičko značenje višerječnih izraza u usporedbi s drugim često korištenim modelima kompozicijske distribucijske semantike. Provedena vrednovanja modela ukazala su na različite mogućnosti poboljšanja i proširenja modela.

Ključne riječi: kompozicijska distribucijska semantika, praktični model leksičke funkcije, obrada prirodnog jezika, inkluzija vektora, semantička devijantnost, hrvatski jezik, strojno učenje

Compositional Distributional Semantics based on the Lexical Function Model

Abstract

Compositional distributional semantics deals with vector representations of multiword expressions in high-dimensional vector spaces. Thesis describes some of the most commonly used compositional distributional models, focusing on practical lexical function model and its proposed adaptations. In the thesis, additional adaptations are proposed, based on maximizing the inclusion between pairs of vectors of certain nouns. Practical lexical function model is implemented for Croatian language, as well as evaluated on specially created dataset of phrases containing multiple words. Model is also evaluated on the tasks of semantic compositionality and semantic deviance of longer phrases. Evaluation results have shown that model can successfully model semantic meaning of longer phrases and have also led to some interesting ideas for follow-up work.

Keywords: compositional distributional semantics, practical lexical function model, natural language processing, vector inclusion, semantic deviance, Croatian language, machine learning