TakeLab

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

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 4275

Automatska ekstrakcija semantičkih glagolskih relacija iz korpusa na hrvatskome jeziku

Ivan Sekulić

Zagreb, srpanj 2015.

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

Zagreb, 13. ožujka 2015.

ZAVRŠNI ZADATAK br. 4275

Pristupnik: **Ivan Sekulić (0036472495)**
Studij: Računarstvo
Modul: Računarska znanost

Zadatak: **Automatska ekstrakcija semantičkih glagolskih relacija iz korpusa na hrvatskome jeziku**

Opis zadatka:

Leksičkosemantički jezični resursi nezaobilazni su za semantičku obradu prirodnog jezika i mnoge zadatke u ekstrakciji informacija. Budući da glagoli u tekstu često korespondiraju s predikatnom strukturom teksta, odnosno semantikom događaja, često je potrebno modelirati semantiku glagola odnosno događaja kojima oni odgovaraju. Za takvo je modeliranje vrlo korisna baza semantičkih glagolskih relacija (npr. sličnost, antonimija, omogućavanje), međutim ručna izrada takvog resursa zadovoljavajućeg obima vrlo je zahtjevan posao. Kako bi se zaobišao taj problem, u literaturi je predloženo nekoliko pristupa za statističku ekstrakciju glagolskih relacija iz korpusa.
U okviru završnoga rada potrebno je proučiti postupke za ekstrakciju semantičkih relacija iz korpusa, s naglaskom na postupke za ekstrakciju glagolskih semantičkih relacija te postupke temeljene na sintaktičkim uzorcima. Razraditi model za ekstrakciju glagolskih semantičkih relacija iz korpusa na hrvatskome jeziku, po uzoru na resurs VerbOcean opisan u radu Chklovskog i Pantela (2004). Izgraditi i ručno označiti odgovarajući skup tekstnih podataka na hrvatskome jeziku za razvoj i ispitivanje modela. Razviti programsku implementaciju modela te ga primijeniti na hrvatski web-korpus. Razmotriti prilagodbu i primjenu modela na podatke dobivene internetskom tražilicom. Provesti iscrpno eksperimentalno vrednovanje modela u smislu preciznosti i broja ekstrahiranih relacija, statističku obradu rezultata te analizu pogrešaka. Radu priložiti izvorni i izvršni kod razvijenog sustava, označene skupove podataka i potrebnu dokumentaciju te citirati korištenu literaturu.

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

Mentor:

Doc. dr. sc. Jan Šnajder

Djelovođa:

Doc. dr. sc. Tomislav Hrkać

Predsjednik odbora za završni rad modula:

Prof. dr. sc. Siniša Srbljić

Mami, na neizmjernoj ljubavi i razumijevanju. Mami, kojoj nikad ništa nije teško.

Tati, na ogromnoj podršci tijekom cijelog školovanja. Tati, na motivaciji svojim radom, konstantnim napretkom i pogledom na život.

Bratu, na tome što je moj mlađi brat i što na svoj način pokazuje da je život predivan.

Baki, na pretjeranoj brižljivosti, jajima na špeku i pudingu.

Dedi, na dodatnoj inspiraciji da postanem inženjer.

Prijateljima, starim i novim, koji uspješno održavaju moj socijalni život.

Mentoru, na prilici za učenje uz najbolje i posvećenom vremenu.

Sadržaj

1. Uvod

Prirodni jezik sredstvo je komunikacije ljudi, a podataka spremljenih u obliku teksta i govora sve je više u današnjem svijetu. Za čovjeka je analiza tolikih količina podataka praktički nemoguća, stoga je nužno dio obrade prepustiti računalu. Tu u pomoć uskače obrada prirodnog jezika (engl. *natural language processing, NLP*) – grana računarske znanosti, umjetne inteligencije i računalne lingvistike. Kao jedna od najvažnijih tehnologija modernog informacijskog doba rješava brojne probleme, uključujući strojno prevođenje (engl. *machine translation*), analiza sentimenta (engl. *sentiment analysis*), crpljenje i pretraživanje informacija (engl. *information retrieval, information extraction*) itd.

Leksičkosemantički jezični resursi nezaobilazni su za semantičku obradu prirodnog jezika i mnoge zadatke u ekstrakciji informacija. Budući da glagoli u tekstu često korespondiraju s predikatnom strukturom teksta, odnosno semantikom događaja, često je potrebno modelirati semantiku glagola odnosno događaja kojima oni odgovaraju. Glagol *kupiti* vremenski prethodi glagolu *prodati* te nam taj podatak može ukazati da se neki događaj povezan s kupovanjem određenog objekta dogodio prije prodavanja istog. Mnoga područja obrade prirodnog jezika, uključujući odgovaranje na pitanja (engl. *question answering*), sažimanje teksta (engl. *summarization*) i strojno prevođenje, mogu iskoristiti bazu semantičkih glagolskih relacija (npr. sličnost, antonimija, omogućavanje). Takva baza pomogla bi u rješavanju zadataka vezanih uz događaje ili relacije među entitetima, međutim ručna izrada takvog resursa zadovoljavajućeg obima vrlo je zahtjevan posao. Kako bi se zaobišao taj problem, u nastavku je opisan postupak za statističku ekstrakciju glagolskih relacija iz korpusa.

U okviru završnog rada razrađen je model za ekstrakciju glagolskih semantičkih relacija iz korpusa na hrvatskome jeziku, po uzoru na resurs VerbOcean opisan u radu Chklovski i Pantel (2004). Model je primijenjen na hrvatski web-korpus, a razmotrena je i prilagodba i primjena modela na podatke dobivene internetskom tražilicom. Ručno je označen skup tekstnih podataka na hrvatskome jeziku

te provedeno iscrpno eksperimentalno vrednovanje modela u smislu preciznosti i broja ekstrahiranih relacija, kao i analiza pogrešaka.

U sljedećem poglavlju detaljnije je definiran zadatak te su spomenuti dosadašnji znanstveni doprinosi i različiti pristupi istome. Treće poglavlje opisuje sam postupak ekstrakcije glagolskih semantičkih relacija, koji je podijeljen na dva dijela. Prvi dio bavi se izvlačenjem parova glagola koji su na neki način povezani, dok drugi opisuje traženje konkretne semantičke relacije. Četvrto poglavlje daje kratak uvid u implementacijske probleme i organizaciju cjelokupnog sustava. Eksperimentalno vrednovanje i analiza pogrešaka dani su u petom poglavlju. Zaključak rada nalazi se u šestom poglavlju. U dodatku su prikazani ručno označeni parovi glagola.

2. Semantičke relacije između glagola

2.1. Definicija problema

Prvi korak u rješavanju postavljenog zadatka jest detaljno definirati problem. Potrebno je odrediti što je točno semantička relacija između dva glagola te koliko takvih relacija ima. Model pronalazi četiri semantičke glagolske relacije, koje su, s opisom i primjerima, navedene u nastavku.

Sličnost. Prema Fellbaum (1998), sličnost glagola ne proizlazi samo iz odnosa *to-je* (engl. *is-a*), pa ova skupina obuhvaća više od tako povezanih glagola, ali i više od običnih sinonima. Tako glagoli mogu biti slični iako se razlikuju po stupnju, intenzitetu ili načinu vršenja radnje. Primjeri sličnih glagola ekstrahiranih modelom uključuju: *informirati :: educirati, poboljšati :: unaprijediti, snimiti :: fotografirati.*

Intenzitet. Kao što je gore spomenuto, slični glagoli se mogu razlikovati po intenzitetu, tj. svojoj težini ili jakosti. Ovdje su svrstani parovi glagola kod kojih je drugi glagol po značenju jači od prvoga (npr. *raniti :: ubiti*). Tako povezani glagoli su zapravo podskup skupa sličnih glagola. Neki od intenzitetom povezanih glagola izvučeni algoritmom su: *udvostručiti :: utrostručiti, komentirati :: kritizirati, misliti :: znati.*

Antonimija. Antonimija je leksičko-semantička pojava značenjske opreke između dvaju leksema. Glagoli kao antonimi mogu biti raznokorijenski (primarni, pravi) – ne postoji etimološka veza (npr. *pobijediti :: izgubiti, prihvatiti :: odbaciti*) ili istokorijenski antonimi (tvorbeni) – nastali najčešće prefiksalnim tvorbama riječi (npr. *oružati :: razoružati, imati :: nemati*). Primjeri antonimije ekstrahirani algoritmom: *povećati :: smanjiti, kupiti :: prodati, potvrditi :: demantirati.*

Prethođenje (engl. *happens-before*). Par glagola pripada ovoj relaciji ako se glagoli odvijaju u različitim vremenskim intervalima, tj. ako se radnja koju

opisuje jedan glagol događa prije radnje drugog glagola. Takvi parovi glagola mogu ujedno biti i slični (npr. *diplomirati :: magistrirati*) ili antonimi (npr. *potrgati :: popraviti, kupiti :: prodati*). Primjeri parova glagola izvučeni algoritmom koji pripadaju relaciji *prethođenje*: *napasti :: udariti, diplomirati :: magistrirati, dijagnosticirati :: liječiti*.

Vrijedi spomenuti i relaciju *omogućava* (engl. *enablement*), koju nismo ekstrahirali, ali Chklovski i Pantel (2004) jesu. Barker i Szpakowicz (1995) klasificiraju navedenu relaciju kao kauzalnu. Par glagola pripada ovoj relaciji ako je odvijanje jedne radnje omogućeno odvijanjem druge (npr. *igrati :: pobijediti, pregledati :: ocijeniti*). Relacija *omogućava* direktno označava da je jedan glagol uzrok, a drugi posljedica, što može biti iznimno korisno u analizi događaja. Možemo zaključiti kako je ova relacija specifičan slučaj relacije *prethođenje*, odnosno njen podskup. Neki od glagola povezanih relacijom *prethođenje*, ali ne i relacijom *omogućava* uključuju: *oteti :: ubiti, udariti :: pobjeći*.

Jedan par glagola može istovremeno biti u više semantičkih relacija. Glagoli mogu biti slični ili suprotni, a da uz to vremenski prethode jedan drugome. Također, kao što je već spomenuto, glagoli mogu biti slični, ali različiti po intenzitetu. Primjeri parova glagola za koje je pronađeno više relacija su: *diplomirati :: magistrirati – sličnost, prethođenje, vjerovati :: znati – antonimija, prethođenje*.

Tablica 2.1: Primjeri parova glagola za pojedinu relaciju.

sličnost	informirati :: educirati
	poboljšati :: unaprijediti
	snimiti :: fotografirati
intenzitet	udvostručiti :: utrostručiti
	komentirati :: kritizirati
	misliti :: znati
antonimija	povećati :: smanjiti
	kupiti :: prodati
	potvrditi :: demantirati
prethođenje	napasti :: udariti
	diplomirati :: magistrirati
	dijagnosticirati :: liječiti

2.2. Srodni radovi

Tijekom razvijanja područja obrade prirodnog jezika bilo je nekoliko pokušaja grupiranja glagola i označavanja njihovih tematskih uloga. Vjerojatno najveći resurs leksika engleskog jezika jest WordNet (Miller, 1995). WordNet povezuje imenice, glagole i pridjeve u 117 000 međusobno povezanih skupova kognitivnih sinonima. Od glagolskih relacija identificira sinonime, antonime, troponime (engl. *troponymy*) i omogućavanje. EVCA[1] (Levin, 1993) organizira glagole po sličnosti i pojavljivanju / nepojavljivanju u određenim uzorcima, a sadrži 3200 glagola klasificiranih u 191 klasu. Ručno izrađeni resursi uključuju PropBank (Palmer et al., 2005), FrameNet (Baker et al., 1998), VerbNet (Kipper et al., 2000). Traženje semantičkih relacija između imenica opisuju Hearst (1992), Etzioni et al. (2005), Ravichandran i Hovy (2002).

Pristupi automatskoj ekstrakciji semantičkih relacija temeljeni su uglavnom ili na uzorcima ili na grupiranju (engl. *clustering*). Hearst (1992) je pionir u primjeni uzoraka te njegov pristup usvajaju i nadograđuju mnogi drugi. Berland i Charniak (1999) proširuju njegov pristup hvatajući ne samo *to-je* (engl. *is-a*), već i *dio* (engl. *part-of*) relacije. Daljnja poboljšanja uvodi Girju et al. (2006) kombinirajući metode strojnog učenja s WordNet-om. Pantel i Pennacchiotti (2006) primjenjuju automatski generirane generičke uzorke na ekstrakciju semantičkih relacija u tekstu. Prvi pokušaj organizacije riječi u klase napravio je Caraballo i Charniak (1999) koristeći veznike i apozicijske značajke kako bi izgradio klase imenica. Klasifikaciju glagola rade Kawahara et al. koristeći nenadzirano strojno učenje. U referentnom radu (Chklovski i Pantel, 2004) ekstrahiraju semantičke glagolske relacije koristeći jednostavne, ručno izrađene leksičkosintaktičke uzorke. Uzorke popunjavaju povezanim glagolima, dobivenim algoritmom DIRT (Lin i Pantel, 2001), te ih šalju kao upite na internetsku tražilicu. Ovaj pristup je, uz nekoliko modifikacija i prilagodbu na hrvatski jezik, usvojen u ovom radu.

[1] English Verb Classes and Alternations, Levin (1993)

3. Postupak ekstrakcije semantičkih relacija između glagola

Cjelokupni postupak podijeljen je na dva dijela: ekstrakcija parova vrlo povezanih glagola i traženje konkretnih semantičkih relacija između tako generiranih parova glagola. Prvi dio potreban je kako ne bismo morali provjeravati semantičku povezanost svih kombinacija glagola u pojedinoj rečenici. Ovakav pristup omogućuje i prilagodbu modela na podatke dobivene internetskom tražilicom, kao što je opisano u Chklovski i Pantel (2004). Sama baza povezanih glagola može biti korisna ne samo za određivanje semantičkih relacija, već i za druge zadatke obrade prirodnog jezika. Postupak izgradnje takve baze opisan je u odjeljku 3.2, dok je sama ekstrakcija relacija opisana u 3.3.

3.1. Opis i priprema korpusa

Korišteni korpus je hrvatski web-korpus HrWaC (Ljubešić i Erjavec, 2011), odnosno njegova filtrirana parsana verzija fHrWaC-parsed (Šnajder et al., 2013). Korpus sadrži 50,940,598 ovisnosno parsanih rečenica, od kojih ih je u prvom koraku algoritma korišteno 20%. Parsan korpus zapisan je u formatu CoNLL[1], gdje svaka linija predstavlja jednu riječ s nekoliko oznaka odvojenih tabulatorom, a rečenice su međusobno odvojene praznim retkom. Primjer rečenice zapisane u formatu CoNLL dan je u tablici 3.1. Za ovaj rad relevantne oznake riječi su:

- ID: indeks u rečenici, počinje od 1
- FORM: pojavnica
- LEMMA: lematizirana (engl. *lemmatization*) riječ

[1] Conference on Natural Language Learning

– POS: oznaka vrste riječi (engl. *part-of-speech tag, POS tag*)

– HEAD: indeks roditelja, 0 za korijen

– DEPREL: sintaktička relacija s HEAD riječi

Tablica 3.1: Rečenica "Klijent je suprugu obavijestio da ide na kraći službeni put" parsana i zapisana u formatu CoNLL

ID	FORM	LEMMA	POS	PPOS	FEAT	HEAD	DEPREL
1	Klijent	klijent	N	N-msm	–	4	Sb
2	je	biti	V	Vcr3s	–	4	Aux
3	suprugu	supruga	N	N-fsa	–	4	Obj
4	obavijestio	obavijestiti	V	Vmp-sm	–	0	Pred
5	da	da	C	Cs	–	4	Sub
6	ide	ići	V	Vmr3s	–	5	Pred
7	na	na	S	Sa	–	6	Prep
8	kraći	kraći	V	Vmn	–	7	Atv
9	službeni	služben	A	Agpmsa	–	10	Atr
10	put	put	N	N-msan	–	8	Obj
11	.	.	Z	Z	–	0	Punc

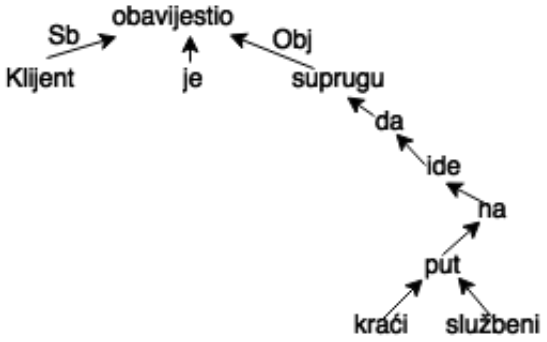
U čitavom modelu korišten je samo lematizirani oblik riječi. Već pri prvom prolasku u nastavku opisanog algoritma primijećeno je izrazito puno šuma, koji nastaje što zbog neispravno napisane izvorne rečenice, a što zbog loše lematizirane riječi. Javila se potreba za izradom rječnika "dobrih" riječi, točnije "dobrih" glagola i "dobrih" imenica. U rječnik "dobrih" glagola dodani su svi glagoli koji se pojavljuju barem tisuću puta u korpusu i završavaju na *ti* ili *ći*. Također su, radi smanjenja šuma, izbačeni glagoli biti, htjeti, moći i morati. Navedenim postupcima dobiven je rječnik s 4997 glagola, što je naoko zadovoljavajuć broj glagola. Za izradu rječnika "dobrih" imenica prebrojan je broj pojavljivanja imenica koje se pojavljuju kao subjekt ili objekt u 12% korpusa. Filtrirane su riječi koje se pojavljuju manje od 10 puta te je dobiven rječnik s 80218 imenica. Procesuirano je i 25% korpusa, ali se tu pojavljivalo puno više šuma, tj. previše nepostojećih imenica. Nakon kratkog eksperimentiranja odlučeno je uzeti imenice s frekvencijom pojavljivanja većom od 10 u 12% korpusa.

3.2. Ekstrakcija semantički povezanih glagola

Mnogo algoritama za traženje sličnih riječi bazira se na principu poznatom kao distribucijska hipoteza (Harris, 1954). Hipoteza tvrdi da riječi koji se pojavljuju u istom kontekstu teže semantičkoj povezanosti. U ovom radu koristi se algoritam DIRT[2], čija osnovna ideja također proizlazi iz navedene hipoteze. Lin i Pantel (2001) primjenjuju distribucijsku hipotezu, umjesto klasičnog pristupa na riječi, na puteve u ovisnosno parsanim rečenicama (engl. *dependency trees*). Ako dva puta u više rečenica povezuju iste riječi, pretpostavlja se da je njihovo značenje slično. U nastavku opisani postupak koristi se za ekstrakciju povezanih glagola, no govorit će se o putevima. U ovome radu jedan glagol odgovara jednome putu, ali ovakva generalizacija omogućava primjenu algoritma ne samo na glagole, već i za traženje parafraza i drugih povezanih sintaktičkih nizova.

3.2.1. Ekstrakcija puteva u ovisnosno parsanim rečenicama

Ovisnosni odnos (Hays, 1964) je asimetrična binarna relacija između riječi *glava* (engl. *head*) i druge riječi *modifikator* (engl. *modifier*). Struktura rečenice može se prikazati kao niz takvih relacija koje oblikuju stablo (engl. *dependency tree*). Svaka riječ može imati više modifikatora, ali može biti modifikator isključivo jednoj riječi. Drugačije rečeno, svaka riječ pamti samo svog roditelja, a ne djecu kao što je slučaj u standardnom pristupu konstruiranja stabla, te oznaku relacije koja opisuje njihov odnos. Slika 3.1 predstavlja primjer ovisnosnog stabla.



Slika 3.1: Ovisnosno stablo za rečenicu "Klijent je obavijestio suprugu da ide na kraći službeni put".

[2]Discovery of Inference Rules from Text, Lin i Pantel (2001)

Tablica 3.2: X obavijestiti Y

Mjesto X			Mjesto Y		
policija	68	0.97	policija	265	2.33
liječnik	34	1.81	javnost	144	3.99
on	32	-1.21	mi	76	0.71
nitko	30	1.73	on	56	-0.65
centar	30	2.25	oni	48	0.09
sud	20	0.86	burza	46	4.16
organizator	18	1.97	medij	40	2.17
sav	16	0.06	vi	32	1.02
odvjetnik	16	2.29	predsjednik	32	2.02
građanin	16	1.09	stanar	24	3.24

Put u stablu definiran je kao spoj riječi i relacija između njih, pod uvjetom da su krajnje riječi na putu imenice. Lin i Pantel (2001) u put uvrštavaju sve riječi koje nose neki sadržaj, tj. imenice, glagole, pridjeve i priloge. Njihov algoritam iz priložene rečenice izvlači dva puta: između *klijent* i *supruga* te između *klijent* i *put*. Krajnje lijeva riječ popunjava mjesto (engl. *slot*) X, a krajnje desna riječ mjesto Y. Opći oblik tako definiranog puta je: $\mathrm{X} \leftarrow [\text{riječi}] \leftarrow \text{korijen} \rightarrow [\text{riječi}] \rightarrow \mathrm{Y}$.

U ovom radu izvlače se samo relacije $\text{subjekt} \leftarrow \text{glagol} \rightarrow \text{objekt}$, pa se put pretvara u oblik $\mathrm{X} \leftarrow \text{glagol} \rightarrow \mathrm{Y}$. Iz priložene rečenice ekstrahira se samo put $\mathrm{X:Sb:V} \leftarrow \text{obavijestiti} \rightarrow \mathrm{V:Obj:Y} \equiv$ "X obavještava Y", gdje riječ *klijent* popunjava mjesto X, a *supruga* mjesto Y. Za svaki put, tj. glagol, pamte se imenice koje popunjavaju mjesto X i one koje popunjavaju mjesto Y. Kako bi mogli izračunati sličnost puteva, potrebno je pratiti frekvenciju pojavljivanja tih imenica i za svaku izračunati mjeru uzajamne informacije (engl. *mutual information*). Mjera uzajamne informacije nam upućuje na snagu povezanosti imenice i mjesta koje popunjava, a detaljnije je objašnjena u sljedećem pododjeljku.

Sve informacije o putu spremamo u bazu podataka organiziranu kao baza trojki (engl. *triple database*). Možemo ju zamisliti kao niz uređenih trojki oblika (put, slot, riječ). Primjer elementa baze podataka dan je u tablici 3.2. Za glagol *obavijestiti* prikazano je po deset najčešćih imenica koje popunjavaju mjesta X i Y, njihova frekvencija te mjera uzajamne informacije. Primjećujemo kako se imenice koje popunjavaju mjesto Y pojavljuju više puta od onih iz mjesta X. Razlog tome jest svojstvo jezika da jedan glagol, tj. predikat u rečenici, uglavnom ima samo

jedan subjekt, ali može imati više objekata. Baza sadrži 4868 puteva, odnosno glagola.

3.2.2. Sličnost između dva puta

Nakon što smo kreirali bazu puteva, možemo izračunati sličnost bilo koja dva puta. Intuitivno zaključujemo da su dva puta sličnija što više značajki dijele. Također je intuitivno jasno da sve značajke nisu jednako važne. Npr. *on* je vrlo česta zamjenica te se pojavljuje u velikoj većini puteva kao subjekt, dok imenica *odvjetnik* i nije tako česta. Značajka *(slotX, on)* manje ukazuje na sličnost dvaju puteva nego značajka *(slotX, odvjetnik)*. Zbog toga uvodimo mjeru uzajamne informacije (engl. *mutual information*) između značajke i puta.

Notacija $|p, slotX, w|$ označava frekvenciju pojavljivanja trojke $(p, slotX, w)$, gdje je p put, $slotX$ mjesto X, a w riječ. $|p, slotX, *|$ označava $\sum_w |p, slotX, w|$, a $|*, *, *|$ označava $\sum_{p,s,w} |p, s, w|$. $T(p_i, s)$ je skup riječi koji popunjavaju mjesto s puta p_i.

Značajka će biti važnija što je mjera uzajamne informacije veća. Mjera za pojedinu riječ w iz $T(p_i, s)$ bit će veća što je broj pojavljivanja te riječi veći u $T(p_i, s)$, a manji ukupno u svim putevima. Također, mjera je veća što je broj riječi u $T(p_i, s)$ manji. Time umanjujemo važnost značajke kod jako čestih puteva koji povezuju velik broj riječi.

Mjeru uzajamne informacije za svaku riječ računamo prema (Chklovski i Pantel, 2004):

$$mi(p, slot, w) = \ln \left(\frac{|p, slot, w| \times |*, slot, *|}{|p, slot, *| \times |*, slot, w|} \right) \tag{3.1}$$

Sličnost između dva puta p_1 i p_2 definirana je kao geometrijska sredina sličnosti njihovih X mjesta i Y mjesta:

$$psim(p_1, p_2) = \sqrt{ssim(slotX_1, slotX_2) \times ssim(slotY_1, slotY_2)} \tag{3.2}$$

Sličnost mjesta s_1 i s_2 definirana je formulom:

$$ssim\,(s_1, s_2) = \frac{\sum_{w \in T(p_1,s) \cap T(p_2,s)} mi\,(p_1, s, w) + mi\,(p_2, s, w)}{\sum_{w \in T(p_1,s)} mi\,(p_1, s, w) + \sum_{w \in T(p_2,s)} mi\,(p_2, s, w)} \tag{3.3}$$

Dva puta su sličnija što je njihov *psim* veći. Definirali smo sve potrebno za izračun sličnosti između dva puta, ali postavlja se pitanje između kojih puteva računati sličnost. S obzirom na to da imamo gotovo 5000 puteva, računanje sličnosti za svaki mogući par je praktički nemoguće. Prvo možemo zaključiti

kako nema smisla ispitivati parove koji ne dijele niti jednu značajku. Tijekom generiranja puteva je uz svaku riječ koja popunjava određeno mjesto spreman skup puteva koji ju dijele.

Algoritam traženja sličnih puteva nekom putu p provodimo u tri koraka:

1. Za neki put p uzimamo u skup S sve kandidate k koji dijele barem jednu značajku, kao što je ranije opisano.

2. Za svaki kandidat k prebrojimo dijeljene značajke. Filtriramo S tako da izbacimo sve k koji s p dijele manje od 1% ukupnog broja značajki p i k. Izračun sličnosti na takav način je manje zahtjevan od izračuna prema formuli 3.2, a putevi koji ne dijele barem 1% značajki ionako ne bi bili označeni kao slični.

3. Za svaki preostali k iz S izračunamo sličnost prema formuli 3.2. Pamtimo 20 najsličnijih k te njihove mjere sličnosti s p.

3.2.3. Izlaz DIRT-a

U prošlom je pododjeljku opisan algoritam izvlačenja sličnih puteva. Budući da se nama putevi sastoje isključivo od jednog jedinog glagola, treba definirati što zapravo znači kada su oni slični. Algoritmom DIRT ekstrahiraju se vrlo povezani glagoli (Chklovski i Pantel, 2004). Za takve glagole pretpostavlja se postojanje semantičke relacije te se samo oni uzimaju u obzir u postupku traženja semantički povezanih glagola, opisanom u odjeljku 3.3. Tablica 3.3 prikazuje semantički povezane glagole s glagolom *voziti*, ekstrahirane algoritmom opisanim u pododjeljku 3.2.2.

Gledajući tablicu 3.3 zaključujemo kako pretpostavka da se DIRT algoritmom izvlače semantički povezani glagoli ima smisla. Vidimo sve četiri relacije opisane u odjeljku 2.1:

- sličnost: *voziti :: upravljati*
- intenzitet: *voziti :: juriti*
- prethođenje: *voziti :: sudariti*
- antonimija: *voziti :: zaustavljati*

Potencijalno semantički povezane glagole prosljeđujemo algoritmu za traženje semantičkih relacija u nadi da će ih pronaći što više, uključujući i ove koje smo

Tablica 3.3: Semantički povezani glagoli s glagolom *voziti*.

juriti	parkirati
istrčati	izletiti
prepriječiti	naletiti
upaliti	sudariti
upravljati	sjesti
odvoziti	letiti
približavati	prevrnuti
dovezati	odvezati
prevoziti	zaustavljati
skrenuti	opkoliti

mi s lakoćom pronašli u tablici 3.3. Detaljna analiza i evaluacija rezultata nalazi se u poglavlju 5.

3.3. Ekstrakcija semantičkih relacija

Model za ekstrakciju glagolskih semantičkih relacija napravljen je po uzoru na resurs VerbOcean, uz neke modifikacije. Za svaku semantičku relaciju opisanu u odjeljku 2.1 definirano je nekoliko leksičkosintaktičkih uzoraka, opisanih u pododjeljku 3.3.1. U referentnome radu, relacija se otkriva upitima uzoraka indikativnih za tu relaciju na web, točnije na Googleovu tražilicu, dok se ovdje uzorci traže u rečenicama korpusa fHrWaC. Prvo je potrebno definirati sintaktičke uzorke te kako ih primjenjivati na rečenicu, što je napravljeno u pododjeljku 3.3.1. Zatim je u pododjeljku 3.3.2 opisan postupak odabira ispravnih semantičkih relacija, ako je za isti par glagola pronađeno više, što je vrlo čest slučaj.

3.3.1. Leksičkosintaktički uzorci

Definirano je 16 leksičkosintaktičkih uzoraka, od kojih svaki ukazuje na određenu semantičku relaciju. Uzorci su navedeni u tablici 3.4. Mjesta X i Y popunjavaju se parovima vrlo povezanih glagola, odnosno onima koje je izbacio algoritam DIRT. Na mjestu asteriska "*" može se pojaviti proizvoljan broj riječi. Potreba uvođenja asterisk simbola proizlazi iz same prirode jezika. Promatrajući složenije sintaktičke nizove i njihov položaj u stvarnim rečenicama, zaključujemo kako se između glagola i niza može pojaviti mnoštvo drugih riječi, a da glagoli ipak budu

semantički povezani. Kao primjer uzmimo rečenicu "Vozio je vrlo neodgovorno i na kraju se sudario". Glagoli *voziti* i *sudariti* semantički su povezani relacijom *prethođenje* koju bez asteriska ne bismo uhvatili. U uzorcima koji se sastoje od samo jedne riječi između para glagola ne dopuštamo pojavljivanje drugih riječi. Uvođenje asteriska kod takvih uzoraka izazvalo bi previše šuma, tj. njihovo pojavljivanje u rečenicama bilo bi prečesto što bi rezultiralo lažno pozitivnim (engl. *false positive*) relacijama.

Uzorak primjenjujemo na rečenicu u obliku regularnog izraza tako da glagole iz rečenice stavimo na mjesta X i Y u uzorku. Kao što je već napomenuto, ispitivat ćemo samo one parove glagola koje smatramo povezanima, pa prvo moramo provjeriti prisutnost para u ranije generiranom skupu semantički povezanih glagola. Ako se par glagola nalazi u skupu, isprobavamo sintaktičke uzorke na rečenici iz koje smo izvukli par. Pronalazak uzorka u rečenici znači postojanje one semantičke relacije na koju taj uzorak ukazuje. Za svaki par glagola pamte se semantičke relacije i njihov broj pojavljivanja. Diskusija o tome je li frekvencija pojavljivanja dovoljan indikator semantičke relacije dana je u odjeljku 5.3.

Tablica 3.4: Semantičke relacije i uzorci koji ukazuju na njih. Broj uzoraka za svaku relaciju prikazan je u zagradama.

sličnost (2)	antonimija (2)	prethođenje (7)	intenzitet (5)
X tj. Y	X ili Y	X * i onda * Y	ne samo * X * već i * Y
X i Y	X * nego * Y	X * zatim * Y	X * čak i * Y
		X pa Y	X * ili barem * Y
		X * i na kraju * Y	Y * ili bar * X
		X * i poslije * Y	X * a čak * Y
		X * i kasnije * Y	
		Y * nakon što * X	

3.3.2. Podrezivanje identificiranih semantičkih relacija

Za jedan par glagola možemo pronaći više semantičkih relacija te se javlja jedan od sljedećih slučajeva (Chklovski i Pantel, 2004):

1. jedna relacija je specifičnija od druge (*intenzitet* je specifičniji od *sličnost*)
2. postojanje jedne relacije ne isključuje postojanje druge relacije (npr. *antonimija* i *prethođenje*)

Tablica 3.5: Parovi glagola, njihove semantičke relacije i njihov broj ponavljanja.

Par glagola	Relacija	Frekvencija
producirati :: režirati	sličnost	28
voziti :: udariti	prethođenje	24
napasti :: udariti	prethođenje	32
	sličnost	16
vjerovati :: znati	antonimija	28
	prethođenje	18
diplomirati :: magistrirati	sličnost	36
	prethođenje	12

3. relacije su međusobno isključive (*sličnost* i *antonimija*)

Vrlo često ekstrahira se, za jedan par glagola, više relacija s različitim brojem pojavljivanja. Odluku koju relaciju zadržati, a koju ignorirati definirali smo s nekoliko jednostavnih pravila. Pravila proizlaze iz semantike relacija koje smo definirali u odjeljku 2.1.

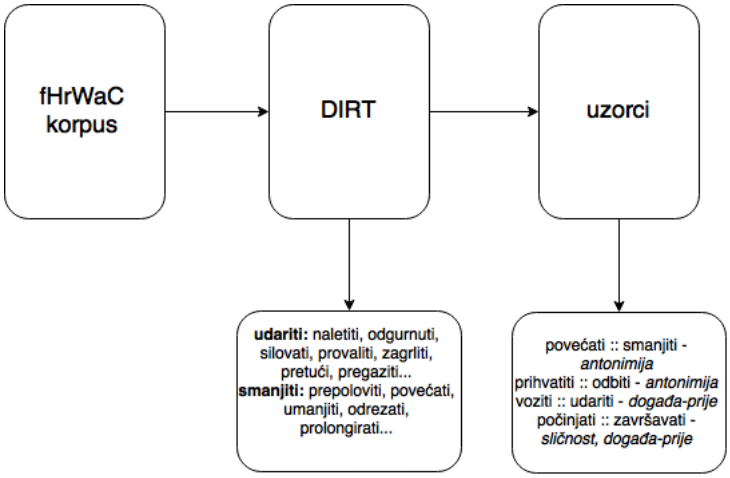
Prisutnost *prethođenje* relacije nam nikada ne smeta te nju nikada ne ignoriramo. Ignoriramo relaciju *sličnost* ako je detektirana relacija *intenzitet* (jer je *intenzitet* specifičnija od *sličnost*). Od preostalih relacija (*sličnost*, *intenzitet* i *antonimija*) uzimamo onu s najviše pojavljivanja, dok ostale ignoriramo. Primjenom navedenih pravila, za svaki par glagola dobivamo najviše dvije relacije. Primjer konačnog izlaza iz opisanog algoritma dan je u tablici 3.5.

Pregled rezultata ukazao je na velik broj lažno pozitivnih relacija te se javila potreba za dodatnim filtriranjem. Dodatno filtriranje napravljeno je nakon ručnog označavanja semantičkih glagolskih relacija. Iz 150 označenih parova glagola izračunat je prosječan broj ponavljanja lažno pozitivnih primjera pojedine relacije (*sličnost* 3, *antonimija* 7, *prethođenje* 6, *intenzitet* 6). Iz skupa su izbačene sve relacije čija je frekvencija pojavljivanja, za određeni par glagola, manja od prosječne frekvencije za tu relaciju. Time je porasla preciznost cijelog modela, uz cijenu pada odziva. Detaljna analiza dana je u poglavlju 5.

3.4. Povezivanje komponenti

Cjelokupni postupak ekstrakcije semantičkih relacija između glagola, podijeljen na komponente *DIRT* i *uzorci*, prikazan je na slici 3.2. Ulaz u algoritam DIRT

je parsan fHrWaC korpus, a izlaz, detaljnije opisan u pododjeljku 3.2.3, potencijalno semantički povezani glagoli. Izlaz DIRT-a je, uz korpus, ulaz nužan za traženje konkretne semantičke relacije sintaktičkim uzorcima. Algoritam DIRT proveden je na 20% korpusa, dok je traženje konkretne relacije provedeno na čitavom korpusu. Konačni izlaz sustava je skup parova glagola i semantičkih relacija koje ih vežu 3.5.



Slika 3.2: Cjelokupni postupak s primjerima izlaza pojedinih komponenti.

4. Implementacija

4.1. Korišteni alati

Za razvoj cjelokupnog sustava izabran je programski jezik Python, verzija 2.7. Razlog tome jest njegova jednostavnost, otvorenost koda (engl. *open source*) te osobno pozitivno iskustvo. Izbor se pokazao odličnim jer tijekom izrade sustava nije bilo potrebno obraćati posebnu pozornost na sam jezik. Nije bilo potrebno razmišljati o sintaksi, bibliotekama i konstrukciji jezika, već se moglo najveći dio vremena usredotočiti na sam zadatak i programsku logiku. Razvoj je bio brz, a modifikacije koda jednostavne, što je omogućilo učinkovito ispitivanje u svakom koraku.

Od esencijalne važnosti za obradu teksta na hrvatskom jeziku jest podrška hrvatskih grafema. Sve operacije nad tekstovnim podatcima obavljaju se u *Unicode* standardu, koji zadovoljava naše potrebe. Pythonova biblioteka *codecs* omogućuje kodiranje znakova standardom *UTF-8*, odnosno zapisivanje znakova u *Unicode*. Valja spomenuti i biblioteku *pickle*, koja omogućava serijalizaciju objekata. Za čuvanje svih podataka korišteni su Pythonovi rječnici (engl. *dictionary*), koji su pomoću navedene biblioteke zapisani u datoteke.

Sustav je razvijan na prijenosnom računalu Dell Inspiron 5720 s:

- 6 GB radne memorije
- Intel Core i5-3210M CPU @ 2.50GHz
- Linux OS - Ubuntu 14.04

4.2. Složenost sustava

Iznimno velika količina podataka, čak 42.6 GB parsanog korpusa, zahtijevala je promišljen pristup obradi. Kao što je spomenuto u odjeljku 3.1 napravljen je rječnik "dobrih" imenica i rječnik "dobrih" glagola. Svakoj riječi u rječniku

pridružena je cjelobrojna vrijednost koja u svim koracima obrade sustava reprezentira tu riječ. Time je značajno smanjena memorijska složenost cjelokupnog sustava i omogućena obrada većeg udjela korpusa. Vremena izvođenja pojedinih komponenti su reda veličine nekoliko sati, što je očekivano s obzirom na količinu podataka. Unatoč višestrukom refaktoriranju koda, cijeli korpus nije mogao biti obrađen algoritmom DIRT. Nedostatak radne memorije računala mogao se riješiti kreiranjem baze podataka na disku, ali to bi dodatno usporilo rad sustava. Druga opcija je korištenje jačeg računala s više radne memorije.

5. Eksperimentalno vrednovanje

5.1. Priprema skupa za testiranje

Vrednovanje sustava zahtijevalo je ručno označene semantičke glagolske relacije. Slučajnim odabirom izabrano je 500 parova glagola generiranih DIRT algoritmom i 500 parova iz izlaza cjelokupnog sustava. Prvi dio skupa iskorišten je za evaluaciju odziva, a drugi za evaluaciju preciznosti sustava (detaljnije objašnjeno u sljedećem odjeljku). Dva nezavisna sudca označila su semantičke relacije svakog para glagola. Oznake za pojedinu relaciju su: *s – sličnost, i – intenzitet, h – prethođenje, a –antonimija, n – nema relacije.* Iako par glagola može imati više relacija, označavana je samo jedna relacija, ona najistaknutija. Oznakom *n* označeni su ne samo parovi glagola koji nemaju semantičku relaciju, već i svi parovi glagola od kojih barem jedan nije dobro lematiziran te ne pripada hrvatskome jeziku. Iako je korišten rječnik dobrih glagola, neke česte pogreške u parsanju (ili pravopisu) rezultirale su pojavom zapravo nepostojećih glagola ili čak imenica (npr. *bavititi, gosti*). Broj takvih riječi znatno je veći u DIRT skupu, nego u konačnom izlazu. Razlog tome vjerojatno su ispravnije napisane rečenice u kojima pronalazimo određeni uzorak.

U uzorak zlatnog standarda (engl. *gold standard sample*) uzeti su parovi glagola za koje su oba sudca označila istu relaciju ili nepostojanje iste. Tako je dobiven skup od 454 parova generiranih DIRT algoritmom i skup od 332 para iz konačnog izlaza sustava. Kompletna lista označenih parova nalazi se u dodatku A. Razlog ovako malog broja glagola u drugome skupu možda je već spomenuto označavanje samo jedne relacije. Moguće je da su sudci označili različitu relacije, a da su oboje bili u pravu, ali to onda nije uzeto u presjeku. Za precizniju evaluaciju potrebno je označiti više parova glagola te dopustiti označavanje više relacija za pojedini par. Skup za vrednovanje preciznosti podijeljen je na dva dijela (150 u prvom te 182 u drugom). S prvim dijelom obavljeno je dodatno filtriranje, opisano u odjeljku 3.3.2. Nad drugim dijelom provedeno je vrednovanje

Tablica 5.1: Matrica zabune DIRT skupa.

		Zlatni standard					
		s	i	h	a	n	
Predviđeno	s	6	0	0	0	2	8
	i	0	0	0	0	0	0
	h	0	0	0	0	1	1
	a	0	0	0	0	0	0
	n	17	0	4	5	420	446
		23	0	4	5	423	

preciznosti, čiji su rezultati u sljedećem odjeljku.

5.2. Evaluacijske mjere

Eksperimentalno vrednovanje sustava provedeno je u smislu preciznosti (engl. *precission*), odziva (engl. *recall*) i F-mjere (engl. *F score*) (Van Rijsbergen, 1974). Za izračun navedenih mjera potrebna nam je matrica zabune (engl. *confusion matrix*) – tablica kojom vizualiziramo performanse određenog algoritma. Redci matrice reprezentiraju klase predviđene od strane sustava, a stupci stvarne klase iz zlatnog standarda. Kod višeklasne klasifikacije, kakvu imamo, za pojedinu klasu definiramo:

- TP_i – ispravno pozitivni (engl. *true positive*), i-ti element dijagonale
- FP_i – lažno pozitivni (engl. *false positive*), zbroj nedijagonalnih elemenata i-tog retka
- FN_i – lažno negativni (engl. *false negative*), zbroj nedijagonalnih elemenata i-tog stupca
- TN_i – ispravno negativni (engl. *true negative*), zbroj po elementima izvan retka i i stupca i (minora)

Preciznost, odziv i F-mjeru računamo za svaku klasu zasebno, a zatim uzmemo prosjek kako bi dobili mjeru za cijeli sustav. Preciznost definiramo kao udio točno klasificiranih primjera u skupu pozitivno klasificiranih primjera, a računamo formulom:

$$P = \frac{TP}{TP + FP} \tag{5.1}$$

Odziv je udio točno klasificiranih primjera u skupu svih pozitivnih primjera:

$$R = \frac{TP}{TP + FN} \tag{5.2}$$

F-mjera je harmonijska sredina preciznosti i odziva. U općem slučaju važnost preciznosti i odziva kontroliramo parametrom β, ali ovdje koristimo F_1 mjeru, gdje je $\beta = 1$:

$$F_1 = \frac{2PR}{P + R} \tag{5.3}$$

Tablica 5.1 prikazuje matricu zabune DIRT skupa označenih parova, pomoću koje računamo odziv sustava. Na ovaj način zapravo računamo odziv drugog dijela algoritma, u odnosu na izlaz iz DIRT-a. Do pravog odziva cjelokupnog modela teško je doći, jer bi morali označavati parove glagola iz cijelog hrvatskog jezika, kojih naravno ima jako puno te je vjerojatnost postojanja semantičke relacije između njih vrlo mala. U matrici zabune vidimo kako većina parova glagola nema semantičku relaciju te da ona nije pronađena ni našim modelom. Ovaj podatak ukazuje na loš rad DIRT algoritma. Pretpostavka je bila da ćemo njime dobiti parove glagola koji su na neki način semantički povezani, ali 93.17% parova klasificiranih kao nepovezani ne potvrđuje tu pretpostavku. Pošto računamo odziv cjelokupnog sustava, klasu *nema relacije* ne uzimamo u obzir. Dobivamo relativno mali odziv od 22.22%. U uzorku je premalo parova glagola koji imaju semantičku relaciju pa možemo zaključiti kako rezultat nije reprezentativan te je potrebno označiti više parova.

Preciznost, odziv i F-mjera čitavog sustava prije i poslije dodatnog podrezivanja dani su u tablici 5.4. Podrezivanje je izvedeno tražeći najveću preciznost i odziv nad 150 parova glagola. Nakon optimiranja rezultata na 150 parova glagola, evaluirano je ostalih 182 para iz zlatnog standarda. Tablica 5.2 prikazuje matricu zabune označenog skupa konačnog izlaza za testiranje (182 para) prije podrezivanja, a tablica 5.3 nakon. Preciznost, odziv i F-mjera prikazani su u tablici 5.5. Iz tablice je vidljivo značajno povećanje preciznosti te neznatni pad odziva, što je i očekivano.

Rezultati jasno pokazuju kako je dodatno filtriranje po frekvenciji pojedine relacije povećalo preciznost modela, koja na kraju za testni skup konačnog izlaza iznosi 52.6%. Referentni VerbOcean ima točnost od 65.5% što također nije predobro. U konačnici je ekstrahirano 1663 para semantički povezanih glagola. Zaključujemo kako model ipak nešto radi, ali mu je potrebno niz poboljšanja.

Tablica 5.2: Matrica zabune testnog skupa prije podrezivanja.

		Zlatni standard					
		s	i	h	a	n	
Predviđeno	s	41	1	5	5	24	76
	i	3	0	0	0	3	6
	h	14	1	8	2	25	50
	a	21	0	1	14	41	77
	n	0	0	0	0	0	0
		155	2	30	32	168	

Tablica 5.3: Matrica zabune testnog skupa nakon podrezivanja.

		Zlatni standard					
		s	i	h	a	n	
Predviđeno	s	25	1	4	2	8	40
	i	0	0	0	0	1	1
	h	4	0	3	3	3	13
	a	3	0	1	5	6	15
	n	36	1	6	12	65	120
		68	2	14	22	83	

Tablica 5.4: Preciznost, odziv i F-mjera skupa za podrezivanje (150 parova).

	Prije podrezivanja	Nakon podrezivanja
P	0.304	0.48
R	0.504	0.471
F_1	0.379	0.475

Tablica 5.5: Preciznost, odziv i F-mjera testnog skupa (182 para).

	Prije podrezivanja	Nakon podrezivanja
P	0.301	0.526
R	0.543	0.526
F_1	0.387	0.526

Tablica 5.6: Preciznost, odziv i F_1 pojedine klase.

	sličnost	intenzitet	prethođenje	antonimija
P	0.625	0.0	0.3	0.333
R	0.367	0.0	0.214	0.263
F_1	0.463	0.0	0.25	0.294

5.3. Analiza pogrešaka i moguće nadogradnje

U ovom odjeljku predstavljeni su propusti u izradi modela te dani prijedlozi za daljnji rad i poboljšanje pojedinih komponenti sustava. Iako je mnogo pogrešaka uočeno i ispravljeno tijekom konstrukcije modela, tek je analiza konačnih rezultata ukazala na neke od mogućih poboljšanja. Naknadne izmjene modela, u okviru završnog rada, zbog nedostatka vremena nažalost nisu bile moguće. Prvo su predložene izmjene algoritma DIRT, a zatim algoritma za ekstrakciju semantičkih relacija te je na kraju odjeljka razmotren model u cjelini.

Rezultate algoritma DIRT teško je vrednovati, ali postoje indicije kako ga poboljšati. Parametar koji je najlakše promijeniti jest broj povezanih glagola sa svakim glagolom. Uz pojedini glagol vežemo 20 drugih, što je naoko, bez detaljne analize, izgledalo kao zadovoljavajuć broj. Uvjet da za par glagola uopće provjerimo postoji li semantička relacija jest da se on nalazi u skupu vrlo povezanih glagola, tj. izlazu DIRT-a. Nameće se pitanje za koliko parova postoji određena relacija, a nije pronađena ne zbog loše definiranih uzoraka ili nekog drugog propusta, već zbog toga što glagoli nisu niti označeni kao povezani. Povećanjem veličine grozda moguće je u konačnici ekstrahirati više semantičkih relacija, odnosno povećati odziv kompletnog sustava. Smanjenje veličine grozda može povećati preciznost modela, ali vjerojatno nije isplativo zbog velikog smanjenja odziva modela. Potrebno je evaluirati rezultate konačnog izlaza s različitim veličinama grozdova i zatim odabrati optimalnu.

Prilikom ekstrakcije puteva, za put je uzet glagol u osnovnom, neodređenom obliku, tj. infinitivu. Time se odbacuje glagolsko vrijeme i lice te gubimo informacije koje nam mogu pomoći u otkrivanju povezanih glagola. Ovakav pristup ne uzima u obzir niti pasivni oblik prijelaznih glagola niti perfekt. Prošlo vrijeme tvori se od nenaglašenog prezenta pomoćnog glagola *biti* i glagolskog pridjeva radnog. Ako je rečenica parsana tako da je glagol *biti* povezan sa subjektom i

objektom rečenice, gubimo te riječi na mjestima X i Y za stvarni, sprezani glagol. Prvi korak u rješavanju navedenog problema bio bi da svaki put kada su subjekt i objekt povezani s glagolom *biti* provjerimo postoji li još neki glagol vezan uz *biti* te, ako postoji, uzmemo taj glagol. Tada bi opet glagol uzimali u infinitivu, ali bi imali više informacija o njemu. Drugi korak je razvijanje modela koji će iskoristiti glagolska vremena. Ona mogu gotovo direktno ukazivati na *prethođenje* relaciju (npr. *Nedavno je kupio auto, ali će ga prodati.*) te vrijedi razmotriti i njihovo povezivanje sa sintaktičkim uzorcima.

Ekstrakcijom samo relacija subjekt-glagol-objekt u rečenici dobivamo puteve koji su praktički samo jedan glagol. Moguća je nadogradnja algoritma tako da u put uzimamo sve riječi između dvije imenice koje nose neko značenje (imenice, glagoli, pridjevi, prilozi), kao Lin i Pantel (2001). Tako vodimo brigu o pasivnom obliku te proglašavamo puteve kao npr. *X rješava Y* i *Y je riješen s X* sličnima. Ako bismo iz takvih, dužih puteva izvukli samo glagol, vjerojatno bismo dobili podskup originalno ekstrahiranih sličnih glagola. Moguće je da je taj podskup bolji, tj. precizniji, što treba provjeriti evaluacijom konačnih rezultata. No, želimo li iskoristiti cijeli put za ekstrakciju semantičkih glagolskih relacija morat ćemo promijeniti pristup problemu. Baza sličnih dugačkih puteva može biti iznimno korisna u rješavanju različitih zadataka obrade prirodnog jezika, uključujući parafraziranje teksta i one spomenute u uvodu. Izrada takvog resursa nije tema ovog završnog rada, ali možda jest nekog budućeg.

Algoritam ekstrakcije konkretnih semantičkih glagolskih relacija ima puno više prostora za nadogradnju od DIRT-a, a ujedno i više propusta. Prva promjena koja se nameće jest dodavanje relacije *omogućava*, koju u referentnom radu ekstrahiraju, a opisana je u odjeljku 2.1. Ona nije tražena iz jednostavnog i neopravdanog razloga – nije smišljen dovoljno dobar uzorak koji bi ukazivao na nju. Drugi propust je mali broj leksičkosintaktičkih uzoraka i njihova upitna kvaliteta. Kvalitetu uzorka dijelom možemo provjeriti tako da ga izbacimo iz skupa svih uzoraka i provedemo postupak traženja relacija. Ako su rezultati bez uzorka bolji nego s uzorkom, zaključujemo da on nije dobar. Rezultati se mogu promijeniti u smislu povećanja preciznosti sustava i blagog pada odziva. Potrebno je izbacivati pojedine uzorke, pratiti rezultate modela te odlučiti koje uzorke zadržati. Osim ručne izrade i smišljanja uzoraka, postoji mogućnost njihovog automatskog generiranja. Time možda možemo proširiti skup uzoraka i poboljšati cjelokupni sustav, ali su metode kompleksne te njihova primjena nadilazi opseg ovog rada.

Za neki par glagola i detektiranu semantičku relaciju pamtimo samo broj

ponavljanja te relacije, što možda nije dovoljno precizno. Računanje neke vrste mjere uzajamne informacije, kao kod DIRT-a, moglo bi dati točnije rezultate. Mjera bi uzimala u obzir broj pojavljivanja pojedinih konkretnih uzoraka koji su ukazali na određenu relaciju i broj pojavljivanja same relacije. Uzmimo za primjer da je relacija *sličnost* pronađena između 5000 parova glagola, relacija *antonimija* između 500 te je broj pojavljivanja prve za glagole *kupiti :: prodati* 50, a druge 45. Trenutni algoritam ignorirao bi relaciju *antonimija*, ali možemo naslutiti da je, zbog značajno manjeg ukupnog broja pojavljivanja te relacije, ona ispravnija odluka. Istu logiku možemo primijeniti i na odnose s pojedinim uzorcima. Što se više puta uzorak pojavio, to je vjerojatnije da izaziva šum. Formula koja bi objedinila sva navedena razmatranja nije smišljena, ali je u (Chklovski i Pantel, 2004) dana slična formula koja pokušava riješiti navedeni problem.

Chklovski i Pantel (2004) primjenjuju svoj model na podatke dobivene internetskom tražilicom. Par glagola popunjava mjesta X i Y u uzorku te se takav uzorak šalje kao upit na tražilicu. Indikacija za postojanje semantičke relacije je broj pronađenih rezultata za uzorak koji ukazuje nju. Iako djeluje vrlo jednostavno, ovakav pristup rezultirao je skupom od 29165 glagola sa 65.5% točno označenih relacija. Primjena ovakvog modela na hrvatski jezik i prilagodba našem je moguća i jednostavna. Broj rezultata dobivenih tražilicom zamijenio bi broj ponavljanja pojedine relacije u korpusu. Rezultate takvog sustava teško je procijeniti, ali vjerojatno točnost ne bi bila veća od originalnog sustava.

Jedna od ideja za budući rad jest i primjena posebnih uzoraka direktno na ovisnosno stablo. Pretpostavka je da se neke glagolske relacije u rečenici mogu identificirati prema mjestu na kojem se glagoli nalaze i onome što se nalazi između njih u stablu. Ovakav pristup primjenjivali bismo uz provjere pojavljuje li se regularni izraz u rečenici ili potpuno zasebno. U našem algoritmu tražimo uzorke u neparsanoj rečenici, iako iteriramo po parsanom korpusu, što je šteta ne iskoristiti. Izrada uzoraka za primjenu na stablo zahtijeva proučavanje odnosa određenog broja parova glagola za koje znamo semantičku relaciju u ovisnosnom stablu.

Skup za testiranje sadrži poprilično malo primjera te zbog toga rezultati možda nisu reprezentativni. Na smislenost rezultata utječe i označavanje samo jedne relacije za određeni par glagola, dok ih u stvarnosti može biti više. Za zadovoljavajuće vrednovanje sustava potrebno je označiti puno više parova glagola. S više primjera u uzorku zlatnog standarda bi i podrezivanje identificiranih semantičkih relacija bilo preciznije.

Iz priloženog vidimo da je napravljeno nekoliko propusta te da postoji mnogo prostora za poboljšanje modela. Kako bi se dobio što bolji model, potrebno je implementirati nekoliko verzija svake komponente te provjeriti rezultate za sve kombinacije takvih verzija. Moguće je mijenjati različite parametre (npr. veličina grozda kod izlaza DIRT-a, broj uzoraka) te promjena svakoga može promijeniti preciznost i odziv sustava. Jedna od gore nespomenutih modifikacija je ubacivanje strojnog učenja u model. Tijekom izrade rada bilo je razmišljanja o tome, ali nije definirano koji se dijelovi i na koji način mogu poboljšati strojnim učenjem.

6. Zaključak

Baza semantičkih glagolskih relacija može biti od velike koristi u raznim područjima obrade prirodnog jezika. U ovom završnom radu razvijen je model za automatsku ekstrakciju sličnosti, antonimije, intenziteta i prethođenje relacija između parova glagola. Po uzoru na VerbOcean (Chklovski i Pantel, 2004) prvo je stvorena baza od 82920 vrlo povezanih glagola, odnosno onih za koje sumnjamo da imaju nekakvu semantičku relaciju. Takav resurs sam po sebi može poslužiti i za neke druge zadatke, a ne samo za traženje semantičkih relacija. Zatim su smišljeni leksičkosintaktički uzorci koji ukazuju na pojedine relacije te su primjenjeni na korpus hrvatskoga jezika. Dobiven je poprilično malen skup od 1663 para semantički povezanih glagola, no postoji mnogo prostora za nadogradnju.

Budući rad obuhvaća prilagođavanje sintaktičkih uzoraka u smislu smanjenja šuma i povećanja preciznosti modela. Drugačiji pristup kreiranju uzoraka bio bi njihovo automatsko generiranje, što može rezultirati velikim brojem uzoraka. Model se može proširiti na puteve koji sadrže više riječi, čime bi, ne samo otvorili prostor za definiranje više specifičnijih uzoraka, već i stvorili korisnu bazu parafraza. Konačno, vrijedi razmotriti primjenu modela na web – najveći jezični resurs na svijetu.

Literatura

Collin F Baker, Charles J Fillmore, i John B Lowe. The berkeley framenet project. U *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, stranice 86–90. Association for Computational Linguistics, 1998.

Ken Barker i Stan Szpakowicz. Interactive semantic analysis of clause-level relationships. U *Proceedings of the Second Conference of the Pacific Association for Computational Linguistics (PACLING 95), pages 22 30, Brisbane, Australia*, 1995.

Matthew Berland i Eugene Charniak. Finding parts in very large corpora. U *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, stranice 57–64. Association for Computational Linguistics, 1999.

Sharon A Caraballo i Eugene Charniak. Determining the specificity of nouns from text. U *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, stranice 63–70. Citeseer, 1999.

Timothy Chklovski i Patrick Pantel. Verbocean: Mining the web for fine-grained semantic verb relations. U *EMNLP*, svezak 2004, stranice 33–40, 2004.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, i Alexander Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1):91–134, 2005.

Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.

Roxana Girju, Adriana Badulescu, i Dan Moldovan. Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135, 2006.

Zellig S Harris. Distributional structure. *Word*, 1954.

David G Hays. Dependency theory: A formalism and some observations. *Language*, stranice 511–525, 1964.

Marti A Hearst. Automatic acquisition of hyponyms from large text corpora. U *Proceedings of the 14th conference on Computational linguistics-Volume 2*, stranice 539–545. Association for Computational Linguistics, 1992.

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

Karin Kipper, Hoa Trang Dang, Martha Palmer, et al. Class-based construction of a verb lexicon. U *AAAI/IAAI*, stranice 691–696, 2000.

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

Dekang Lin i Patrick Pantel. Discovery of inference rules for question-answering. *Natural Language Engineering*, 7(04):343–360, 2001.

Nikola Ljubešić i Tomaž Erjavec. hrwac and slwac: Compiling web corpora for croatian and slovene. U *Text, Speech and Dialogue*, stranice 395–402. Springer, 2011.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Martha Palmer, Daniel Gildea, i Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106, 2005.

Patrick Pantel i Marco Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. U *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, stranice 113–120. Association for Computational Linguistics, 2006.

Deepak Ravichandran i Eduard Hovy. Learning surface text patterns for a question answering system. U *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, stranice 41–47. Association for Computational Linguistics, 2002.

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.

Cornelis Joost Van Rijsbergen. Foundation of evaluation. *Journal of Documentation*, 30(4):365–373, 1974.

Dodatak A
Ručno označene semantičke relacije

U ovom poglavlju nalaze se ručno označene semantičke relacije parova glagola. Tablica A.1 sadrži 454 para uzetih iz izlaza DIRT algoritma, a tablica A.2 332 para iz konačnog izlaza postupka. Oznake relacija su: *s – sličnost, i – intenzitet, h – prethođenje, a – antonimija, n – nema relacije.*

Tablica A.1: Ručno označene glagolske relacije izlaza DIRT-a.

a olakšavati :: komplicirati	a oraspoložiti :: obeshrabriti	a pomilovati :: osumnjičiti
a rashoditi :: prihodovati	a razveseliti :: uznemiriti	h dizati :: podignuti
h doživjeti :: pamtiti	h kleknuti :: prositi	h utemeljiti :: izgraditi
n angeti :: okomiti	n anketirati :: mobilizirati	n artikulirati :: opstruirati
n avati :: nematiti	n bavititi :: zakinuti	n bijasati :: otezati
n bijati :: glupiti	n bitnjeti :: krviti	n bježati :: zaostajati
n boli :: svađati	n boliti :: savjetivati	n citirati :: prevesti
n crtati :: odveli	n dаći :: otprijati	n dani :: redati
n davati :: jamčiti	n debeti :: dijagonati	n debeti :: omesti
n debljati :: propadati	n definirati :: zanemarivati	n diktirati :: prevagnuti
n dočarati :: nabijati	n doći :: krenuti	n doći :: pamtiti
n dodajtiti :: oboliti	n dodirovati :: micati	n doljati :: reproducirati
n dometnuti :: izbaci	n dometnuti :: prepričavati	n doprinjeti :: sputati
n dospjeti :: pripraviti	n dostizati :: dešavati	n doti :: stavimti
n dozirati :: privucati	n dozivati :: zarasti	n dragti :: leći
n držeći :: uvlačiti	n dužiti :: dostupeti	n eksponirati :: fiksirati
n emigrirati :: oboljeti	n formulirati :: postrožiti	n forsirati :: prokleti
n gaji :: dopusti	n garažiti :: nedati	n glasovati :: odraziti

n gledatiti :: mucati
n gušiti :: podcjenjivati
n idati :: pretiti
n isfurati :: ginuti
n ispitivati :: prepisivati
n isplivati :: zaostati
n istupiti :: sročiti
n izdići :: boljeti
n izmisliti :: initi
n izraditi :: analizirati
n izudarati :: zamotati
n jačati :: uništati
n jamčiti :: poboljšati
n kandidirati :: izbrojati
n koći :: kožati
n kotirati :: budeti
n kumovati :: dopadati
n kupujetiti :: usisavati
n lajati :: gubititi
n libiti :: tiskati
n ljepšiti :: uzrujavati
n miješati :: sjetititi
n molimti :: promotriti
n mrziti :: zavući
n nadodati :: daviti
n nadvisiti :: šutirati
n nagraditi :: ispratiti
n najgori :: vozeći
n naniti :: faliti
n nasjedati :: izabirati
n nasrnuti :: kombi
n navati :: oprosti
n nazovati :: opraštati
n negoliti :: usidriti
n obećavati :: unaprijediti
n objedinjavati :: očajavati
n gosti :: objasni
n hapsiti :: odraziti
n ilustrirati :: iskristalizirati
n ispadati :: ginuti
n ispitivati :: zanemarovati
n ispravljati :: evidentirati
n izazvati :: čuti
n iziskovati :: bijasati
n iznjedriti :: ustaliti
n izraditi :: utvrđivati
n izumiti :: promotriti
n jahati :: natuknuti
n jesti :: odmahivati
n kititi :: pogledatiti
n kornati :: fakati
n krivi :: dobjeti
n kumovati :: veseli
n kvaliteti :: ucati
n lezati :: uteći
n likovati :: zanemarovati
n mami :: ispariti
n mirisati :: pripomoći
n moli :: svaliti
n nadahnuti :: odviti
n nadovezati :: progovoriti
n nagovarati :: poslali
n naigrati :: grliti
n naletiti :: pretjecati
n napati :: progledati
n naslućivati :: prepirati
n naštetiti :: začuđivati
n navršiti :: lagati
n nebi :: ubojiti
n nudimti :: opominjati
n obilovati :: iziskovati
n objesiti :: zakinuti
n gotoviniti :: anti
n hrabriti :: prijavititi
n intervenirati :: poskupljivati
n ispati :: paraziti
n isplaćivati :: uprihoditi
n istući :: uzvrati
n izbijati :: uzrujavati
n izjednačiti :: primaknuti
n izostajati :: opadati
n izreći :: sazvati
n izvući :: prevezti
n jaknuti :: usavršavati
n kaknuti :: uklopiti
n klicati :: usuđivati
n kositi :: gibati
n krokoditi :: portfoliti
n kupititi :: nabavititi
n kvalitetniji :: najbržati
n ležiti :: preoteti
n listi :: isploviti
n medicinski :: pojednostaviti
n mobiteti :: razlikivati
n motiviti :: siviti
n nadmašiti :: doseći
n nadovezivati :: replicirati
n nagovijestiti :: kapitalizirati
n najbitnjeti :: minimalizirati
n namjeravati :: usuglasiti
n napišeti :: interesirati
n naspavati :: tugovati
n naučiti :: zanemarivati
n navršiti :: letiti
n negoli :: alati
n obati :: visati
n objašnjavati :: proučiti
n obložiti :: izvucati

n obmanjivati :: iznenađivati	n obogaćivati :: nakupljati	n obojati :: izgubititi
n obraniti :: nizati	n obrisati :: citati	n obrnuti :: skresati
n obuzeti :: prisiti	n očajavati :: oprosti	n očekivati :: obećavati
n očitati :: nuti	n očitovati :: udavati	n odašiljati :: sagledavati
n odatti :: pitatiti	n odbaci :: raspraviti	n odijeti :: limitirati
n odjuriti :: dragati	n odletiti :: najljepši	n odmaknuti :: svađati
n odobravati :: obavještavati	n odrediti :: razmatrati	n odrgati :: ustajati
n odšetati :: polomiti	n odugovlačiti :: omalovažavati	n odupirati :: poistovjećivati
n odužiti :: zatrebati	n odvijati :: utjecati	n oglušiti :: obazirati
n okititi :: sakupiti	n okrugli :: okrući	n opozvati :: poništi
n ordinirati :: postavititi	n ordinirati :: primjeniti	n oslabjeti :: navješćivati
n osloniti :: misati	n osmjehnuti :: prevoditi	n ostvarovati :: cijeniti
n osvrnuti :: demantirati	n osvrnuti :: smijati	n otapati :: propagirati
n otkinuti :: pusati	n otplaćivati :: naknaditi	n outi :: podrediti
n ovjeriti :: opominjati	n ovlašćivati :: otplaćivati	n ovlašćivati :: uručivati
n ozljediti :: povucati	n pameti :: uzburkati	n paraziti :: boli
n parirati :: dinati	n pauzirati :: šesti	n pecati :: kleknuti
n pecati :: sjetititi	n pitati :: podsjetiti	n pješačiti :: sisati
n plačati :: natopiti	n plešiti :: kasni	n pljačkati :: opredijeliti
n pljunuti :: uvezti	n pljuvati :: dešavati	n pobijati :: izvoliti
n počastiti :: očarati	n poći :: masti	n podnijeti :: izvijestiti
n podučiti :: apetiti	n podupirati :: pozdravljati	n poduprati :: razlikivati
n podvaliti :: preprodati	n pohvati :: opraštati	n poigrati :: kloniti
n poigravati :: mlatiti	n poistovjećivati :: kažiti	n pokoriti :: baštiniti
n pokositi :: mješoviti	n pokositi :: uspjevati	n polijevati :: ludi
n politi :: odvući	n politi :: uzbuđivati	n poljuljati :: zloupotrijebiti
n polučiti :: odmicati	n pomoziti :: moljeti	n pomrsiti :: lošiji
n pomutiti :: grudi	n poravnati :: načeti	n pošaljiti :: očajavati
n pospremati :: ordinirati	n posuđivati :: privučiti	n posustati :: zanimljivti
n posviti :: manipulirati	n potkrepljivati :: pribavljati	n potrajati :: opustiti
n potući :: jati	n potući :: osvrtati	n povegati :: očarati
n povisiti :: okruniti	n povisivati :: ći	n povoditi :: omogučiti
n pozabaviti :: preispitati	n praksiti :: pošaljati	n prašiti :: klasiti
n prebaciti :: provocirati	n preboliti :: tajiti	n predbilježiti :: uveseljavati
n prediti :: aplicirati	n premjestiti :: svađati	n preraditi :: zakinuti
n preskakati :: letjeti	n preskakati :: pljačkati	n preskočiti :: rušiti

n preslušati :: izuzeti
n prikazivati :: zamisliti
n primaknuti :: plocati
n prisiliti :: rastužiti
n prisustvivati :: dokumentirati
n prkositi :: omesti
n procjenjivati :: razmotriti
n progovoriti :: sastati
n propitivati :: zagledati
n proširivati :: zaokruživati
n prouzročiti :: potresti
n rangirati :: kloniti
n raspoloživjeti :: svati
n razbuktati :: dodušiti
n razlučivosti :: dijagonati
n restrukturirati :: kožati
n rukovoditi :: pripraviti
n rusati :: vatrati
n sakupiti :: primaknuti
n senzibilizirati :: motivirati
n sjediti :: svađati
n skloniti :: zasnovati
n skupljati :: podmirivati
n slavljati :: potpomoći
n smiti :: najskupljati
n sresti :: škoditi
n stigati :: plivati
n štrajkati :: kopasti
n sučeliti :: participirati
n svojedobnti :: pomilovati
n tituti :: ohrabrivati
n transportirati :: iznaći
n ubrizgati :: zakinuti
n udomiti :: sačiniti
n ugrizati :: pozliti
n uletjeti :: osloniti
n prestati :: klanjati
n priklonijeti :: nabavititi
n primicati :: srozati
n prisluškovati :: uspjevati
n prisvojiti :: vinuti
n probati :: važiti
n prodirati :: progledati
n promjeniti :: znakoviti
n propuštati :: raznati
n proslavljati :: dežurati
n prozboriti :: naucati
n rashoditi :: kreditirati
n raspršivati :: poskupljivati
n razbuktati :: smjenjivati
n razmijeniti :: pržiti
n reti :: najjacati
n rusati :: nedati
n sadrgati :: dostupeti
n šaliti :: osvrtati
n sjajiti :: aferiti
n skijati :: pohitati
n školovati :: odvažiti
n skupti :: pogurati
n slući :: kockati
n smiti :: prokleti
n štati :: jeli
n stopasti :: iščitati
n stražiti :: montirati
n sustići :: usaditi
n svratiti :: presjeći
n točiti :: zaključavati
n ubaci :: otključavati
n učestati :: idemti
n udovoljiti :: obiti
n ugurati :: pospremiti
n umiješati :: gostiti
n prestati :: razboljeti
n prikupljati :: rasporediti
n pripisivati :: napominjati
n pristupajati :: pjevušiti
n privucati :: usuđivati
n probiti :: popeti
n produljivati :: zarasti
n propati :: kontroliti
n proraditi :: odšetati
n protjecati :: ohladiti
n rađati :: povraćati
n raspolagati :: zaplijeniti
n razbijati :: provlačiti
n razljutiti :: etiketirati
n reprogramirati :: obustavljati
n riječjati :: praksati
n rusati :: protegati
n sadržati :: osiguravati
n sastojati :: povezivati
n sjećati :: razumiti
n skloniti :: usuđivati
n skrbiti :: najbližiti
n sladiti :: posuti
n smijetiti :: diskriminirati
n spriječavati :: napameti
n stavljati :: snositi
n stopirati :: dasti
n stti :: kampirati
n sviditi :: različiti
n svrstavati :: nebi
n topiti :: prelazi
n ubrizgati :: imenivati
n udaljiti :: vežiti
n udruzi :: obavještavati
n ukazati :: dirnuti
n umrijeti :: patiti

n uništavati :: igramti	n unositi :: kupujetiti	n upamtiti :: povraćati
n upotpunjavati :: stiskati	n uputititi :: znaci	n uravnotežiti :: signati
n uručivati :: suglasiti	n uskratiti :: zloupotrijebiti	n uspiti :: frustrirati
n uspjeti :: saznati	n uspoređivati :: spočitavati	n ustajati :: jati
n ustrojavati :: zaduživati	n utećі :: opominjati	n utrčati :: ocati
n uvlačiti :: prešućivati	n uvlačiti :: urlati	n uživati :: čivati
n valjdati :: citati	n valorizirati :: iskristalizirati	n valorizirati :: razapeti
n veseti :: zapostavljati	n višiti :: uzletjeti	n viti :: najdražiti
n viti :: užiti	n volontirati :: transformirati	n vući :: trubiti
n zabijati :: opaliti	n zabiti :: izboriti	n zafrkavati :: osvrtati
n zahvalni :: podcijeniti	n zakopasti :: zavrtiti	n zaliti :: simpatizirati
n zamrijeti :: pripovijedati	n zamrijeti :: segati	n zanemarovati :: likovati
n zanimljiviji :: pokrati	n zanimljivti :: najljepsati	n zanositi :: omogučiti
n zapaliti :: opljačkati	n zaplivati :: nakupiti	n zaplivati :: plaknuti
n zasvirati :: asfaltirati	n zatrebati :: sprijateljiti	n zaustavljati :: nedati
n zavezati :: istrošiti	n zavisiti :: nagađati	n zavoditi :: klati
n zazivati :: prositi	n zazvonijeti :: podsjeti	n zbrinuti :: dokapitalizirati
n zloupotrijebiti :: zvuči	n žuti :: zvoniti	s analizirati :: ispitati
s baciti :: bacati	s izvjestiti :: izvještati	s naletiti :: sudariti
s napustiti :: okončati	s naučiti :: podučiti	s negirati :: odbaciti
s negirati :: poricati	s odraditi :: realizirati	s opsovati :: psovati
s pobijediti :: osvojiti	s poharati :: uništiti	s polomiti :: otrgnuti
s potpisati :: podržati	s potresati :: uzdrmati	s prenositi :: širiti
s preraditi :: reciklirati	s promovirati :: reklamirati	s sprovesti :: sprovoditi
s stupiti :: ušetati	s teretiti :: optuživati	s vratiti :: vraćati

Tablica A.2: Ručno označene glagolske relacije izlaza cjelokupnog modela.

s misliti :: razmišljati	h analizirati :: otkriti	s voljeti :: voliti
n najavljivati :: tražiti	n alati :: psi	s pojaviti :: pojavljivati
n zabilježiti :: pridonijeti	s najavljivati :: prognozirati	s boriti :: natjecati
h pročitati :: prokomentirati	n pratiti :: prilagođavati	n uslijediti :: kulminirati
s silovati :: tući	n tražiti :: predlagati	h vježbati :: naučiti
s otpjevati :: izvesti	s otežavati :: pogoršavati	n uvidjeti :: iziskivati
s nadzirati :: kontrolirati	s naručiti :: naručivati	a ustvrditi :: demantirati
s održavati :: njegovati	n trati :: nagađati	h čekati :: dobiti

n znati :: pročitati
s pogoditi :: zabiti
h primijetiti :: ispraviti
s povećati :: pojačati
n govoriti :: donijeti
n napadati :: preskočiti
n iskoristiti :: shvatiti
n vrijediti :: dosegnuti
s planirati :: organizirati
n razumjeti :: željeti
n izbaciti :: igrati
h pitati :: razumjeti
s brinuti :: zabrinuti
s omogućiti :: olakšati
s nastojati :: truditi
a popeti :: spustiti
s uzvratiti :: vratiti
s dodati :: ubaciti
n razumjeti :: oprostiti
n donositi :: uključivati
a roditi :: umrijeti
a poticati :: pogoršavati
s čuvati :: njegovati
n kupivati :: željeti
s poznavati :: razumjeti
n znati :: dogovoriti
s naći :: pronalaziti
s prezentirati :: prikazati
s sastavljati :: izrađivati
s određivati :: odrediti
n odvesti :: truditi
a sklapati :: raskidati
n slaviti :: zabiti
n nositi :: bojati
h posjedovati :: koristiti
n ugovoriti :: kontrolirati
s utvrđivati :: istraživati
n rušiti :: preispitivati
a uvoziti :: izvoziti
s izboriti :: plasirati
s plaćati :: snositi
n ograničiti :: planirati
a propustiti :: realizirati
s tužiti :: optužiti
s pretvarati :: stvarati
n urediti :: darovati
a smanjiti :: povećati
h razumjeti :: objasniti
n označavati :: svoditi
n pripremiti :: obogatiti
n spavati :: pitati
s izgubiti :: propustiti
n vapiti :: istjerati
s promovirati :: promicati
a kupiti :: prodati
n nastavljati :: trajati
s ostvariti :: ostvarivati
n usuditi :: pomisliti
h odigrati :: zabiti
n osušiti :: izravnati
n dirati :: povući
a dobiti :: platiti
s milovati :: maziti
s povesti :: voditi
n vjerovati :: dogoditi
s analizirati :: usporediti
s nadigrati :: nadjačati
s zahvaćati :: zahvatiti
h zaslužiti :: igrati
s shvatiti :: razumjeti
s sklopiti :: dogovoriti
n ovisiti :: kretati
n sreti :: fotografirati
s definirati :: propisivati
n stjecati :: razvijati
s prodavati :: prodati
s dijeliti :: darivati
s obećati :: dogovoriti
s potvrditi :: istaknuti
n shvatiti :: vidjeti
n odrasti :: misliti
s prenijeti :: komentirati
s naglasiti :: istaknuti
n odlučiti :: potražiti
s kretati :: prelaziti
s uzimati :: uzeti
n povesti :: napraviti
n stići :: izboriti
s postići :: osvojiti
n zahvaliti :: kazati
s suspendirati :: kazniti
s objasniti :: opravdati
h gledati :: naučiti
s plati :: naplaćivati
h zaprijetiti :: pristati
h uzeti :: baciti
s kazati :: ispričati
n shvatiti :: iskorištavati
s staviti :: ubaciti
s otići :: odlaziti
n igrati :: otići
h pucati :: promašivati
n nastaviti :: planirati
s prikupiti :: osigurati
h zamisliti :: stvoriti
s pozdravljati :: podržavati
s voditi :: kontrolirati
s čestitati :: poželjeti

n nalagati :: vežiti
n stići :: pasti
n pobijediti :: doći
n snaći :: dokazati
n nemojti :: vjerujti
n isporučivati :: pokretati
n placati :: smjeti
n dati :: donijeti
a završavati :: početi
s putovati :: ići
s omogućavati :: nuditi
n nalaziti :: prihvatiti
a povećati :: smanjiti
h držati :: spustiti
s pjevati :: zabavljati
n shvatiti :: naći
h odigrati :: izgubiti
s pomagati :: stimulirati
n dolaziti :: truditi
n živjeti :: spavati
n obilježiti :: obogatiti
s nabaviti :: naručiti
n doznati :: ostajati
n pripisivati :: utjecati
n nalaziti :: očekivati
s financirati :: sufinancirati
s poduzeti :: poduzimati
s proširiti :: osvježiti
s odvijati :: vršiti
s kontrolirati :: regulirati
n saznati :: naći
n donijeti :: govoriti
a plaćati :: primati
s ostati :: ostajati
a razveseliti :: rastužiti
n pokrenuti :: osmisliti
n obnašati :: imenovati
s djelovati :: utjecati
s doživjeti :: doživljavati
s uključivati :: obuhvaćati
s povećati :: povećavati
n isporučivati :: integrirati
s demonstrirati :: dokazati
n nalaziti :: značiti
s utvrditi :: otkriti
a mrziti :: obožavati
s oduševiti :: impresionirati
n složiti :: priznati
s pokazati :: prikazati
n zaključivati :: prenositi
n voljeti :: kupivati
n složiti :: upozoriti
n stidjeti :: otplaćivati
a napustiti :: priključiti
n koštati :: potrošiti
h osvojiti :: slaviti
n upisivati :: stjecati
n objedinjivati :: postojati
s izvući :: spasiti
s operirati :: odstraniti
s dostignuti :: narasti
h razviti :: testirati
n otići :: sjediti
n iznenaditi :: nasmijati
n odvezati :: krenuti
n ubilježiti :: popeti
n truditi :: brinuti
a zabijati :: promašivati
h planirati :: napraviti
n uspjeti :: saznati
s kritizirati :: prozivati
s usporiti :: otežati
n primjećivati :: misliti
s fotografirati :: snimati
h razvijati :: testirati
n odgovoriti :: ponoviti
n mučiti :: ostati
s istraživati :: ispitivati
n podržavati :: donositi
s formirati :: osnivati
h napasti :: pobjeći
s provesti :: realizirati
s obnoviti :: obnavljati
s vrijeđati :: dobacivati
a spustiti :: podignuti
h dodati :: zabiti
n dostići :: udvostručiti
n osjećati :: shvaćati
n kanati :: alati
n srušiti :: upati
i najavljivati :: obećavati
n konzumirati :: voljeti
s navesti :: navoditi
n patiti :: usuđivati
n izgledati :: pokazati
n predstavljati :: razviti
n saznati :: izaći
n osvojiti :: zabiti
s promatrati :: percipirati
n dići :: doći
s obratiti :: reći
n izraditi :: odabirati
h ustati :: pohitati
n trebati :: objaviti
s ispitivati :: istraživati
n znati :: nalaziti
s otkloniti :: ublažiti
n okrenuti :: truditi

n dogovoriti :: udružiti	s dodati :: dodavati	a primicati :: odmicati
h dobijati :: posjedovati	a zagrijavati :: hladiti	n zatražiti :: reagirati
s uplatiti :: donirati	n povećati :: ostvariti	a spustiti :: popeti
a baciti :: nositi	n izdati :: povući	n stići :: prebaciti
n osvanuti :: paliti	s poskupjeti :: poskupiti	s završiti :: prekinuti
s obnavljati :: modernizirati	n nosititi :: držiti	a smijati :: žaliti
n pustiti :: maknuti	n otići :: pristati	n obavijestiti :: obvezati
s uzeti :: izvaditi	s uspostaviti :: uspostavljati	s spriječavati :: usporavati
s dolaziti :: stizati	s podržavati :: promicati	n značiti :: omogućavati
s nasmijati :: oduševiti	s kontrolirati :: držati	n pozdraviti :: istaknuti
n ticati :: jeti	s inspirirati :: intrigirati	a promašivati :: pogađati
s podržati :: poduprijeti	s izdvajati :: isticati	s poslati :: predati
n razviti :: isporučiti	s predati :: predočiti	n izazvati :: pridonijeti
a pojačavati :: smanjivati	s prisiliti :: natjerati	s otežavati :: usporavati
s naložiti :: odobriti	n zabilježiti :: očekivati	s obožavati :: voljeti
n smjeti :: pomagati	s donirati :: darovati	n razumjeti :: naći
n smijati :: misliti	s potvrđivati :: opravdavati	s usvajati :: prihvaćati
n zalagati :: zaklinjati	h percipirati :: rješavati	n rađati :: doživljavati
n zamjeriti :: odgovoriti	s zaplesati :: zapjevati	s pucati :: ispaliti
n crveniti :: kožiti	s istrčati :: izletiti	n trebati :: reći
s uložiti :: ulagati	n održati :: pripremiti	i prijetiti :: pretući
n predvoditi :: slaviti	h osmisliti :: pripremiti	s birati :: izabrati
n zahvaljivati :: pozdravljati	s uništavati :: uništiti	n isticati :: razmatrati
a ubrzati :: kočiti	a poboljšati :: pogoršati	s provjeravati :: potvrđivati
n pokušavati :: željeti	h odvesti :: vratiti	n koristiti :: donijeti
s obučiti :: educirati	n izgledati :: učiniti	n odgajati :: vjerovati
a usporiti :: ubrzati	s gledati :: promatrati	s preispitati :: razmotriti
h kupiti :: posjedovati	h dijagnosticirati :: liječiti	n izgledati :: funkcionirati
s proći :: krenuti	s olakšati :: poboljšati	n uplašiti :: uspjevati
n slaviti :: odigrati	n ispitati :: dovršiti	a kupovati :: prodavati

Automatska ekstrakcija semantičkih glagolskih relacija iz korpusa na hrvatskome jeziku

Sažetak

Leksičkosemantički jezični resursi nezaobilazni su za semantičku obradu prirodnog jezika i mnoge zadatke u ekstrakciji informacija. Budući da glagoli u tekstu često korespondiraju s predikatnom strukturom teksta, odnosno semantikom događaja, često je potrebno modelirati semantiku glagola odnosno događaja kojima oni odgovaraju. U sklopu završnog rada napravljen je model za ekstrakciju semantičkih glagolskih relacija (sličnost, antonimija, intenzitet i prethođenje) iz korpusa na hrvatskome jeziku. Ručno je označen skup tekstnih podataka te je provedeno eksperimentalno vrednovanje modela.

Ključne riječi: obrada prirodnog jezika, semantičke glagolske relacije, hrvatski jezik

Extraction of Semantic Verb Relations from Croatian Corpora

Abstract

Lexico-semantic language resources are a must for semantic processing of natural language, and many tasks in information extraction. As the verbs in the text often correspond with the predicate structure of the text, or the semantics of events, it is often necessary to model the semantics of the verb or event which they correspond with. This bachelor's thesis proposes a statistical approach for extracting verb relations (similarity, antonymy, intensity and happens-before) from Croatian web corpus.

Keywords: natural language processing, verb sematics, Croatian language