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. 1322

Identifikacija višerječnih izraza zasnovana na kombinaciji jezičnih značajki

Maja Buljan

Zagreb, srpanj 2016.

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

Zagreb, 11. ožujka 2016.

Predmet: **Analiza i pretraživanje teksta**

DIPLOMSKI ZADATAK br. 1322

Pristupnik: **Maja Buljan (0036451668)**
Studij: Računarstvo
Profil: Računarska znanost

Zadatak: **Identifikacija višerječnih izraza zasnovana na kombinaciji jezičnih značajki**

Opis zadatka:

Višerječni izrazi, uključivo frazemi, stručno nazivlje, leksičke kolokacije i ustaljene fraze, od posebne su važnosti u prirodnome jeziku zbog razmjerno nepredvidivih sintaktičkih, semantičkih i statističkih obilježja. U području računalne lingvistike, velika je pažnja posvećena automatskoj identifikaciji višerječnih izraza iz korpusa na temelju statističkih postupaka. Predložen je niz modela temeljenih na statističkim obilježjima višerječnih fraza, kao i modela koji razmatraju sintaktička i semantička obilježja, poput sintaktičke rigidnosti, semantičke netransparentnosti, leksičke okamenjenosti i sl.
U okviru diplomskoga rada potrebno je proučiti pristupe identifikaciji višerječnih izraza i pristupe vrednovanju tih postupaka. Posebno razmotriti pristupe temeljene na metodama strojnog učenja kao i pristupe koji razmatraju više jezičnih značajki, uključivo generativan model Tsvetkove i Wintnera (2014). Razviti programsku implementaciju modela za ekstrakciju višerječnih izraza i primijeniti ga na korpus tekstova na hrvatskome jeziku. Izraditi prikladan ispitni skup podataka te provesti iscrpno vrednovanje modela, uključivo analizu interakcije jezičnih značajki, analizu pogrešaka te usporedbu s referentnim modelima i slobodno dostupnim rješenjima. Radu priložiti izvorni i izvršni kod razvijenog sustava, skupove podataka i programsku 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

SADRŽAJ

1. Uvod

Višerječni izraz (engl. *multiword expression*, MWE) jezična je jedinica sastavljena od dva ili više leksema. Višerječnim izrazima smatraju se idiomi, ustaljene fraze, kolokacije, stručno nazivlje, strani izrazi, koji se međusobno razlikuju po svojstvima poput specifične morfologije i sintaktičke fleksibilnosti (sastavnica unutar fraze ili fraze u kontekstu rečenice) te razine semantičke neprozirnosti (Baldwin i Kim, 2010), no za sve je skupine višerječnih izraza karakteristično specifično ponašanje i značenje u jeziku, koje odstupa od predvidljivog načina zasebne upotrebe sastavnica fraze. Zbog svoje idiomatske prirode, višerječni izrazi nosioci su "duha jezika", a prema gruboj procjeni (Jackendoff, 1997) u vokabularu izvornog govornika prisutni su u istom redu veličine kao i individualne riječi. Pojavljuju se u svakodnevnom govoru i žargonu, u diskurzivnim, informativnim i književnim djelima, a posebno su česti u stručnoj terminologiji raznovrsnih područja.

Visoka prisutnost višerječnih izraza ukazuje na korisnost njihove točne identifikacije i upotrebe u zadacima obrade prirodnog jezika, poput određivanja značenja riječi, ekstrakcije informacija, sažimanja i generiranja teksta te strojnog prevođenja, no problem ispravne identifikacije otežan je njihovom jezičnom osebujnošću. Višerječni se izrazi, stoga, uvrštavaju u leksikone sustava za obradu prirodnog jezika, pri čemu je automatizacija procesa poželjna radi vremenske efikasnosti i visokog odziva.

U ovome radu razvija se i istražuje djelovanje sustava za identifikaciju višerječnih izraza, koji primjenjuje algoritme strojnog učenja za binarnu klasifikaciju mogućih višerječnih izraza (predložena sintagma *jest* ili *nije* višerječni izraz u hrvatskome jeziku). Sustav objedinjuje statistički i pristup temeljen na lingvističkim pravilima, kao što predlažu Sag et al. (2002), a sastoji se od dva dijela: u prvom koraku, iz korpusa tekstova ekstrahiraju se potencijalni primjeri višerječnih izraza (u daljnjem tekstu: kandidati), koristeći sintaktičke uzorke – sljedove vrsta riječi – karakteristične za višerječne izraze u hrvatskome jeziku. U drugom se koraku, po uzoru na (Tsvetkov i Wintner, 2014), statistički određuju vrijednosti mjera koje odgovaraju jezičnim značajkama karakterističnim za višerječne izraze. Mjere se potom organiziraju u strukturu

Bayesove mreže, kojom se vrši binarna klasifikacija kandidata i identificiranje pozitivnih i negativnih primjera višerječnih izraza.

Sustav je učen i vrednovan na primjerima koji zadovoljavaju najširu moguću definiciju višerječnog izraza; ne specijalizira se za specifičnu vrstu, konstrukciju ili duljinu (n-gram) višerječnog izraza. Jezične značajke koje označavaju kandidate većinom su univerzalno primjenjive, neovisno o jeziku korpusa, no u opsegu ovoga rada prilagođen je za rad, učen i vrednovan samo na hrvatskome jeziku.

Iduće poglavlje daje pobliži opis vrsta i značajki višerječnih izraza te pregled radova na području ekstrakcije i klasifikacije višerječnih izraza. U trećem poglavlju izložen je razvoj i način rada svakog dijela sustava, uz detaljan opis svake od jezičnih značajki koje čine okosnicu klasifikacije kandidata. Četvrto poglavlje opisuje pripremu korpusa i podataka za učenje i vrednovanje, postupak i rezultate označavanja pozitivnih i negativnih primjera za učenje, a potom su izloženi rezultati vrednovanja sustava, uz ovisnosti o promjenjivim parametrima. Diskusija rezultata nastavlja se u zadnjem poglavlju, uz diskusiju mogućih izmjena i proširenja te izvedenih zaključaka.

2. Pregled područja

U ovome poglavlju daje se kratak pregled dostignuća disciplina u čijem je području interesa rad s višerječnim izrazima, razmatrajući temu iz lingvističke perspektive i iz perspektive područja obrade prirodnog jezika (engl. *natural language processing*, NLP). Prvo potpoglavlje razmatra definiranje i klasificiranje višerječnih izraza, kako za specifične jezike, tako i s općeprimjenjivog stajališta. U drugome potpoglavlju predstavlja se uzorak značajnijih radova iz područja NLP-a koji se različitim pristupima bave problemom identifikacije višerječnih izraza.

2.1. Definicije višerječnih izraza

Radi boljeg razumijevanja prirode problema i značajnih aspekata postupka identifikacije višerječnih izraza, potrebno je pobliže definirati svojstva koja određuju takve pojmove. Iako se višerječne izraze općenito može definirati kao niz od dviju ili više punoznačnih riječi koje zajedno čine semantičku cjelinu, a mogu se zamijeniti jednorječnim sinonimom, teško je izreći cjelovitu definiciju koja bi pokrila i u potpunosti opisala sve vrste višerječnih izraza, bilo unutar jednog ili skupine jezika, bilo općenito. Zbog raznolikosti svojstava višerječnih izraza, čak ni specifičnosti ponašanja u morfološkom, sintaktičkom i semantičkom smislu ne mogu se proglasiti dijeljenim svojstvom osim u najširem mogućem tumačenju, s obzirom na to da i te specifičnosti variraju od jedne do druge vrste te od jezika do jezika. Potrebno je, stoga, definirati višerječne izraze kroz nekoliko zasebnih skupina, međusobno različitih po prirodi navedenih svojstava.

2.1.1. Klasifikacija u hrvatskome jeziku

Leksikografska literatura hrvatskoga jezika za višerječne izraze koristi brojne nazive: kolokacije, sveze riječi, višeleksičke jedinice, višeleksičke sveze, višerječne natuknice, višerječne jedinice, višečlani nazivi, višerječne sveze, skupine riječi, skupovi riječi,

sintagmemi, frazemi. Radi razumljivosti i lakšeg povezivanja sa stranom literaturom, a i zato što su pojedini nazivi iz hrvatske frazeologije uži od ovdje korištene radne definicije višerječnog izraza (npr. kolokacija, frazem), u nastavku rada koristit će se termin *višerječni izraz*.

Menac (2010) i Bartolec (2012) nastoje ustanoviti klasifikaciju višerječnih izraza u hrvatskome jeziku i dijele ih u dvije osnovne skupine: (1) gramatičke skupove riječi i (2) leksičke sveze.

Gramatički skupovi riječi Pretpostavljajući da se svaki (po tvorbi neograničen) niz riječi u najširem smislu može nazivati skupom riječi, gramatički skup označava svaku sintagmu nastalu prema strogim gramatičkim pravilima jezika. Budući da je glavno definirajuće obilježje gramatičkih skupova sintaktičko ustrojstvo riječi, takvi skupovi općenito ne zadovoljavaju definiciju višerječnih izraza u obliku u kojem su oni predmet ovoga rada. Ipak, s obzirom na to da višerječni izrazi s gramatičkim skupovima dijele svojstva gramatičke zavisnosti, u matematičkom je smislu skup[1] višerječnih izraza podskup skupa[1] gramatičkih skupova[2] riječi. Radi ilustracije opsega gramatičkih skupova i ograničavajućih gramatičkih pravila koja određuju njihovu tvorbu, u nastavku je dan opis značajki pojedinih kategorija gramatičkih skupova.

Zavisnosti Svojstva koja odvajaju gramatičke skupove od nasumičnih nizova riječi jesu veze gramatičke (valencija) i morfosintaktičke (reakcija) zavisnosti, koje određuju vrste riječi u sintagmi i njihovu sintaksu, i sročnost (kongruencija), koja zahtijeva podudarnost sastavnica u rodu, broju i padežu. Tvorba gramatičkih skupova vrijedi i na formalnoj razini ostvarivanja višerječnog izraza.

Složenice Iako je u grafičkom smislu jedna riječ, složenica se kao spoj dviju leksičkih sastavnica smatra prijelaznim oblikom između jedne riječi i kolokacije (dvočlanog gramatičkog skupa). Ipak, zbog ograničavanja radnog zadatka na skupove riječi odvojenih bjelinom, složenice se neće dalje razmatrati.

Gramatički frazemi Na granici između gramatičkog skupa i višerječnog izraza nalaze se gramatički frazemi – skupovi nepromjenjivih vrsta riječi i složeni glagolski oblici koji nose značenje i imaju odnos prema drugim riječima ili skupovima unutar rečenice i među rečenicama. Takvi su izrazi, primjerice: *zato što*, *budući da*, *kakav*

[1] *mat.* kolekcija izraza
[2] *lingv.* nizovi riječi

god, *sve u svemu*, *morati reći* i dr. Takvi će se izrazi u radu također zanemariti, s obzirom na to da zbog svoje leksičke i kontekstualne nespecifičnosti odskaču od definicije višerječnog izraza kao punoznačne jezične jedinice (također sastavljene od dviju ili više punoznačnih riječi).

Leksičke sveze Ono što se za potrebe rada smatra višerječnim izrazom Bartolec (2012) naziva leksičkim svezama i definira kao: "spojeve riječi temljene na značenjskom odnosu pojedinačnih sastavnica, odnosno čije pojedinačne sastavnice imaju samostalno značenje koje unutar sveze ostaje nepromijenjeno ili se desemantizira". Time se potvrđuje da je glavno zajedničko svojstvo višerječnih izraza djelomično ili potpuno preuzimanje, ili djelomično ili potpuno prekrajanje značenja sastavnih riječi. Leksičke sveze pobliže su opisane trima najzastupljenijim vrstama: nazivima, onimima i frazemima.

Nazivi Naziv je višerječni izraz koji jednoznačno određuje neki pojam, definiran u strogo određenom znanstvenom području. Primjerice: *torzijska vaga*, *solna kiselina*.

Onimi Onim je svako višerječno vlastito ime (ime osobe, geografskog pojma ili institucije) koje služi za univerzalnu identifikaciju pojedinih objekata. U hrvatskome jeziku, onimima se smatraju i svi skupovi riječi koji se u jeziku ponašaju kao onimi, a na opisan ili prenesen način zamjenjuju jednorječni sinonim (npr. *crno zlato* (nafta), *kralj životinja* (lav)).

Frazemi Frazem sačinjava niz riječi koji u govoru ne nastaje spontano, ovisno o situaciji, već u jeziku postoji u gotovom obliku. Frazem je ustaljeni skup riječi čija je barem jedna sastavnica promijenila značenje tako da značenje frazema nije zbroj značenja sastavnih riječi (djelomična ili potpuna semantička neprozirnost). Primjerice: *labuđi pjev*, *bolja polovica*.

2.1.2. Univerzalna klasifikacija

Univerzalnu klasifikaciju daju Sag et al. (2002), prema (Bauer, 1983), koji u najširem smislu svrstavaju višerječne izraze u dvije skupine: (1) leksikalizirane izraze i (2) institucionalizirane izraze. U prvome nazivu, riječ *leksikalizacija* označava pretvaranje skupine leksema u jedinstvenu leksičku cjelinu, dok u drugome riječ *institucionalizacija* označava uvriježavanje oblika pojavljivanja određenog izraza učestalim korištenjem u jeziku. U daljnjem tekstu dân je pobliži opis obaju kategorija.

Leksikalizirani izrazi Leksikalizirani izrazi leksičke su cjeline specifičnih, ponekad neočekivanih sintaktičkih i/ili semantičkih svojstava koji nemaju nužno poveznica s gramatičkim ponašanjem riječi sastavnica, a za njih je karakterističan određeni stupanj semantičke neprozirnosti. U većine primjera pripadnika ove skupine prisutan je fenomen duha jezika i izvornog govornika – mogu sadržavati riječi koje se u jeziku ne pojavljuju izvan konteksta izraza (strane riječi i riječi-fosile; arhaizme, neologizme i izmišljene riječi). Prema stupnju semantičke strogosti, mogu se dalje raščlaniti (silazno) na (1) stroge izraze, (2) djelomično promjenjive izraze i (3) sintaktički fleksibilne izraze.

Strogi izrazi (engl. *fixed expressions*) Strogi izrazi nepromjenjivi su skupovi riječi koji se ne mogu interpretirati analizom sastavnica, ne dozvoljavaju zamjenu sastavnica sinonimom, ograničeni su u unutarnjim i vanjskim morfosintaktičkim varijacijama, a katkad odstupaju i od gramatičkih pravila jezika (npr. *kud puklo da puklo*). U stroge izraze ubrajaju se i višerječne posuđenice (npr. *in situ*).

Djelomično promjenjivi izrazi (engl. *semi-fixed expressions*) Djelomični izrazi ograničeni su u smislu strogo određenog redoslijeda i nepromjenjivosti sastavnica, ali podržavaju određenu razinu morfološke promjenjivosti. U ovu skupinu ubrajaju se:

- tzv. **nerastavljivi idiomi** (*hodati po jajima*, *naći se na tankom ledu*), kod kojih je morfološka varijacija ograničena na konjugaciju, a interne izmjene (*hodati po krhkim jajima*) i pasivizacija (*biti nađen na tankom ledu*) nisu moguće;
- **imenske skupine** (*zračna luka*, *javni bilježnik*), koje su nalik na nerastavljive idiome po sintaktičkoj nepromjenjivosti, ali toleriraju varijacije u broju;
- **vlastita imena** – imena osoba, institucija, geografskih pojmova i dr. – koja također u pravilu ne podržavaju sintaktičke varijacije, a morfološki su ograničena na sklonidbu imenica i pridjeva (npr. *iz Slavonskog Broda*, *u Hrvatskom zavodu za zdravstveno osiguranje*).

Sintaktički fleksibilni izrazi Iako ovakvi izrazi općenito imaju jedan preferirani oblik pojavljivanja, moguće su varijacije u poretku sastavnica, kao i u izboru samih sastavnica. U ovu skupinu ubrajaju se:

- **rastavljivi idiomi** do određene su razine sintaktički fleksibilni, iako ponašanje ovisi od jednog do drugog slučaja i ne može se sustavno odrediti (npr. *staviti ruku u vatru*);

– **glagolski skupovi** označavaju uvriježene kombinacije imenica i glagola među kojima postoje blage preferencije i velika morfosintaktička fleksibilnost (u vidu poretka riječi, pasivizacije glagola i sl.), a značenje sastavnica je doslovno (u slučaju imenice) i više prilagođeno nego preneseno (u slučaju glagola) (npr. *dati glas*).

Institucionalizirani izrazi S druge strane, institucionalizirani izrazi semantički su potpuno prozirni, ali svojstvena im je statistička specifičnost. Riječ je o frazama koje označavaju pojmove koje bi se s istom sintaksom moglo drugačije označiti, ali samo je jedan specifični oblik postao uvriježen u govoru. Takvi su, primjerice, izrazi *željeznički kolodvor* ili *pješački prijelaz*.

2.1.3. Radna definicija

Za potrebe rada, višerječni izrazi definiraju se kao gramatički smislene jezične jedinice sastavljene od dviju ili više punoznačnih riječi, a od nespecifičnih sintagmi razlikuju se po jedinstvenim, iako međusobno različitim ili čak konfliktnim svojstvima po kojima se mogu razvrstati u nekoliko karakterističnih skupina. Podjela je izrađena kombinacijom gore opisanih stručnih klasifikacija višerječnih izraza, a prilagođena po uzoru na (Delač, 2009):

– **idiomi** – djelomično ili potpuno semantički neprozirni izrazi čije se značenje i kontekst ne mogu predvidjeti iz sastavnih riječi;

– **ustaljeni izrazi** – uvriježene sintagme koje označavaju neki pojam, čije značenje izravno izvire iz sastavnica, no u upotrebi nije uobičajeno bilo koju sastavnicu zamijeniti sinonimom;

– **stručni izrazi** – pojmovi iz stručne terminologije specifični za određeno područje;

– **strani izrazi** – sve fraze preuzete iz drugog jezika, uz izmišljene fraze;

– **vlastita imena** – imena osoba, institucija, geografskih pojmova i dr. sastavljena od dviju ili više riječi.

U tablici 2.1 prikazana je veza između univerzalne klasifikacije višerječnih izraza i radne definicije, uz prikaz dijeljenih svojstava. Kratice korištene u tablici 2.1 pojašnjene su u tablici 2.2.

Tablica 2.1: Klasifikacija višerječnih izraza

	LS			**LI**						**II**
				S	*DP*			*SF*		
	N	O	F		NI	IS	VI	RI	GS	
Idiomi		X	X	X	X			X		
Ustaljeni izrazi		X				X			X	X
Stručni izrazi	X					X			X	X
Strani izrazi				X			X			
Vlastita imena		X					X			

Tablica 2.2: Kratice korištene u tablici 2.1

LS	Leksičke sveze
N	Nazivi
O	Onimi
F	Frazemi
LI	Leksikalizirani izrazi
SI	Strogi izrazi
DP	Djelomično promjenjivi
NI	Nerastavljivi idiomi
IS	Imenske skupine
VI	Vlastita imenica
SF	Sintaktički fleksibilni
RI	Rastavljivi idiomi
GS	Glagolski skupovi
II	Institucionalizirani izrazi

2.2. Srodni radovi

Zadatak prepoznavanja višerječnih izraza dvojak je i razlikuje pojmove identifikacije i ekstrakcije. Kim i Baldwin (2006) opisuju tu razliku definirajući identifikaciju kao zadatak prepoznavanja pojave višerječnih izraza u korpusu tekstova, dok je ekstrakcija zadatak izgradnje leksikona višerječnih izraza temeljem analize individualnih pojava u tekstu.

Rani radovi na području ekstrakcije i identifikacije višerječnih izraza oslanjaju se na statističke metode te pristupaju problemu razmatrajući kolokacijsko ponašanje izraza (statističku preferenciju dviju riječi da se u tekstu pojavljuju zajedno, u odnosu na druge kombinacije). Church i Hanks (1990) razmatraju asocijativnost riječi i problem hipergeneracije kod pristupa slobodnog kombiniranja riječi za koje se potom izračunava asocijativnost. Jedan od prvih razvijenih alata za ekstrakciju kolokacija, Xtract (Smadja, 1993), dodatno u obzir uzima vrste riječi sastavnica i neposredne okoline – u prvom se koraku iz korpusa ekstrahiraju česte dvorječne kolokacije, koje se zatim filtriraju i rangiraju prema frekvenciji kolokacije i njenih sastavnica (ilustriranim histogramima distribucije), međusobnoj poziciji sastavnica i vrstama riječi kontekstualne okoline. Lin (1999) koristi statističke mjere (posebno međusobnu informativnost (engl. *pointwise mutual information*, PMI) kako bi identificirao semantički neprozirne izraze.

Pecina (2010) analizira i uspoređuje 55 statističkih mjera asocijacije te pokazuje da pojedine kombinacije mjera asocijacije daju bolje rezultate nego kada se na istom zadatku koristi samo jedna mjera. Daljnja razmatranja istražuju koje su mjere asocijacije pogodnije za identifikaciju višerječnih izraza, i s kakvim raspoloživim resursima. Petrović et al. (2010) proširuju digramske mjere asocijacije za upotrebu u zadacima vrednovanja n-grama.

Same mjere asocijativnosti ne pokazuju se dovoljnima za rješavanje problema identifikacije višerječnih izraza. Piao et al. (2005) uzima alat za semantičko označavanje i klasifikaciju (USAS) te ga unaprjeđuje statističkim mjerama. Sag et al. (2002) razmatraju jezična svojstva višerječnih izraza i argumentirano izlažu nužnost objedinjavanja statističkih mjera i jezičnih značajki, pokazujući da statistički i pristupi temeljeni na pravilima daju slabije rezultate kada se neovisno primjenjuju.

Uvođenjem sintaktičkih pravila tvorbe višerječnih izraza, zadatak identifikacije i ekstrakcije uz statističke se značajke proširuje i lingvističkim pravilima. Cook et al. (2007) identificiraju idiome promatrajući sintaktičko ponašanje izraza, uz argument da su doslovne kombinacije riječi sintaktički i morfološki manje ograničene od idiomatskih izraza, a Baldwin (2005) koristi, između ostalog, parser zavisnosti kako bi u vrednovanje uključio informaciju o valenciji sintagme. Ljubešić et al. (2015) koriste parser zavisnosti i sintaktičke obrasce višerječnih izraza za izgradnju leksikona potencijalnih višerječnih izraza u hrvatskom, srpskom i slovenskom jeziku. Green et al. (2011) identificira višerječne izraze u francuskom jeziku koristeći kontekstno nezavisnu gramatiku u formi stabla (engl. *tree substitution grammar*, TSG).

Semantičkim svojstvima višerječnih izraza razlikuju se prozirni i djelomično ili

potpuno neprozirni izrazi. Baldwin et al. (2003) koristi latentnu semantičku analizu (engl. *latent semantic analysis*, LSA) da bi usporedio semantički kontekst izraza i semantičke kontekste njegovih sastavnica, a u hrvatskom jeziku to čine Šnajder i Almić (2015).

Za potrebe brze i jednostavne izgradnje visokoodzivnog leksikona višerječnih izraza koji se, uz minimalne intervencije, može koristiti za učenje i vrednovanje kompleksnijih sustava, Tsvetkov i Wintner (2012) demonstriraju izgradnju leksikona promatranjem nesuglasja u rečeničnim elementima paralelnih dvojezičnih korpusa. Isti pristup Ljubešić et al. (2011) primjenjuju na engleski i slovenski jezik.

Koristeći izgrađeni leksikon s kandidatima rangiranim po asocijativnosti kao skup pozitivnih i negativnih primjera za vrednovanje, Tsvetkov i Wintner (2014) razvijaju sustav koji kombinira statistička i jezična svojstva fraza za identifikaciju višerječnih izraza. Pritom za izražavanje veza među značajkama i klasifikaciju primjera koriste model Bayesove mreže, široko prisutne u područjima strojnog učenja i obrade prirodnog jezika. Učenje i vrednovanje sustava vrše na tri jezika: engleskom, kao primjer morfološki siromašnog jezika s bogatim resursima za računalnu lingvistiku; francuskom, morfološki bogatom jeziku također kvalitetnih resurasa; i hebrejskom, morfološki bogatom jeziku oskudnih resurasa. Rad se primarno usredotočuje na razvijanje sustava za hebrejski jezik, uz ograničenje jednostavnih resurasa (pretpostavlja se raspoloživost tokeniziranog i lematiziranog korpusa s oznakama vrsta riječi, bez kompleksnijih alata ili bogatijih baza podataka) i uz izazov morfološki vrlo bogatog i sintaktički fleksibilnog jezika. Na tom se radu uvelike temelji razvoj ovdje opisanog sustava.

3. Sustav za identifikaciju višerječnih izraza

Razvijeni sustav za identifikaciju višerječnih izraza objedinjuje dva samostalna sustava – sustav za izdvajanje kandidata, potencijalnih višerječnih izraza, i sustav za binarnu klasifikaciju kandidata. U prvom potpoglavlju dân je opis vanjskog sustava za izdvajanje kandidata – alata DepMWEx. Drugo potpoglavlje opisuje novoizgrađeni sustav za klasifikaciju, zasebno razmatrajući funkcije leksičkih značajki i Bayesovu mrežu kojom su zavisno povezane.

3.1. Izdvajanje kandidata

Prvi korak postupka identifikacije višerječnih izraza u tekstu jest pretraživanje korpusa tekstova i identificiranje kandidata za klasifikaciju. Taj zadatak obavlja se korištenjem alata DepMWEx (Ljubešić et al., 2015), koji koristi sintaktičke uzorke zavisnosti kako bi u stablima parsanja pronašao kandidate za višerječne izraze.

Kod pristupa ekstrakciji višerječnih izraza korištenjem morfosintaktičkih uzoraka, problem stvara djelomično neograničen redoslijed riječi u izrazima i rečenicama karakterističan za hrvatski i slične jezike, kao i ograničenje koje stvara kompleksnost definiranja morfosintaktičkih uzoraka za nizove dulje od dviju riječi. Iako prijašnji radovi pokazuju da preciznost morfosintaktičke analize slavenskih jezika nadilazi vrijednost od 90%, dok preciznost za sintaktičku analizu dostiže 80%, alat se ipak priklanja analizi sintaktičkim uzorcima, s obzirom na to da su ciljevi (1) izgradnja leksikona kandidata za višerječne izraze u kojem je naglasak na odzivu umjesto na preciznosti i (2) zahvaćanje sintagmi sastavljenih od tri ili više riječi.

Formalno (Jozić, 2013), u hrvatskoj gramatici sastavnice sintagme nazivaju se tagmemima, od kojih je jedan u službi glavnog tagmema (određenica), dok su ostali zavisni (odrednice). Alat koristi ovakvu raščlambu pri definiranju strukture parsera i rezultantnog leksikona. Koristeći formalnu gramatiku i korpus s oznakama zavisnosti,

za svaku pronađenu određenicu (u daljnjem tekstu: ključnu riječ) stvara skup najjačih kolokacijskih pojmova.

Gramatika Gramatika se sastoji od skupa gramatičkih odnosa, opisanih pomoću jednog ili više stabala uzoraka. Primjerice, izraz *graditi kule u oblacima* (radni primjer sintagme ali ujedno i pozitivni primjer višerječnog izraza) opisan je podstablom koje za glavni glagol uzima predikat, uz koji su vezani direktni objekt i prijedložni izraz. Glavni glagol smatra se ključnom riječi, te se ostale sastavnice izraza dodaju u pripadni popis njegovih mogućih kolokacija.

Gramatička relacija koja opisuje dani primjer glasi: `GBZ sbz4 u sbz6`. Slovne oznake predstavljaju vrste riječi (`gbz` za glagol, `sbz` za imenicu, `pbz` za pridjev, `rbz` za prilog), dok brojevne oznake predstavljaju padež (`1` za nominativ, `2` za genitiv, itd.). Velika tiskana slova označavaju određenicu izraza.

Postupak ekstrakcije kandidata vrši se nad svakom parsanom rečenicom u korpusu. Nad rečenicom se obavlja iscrpna pretraga za podstablima koja odgovaraju definiranim uzorcima. Sva zadovoljavajuća podstabla upisuju se u leksikon kao par (podstablo, gramatički odnos). Da bi podstablo bilo kvalificirano za upis kao vjerodostojni kandidat za klasifikaciju, u korpusu se mora pojaviti barem pet puta.

hrMWELex Postupak ekstrakcije kandidata obavljen je nad hrvatskim Web korpusom (hrWaC) (Ljubešić i Erjavec, 2011), čime je nastao leksikon kandidata hrMWELex, koji se koristi kao izvor pozitivnih i negativnih primjera u daljnjem radu. Nad manjim uzorkom leksikona provedena je analiza i utvrđena je preciznost od nešto više od 50%. Odziv nije moguće procijeniti bez iscrpne ljudske pretrage korpusa, ali pretpostavlja se da je zadovoljavajuće visok s obzirom na to da su sintaktičnim uzorcima ekstrahirane sve gramatički ispravne sintagme u korpusu.

U tablici 3.1 dân je uzorak leksikona, iz kojeg je vidljivo da alat jednako zahvaća i općenite sintagme u hrvatskom jeziku, i višerječne izraze na način na koji su ranije definirani. Izvor nepreciznosti u leksikonu prvenstveno izvire iz problema neispravno označene vrste riječi ili neispravno dodijeljene leme kod višeznačnih morfoloških oblika, što kasnije može biti uzrok nepreciznosti pri klasifikaciji kandidata.

Tablica 3.1: Primjer hrMWELex leksikona

Ključna riječ	Sintaksa	Izraz
lakat		
	gbz do SBZ2	
		imati do lakat
		biti do lakat
	pbz0 do SBZ2	
		umočen do lakat
		krvav do lakat
vulkanski		
	PBZ0 sbz0	
		vulkanski erupcija
		vulkanski pepeo
		vulkanski logika
soliti se		
	GBZ sbz4	
		soliti se pamet
	GBZ u sbz4	
		soliti se u stup
uš		
	gbz na SBZ5	
		sjediti na uš
		spavati na uš
		držati na uš
		stajati na uš
	Vez-gbz SBZ5	
		biti uš
	gbz sbz4 u SBZ5	
		imati slušalica u uš

3.2. Klasifikacija kandidata

Drugi korak postupka identifikacije višerječnih izraza jest sustav za klasifikaciju kandidata. Podsustav se sastoji od dva dijela: jezičnih značajki za svaku od kojih svaki kandidat poprima određenu vrijednost i Bayesove mreže koja objedinjuje značajke i izražava povezanost i zavisnost među njima. Motivirane prijašnjim radovima (Sag et al., 2002; Tsvetkov i Wintner, 2014) analizom svojstava višerječnih izraza u hrvatskom i drugim jezicima (engleskom, francuskom i hebrejskom), razvijene jezične značajke kombiniraju statističku i lingvističku analizu kandidata, a čije se rezultantne vrijednosti koriste pri binarnoj klasifikaciji.

Za svaki kandidat za koji se vrši klasifikacija, pretražuje se korpus i analiziraju rečenice u kojima se kandidat pojavljuje, uzimajući u obzir i širi kontekst u značajkama u kojima je to prikladno. Pri traženju pojava kandidata, dopuštena je varijacija u redoslijedu sastavnica i umetanje riječi između sastavnica, a specifičnosti tog ponašanja također su uhvaćene prikladnim značajkama. Očekivani format kandidata koji se predaje značajkama na obradu i pretraživanje uređeni je par (x, y), gdje x predstavlja izraz s potpuno lematiziranim sastavnicama, a y niz oznaka vrsta riječi sastavnica. Primjerice: `([roditi, se, pod, sretna, zvijezda], VPSAN)`

U nastavku potpoglavlja pobliže je opisana motivacija, način rada i očekivani rezultat svake jezične značajke, a razvoj drugog dijela sustava za klasifikaciju opisan je u idućem potpoglavlju.

3.2.1. Jezične značajke

Ukupno trinaest jezičnih značajki razvijeno je za provođenje klasifikacije kandidata. Neke su preuzete iz (Tsvetkov i Wintner, 2014) i prilagođene primjeni nad hrvatskim jezikom, a neke su stvorene po uzoru na svojstva višerječnih izraza u hrvatskome jeziku.

Veliko početno slovo Vlastita imena smatraju se višerječnim izrazima, stoga svaki imenovani entitet u tekstu predstavlja pozitivan primjer. U hrvatskome jeziku različita su pravopisna pravila za veliko početno slovo u vlastitim imenima sastavljenim od više od jedne riječi: *Antun Gustav Matoš*; *New York, Velika Jabuka*; *Ulica Bašćanske ploče*; *Hrvatski zavod za zapošljavanje*.

Značajka `caps` (engl. *capitalisation*), preuzeta iz (Tsvetkov i Wintner, 2014), kao rezultat daje binarni vektor s vrijednošću 1 na i-toj poziciji akko i-ta sastavnica kandidata započinje velikim slovom. Primjerice, *Sveti Petar u Šumi* imat će vrijednost

$[1, 1, 0, 1]$.

Poredba Među ustaljenim frazama u hrvatskome jeziku, česta je upotreba sintagmi u obliku poredbe. Primjerice: *kao riba u vodi*, *plakati kao ljuta godina*. Važno je napomenuti da se višerječnim izrazom ne smatra nužno svaka fraza koja podrazumijeva poredbu, nego ustaljeni izrazi s naglašenim frazeološkim duhom, tj. prenesenim značenjem.

Novouvedena značajka `simile` (engl.) binarna je mjera čija je vrijednost 1 akko kandidat sadrži prijedlog *kao* ili *poput*.

Spojnica Pojedini višerječni izrazi sadrže sastavnice povezane spojnicom. U nekim se slučajevima podjednako koristi oblik izraza sa spojnicom i bez nje (*placebo-efekt* / *placebo efekt*), a u nekima je jedino oblik sa spojnicom pravopisno ispravan (*Smail-aga*, *Požeško-slavonska županija*).

Značajka `hyphen` (engl.), preuzeta iz (Tsvetkov i Wintner, 2014), binarna je značajka čija je vrijednost 1:

1. ako je kandidat za klasifikaciju predstavljen sustavu bez spojnice, a u korpusu je pronađeno barem pet pojava kandidata sa spojnicom, ili

2. ako kandidat u predstavljenom obliku sadrži spojnicu.

Riječ-fosil Pojedini višerječni izrazi sadrže sastavnice koje van konteksta izraza ne nailaze na upotrebu u jeziku. Obično su u pitanju arhaizmi i izmišljene riječi, ili riječi i nazivi preuzeti iz drugih jezika, kao u primjerima: *stani-pani*, *curriculum vitae* ili *Real Madrid*.

Značajka `fossil` (engl.), preuzeta iz (Tsvetkov i Wintner, 2014), kao rezultat daje binarni vektor s vrijednošću 1 na i-toj poziciji akko se i-ta sastavnica kandidata u korpusu pojavljuje samo u kontekstu ostalih sastavnica n-grama kandidata. Tako će za primjer *curriculum vitae* rezultantni vektor biti $[1, 1]$, dok će za *ad hoc* rezultat biti $[0, 1]$, s obzirom na to da se latinski prijedlog *ad* pojavljuje i u kontekstu drugih latinskih izraza ustaljenih u hrvatskome jeziku.

Strana riječ Kao što je već navedeno, brojni višerječni izrazi u hrvatskome jeziku preuzeti su iz drugih jezika ili spadaju u univerzalne leksikone stručnog nazivlja, stoga je pogodno iskoristiti tu statistički značajnu karakteristiku pri klasifikaciji kandidata za višerječne izraze.

Za razliku od značajke riječ-fosil, koja u obzir uzima kontekst riječi-sastavnice, novouvedena značajka `foreign` (engl.) koristi postojeće oznake vrsta riječi u korpusu i kandidatima te vraća binarni vektor s vrijednošću 1 na i-toj poziciji akko je i-ta sastavnica kandidata označena oznakom za stranu riječ, `X`. Zato će kandidat *New York* imati vektor vrijednosti $[1, 1]$, a *jazz koncert* vektor $[1, 0]$.

Sintaktički slijed Brojne su varijante nizova vrsta riječi koje označavaju čest oblik višerječnog izraza; imenica-imenica (*prst sudbine*), pridjev-imenica (*crna ovca*), glagol-prijedlog-imenica (*pucati od zdravlja*), i dr. Iako su sintaktički obrasci ponajprije bitni u koraku ekstrakcije kandidata za klasifikaciju, korisno je zadržati informaciju o sastavu kandidata pri računanju jezičnih značajki, s obzirom na vjerojatnosne ovisnosti između značajke slijeda i ostalih značajki koje razmatraju sintaktičko ponašanje izraza.

Značajka `syntax` (engl. *syntax pattern*), preuzeta iz (Tsvetkov i Wintner, 2014), vraća oznaku niza vrsta riječi od kojih je sačinjen kandidat za višerječni izraz.

Susjednost sastavnica Za neke višerječne izraze karakterističan je strogi oblik koji ne dozvoljava umetanje drugih riječi i rastavljanje sastavnica, a sintaktička je fleksibilnost nešto rjeđa i obično bolji indikator negativnog primjera. Primjerice, u jeziku se može govoriti o *slanom Mrtvom moru*, ali izvan stilski obilježenih pjesničkih tekstova nije ispravno koristiti *Mrtvo slano more*. S druge strane, kod negativnog primjera *široko more*, koji također odgovara sintaktičkom slijedu pridjev-imenica, značenje se ne mijenja u slučaju pojave konteksta *široko i duboko more*.

Novouvedena značajka `adjac` (engl. *adjacency*) uspoređuje broj razdvojenih i nerazdvojenih pojava kandidata u korpusu i vraća vrijednost 1 akko je omjer u korist potonjim. Kod pretrage pojava izraza kandidata, prihvaća se svaka pojava kod koje duljina izraza po broju riječi od prve do posljednje pojave sastavnice iznosi najviše dvostruk broj riječi samog kandidata.

Permutacije sastavnica Kao i u slučaju susjednosti, pojedini višerječni izrazi stroge su forme i nije ispravno mijenjati redoslijed sastavnica u izrazu, dok su rjeđi slučajevi u kojima je moguće mijenjati redoslijed i češće ukazuju na negativni primjer. Primjerice, *hrvatski narodni preporod* ne može se zamijeniti s *narodni hrvatski preporod*, ali *sunčan topao dan* istovrijedan je *toplom sunčanom danu*.

Novouvedena binarna značajka `perm` (engl. *permutation*) vraća vrijednost 1 akko u korpusu postoji više od pet pojava permutiranog oblika kandidata.

Fiksni oblik Višerječni izrazi katkad se javljaju u fiksnom obliku, gdje su jedna ili više sastavnica ograničene u broju morfoloških transformacija kroz koje prolaze u kontekstu izraza, u usporedbi s morfološkim oblicima koje preuzimaju pri samostalnoj pojavi. Primjerice, u izrazu *hodati po jajima*, imenica mora uvijek biti u dativu, ali iako jezična pravila ne brane pojavu u obliku jednine – *hodati po jajetu* – u kontekstu izraza, imenica sastavnica pojavljuje se isključivo u obliku množine.

Značajka `frozen` (engl. *frozen form*), preuzeta iz (Tsvetkov i Wintner, 2014), vraća binarni vektor s vrijednošću 1 na i-toj poziciji akko se i-ta sastavnica kandidata u kontekstu izraza pojavljuje isključivo u jednom morfološkom obliku. Pri tome se u obzir uzima vrsta riječi sastavnice te se analiza morfoloških oblika zanemaruje u slučaju nepromjenjivih vrsta riječi, poput prijedloga i veznika, tako da će izraz *i škare i sukno* imati vrijednost $[0, 1, 0, 1]$.

Djelomična morfološka transformacija Na tragu opažanja iz prijašnje značajke, višerječni izrazi u pojedinim slučajevima poprimaju samo podskup oblika iz potpunog skupa gramatički ispravnih morfoloških transformacija koje određeni niz riječi smije poprimiti. Izraz *zlatno doba* primjer je višerječnog izraza oblika pridjev-imenica koji bi po jezičnim pravilima mogao poprimiti sve morfološke varijacije po broju i padežu. Ipak, u radnom korpusu pojavljuje se pretežno u nominativu i lokativu jednine, dok su svi množinski oblici u potpunosti izostali. S druge strane, negativni primjer istog sintaktičkog oblika, *gradska ulica*, jednoliko je zastupljen u svim oblicima broja i padeža (uz izuzetak vokativa). Iako ova značajka nije podjednako mjerodavna za sve oblike višerječnih izraza, ranije navedena značajka sintaktičkog slijeda i struktura zavisnosti u mreži značajki težinski prilagođava utjecaj značajke morfološke transformacije u slučajevima u kojima je ona uvjetovana nizom vrsta riječi sastavnica.

Po uzoru na (Tsvetkov i Wintner, 2014), svojstvo morfološke transformacije izraza izražava se histogramom distribucije morfoloških oblika kandidata na čitavom korpusu, brojanjem pojava pojedinih oblika i određivanjem njihove frekvencije. Pri tome naglasak nije na specifičnoj morfologiji izraza, nego na uniformnosti transformacija, pa se vektor distribucije gradi samo s obzirom na silazno sortiranu relativnu frekvenciju morfoloških oblika.

Na 0.1%-tnom podskupu korpusa izračunati su prosječni vektori distribucije oblika za uzorak pozitivnih i negativnih primjera:

$$v_{pos} = [69, 16, 6, 3]$$
$$v_{neg} = [50, 19, 10, 6, 3, 2, 1, 1]$$

Iz vektora distribucije vidljivo je da se pozitivni primjeri višerječnih izraza u prosjeku pojavljuju u četiri različita morfološka oblika, dok usporedivi negativni primjeri poprimaju prosječno osam oblika. Viša vrijednost prvog elementa pozitivnog vektora implicira da pozitivni primjeri poprimaju jedan preferirani morfološki oblik, dok završni niz niskih vrijednosti u negativnom vektoru indicira slabiju preferenciju i veću raznolikost slabije zastupljenih oblika.

Pri izračunu značajke `partial` (engl. *partial inflection*), provodi se postupak izračuna vektora distribucije morfoloških oblika kandidata, a potom se računa udaljenost između vektora kandidata i prosječnih vektora, koristeći mjeru kosinusa kuta između dvaju vektora. Značajka vraća vrijednost 1 u slučaju kada je vektor bliži prosjeku za pozitivni primjer.

Semantički kontekst S obzirom na to da višerječni izrazi predstavljaju semantičku cjelinu istovrijednu jednostrukim leksemima, no uže specijalizacije s obzirom na kontekst primjene, može se pretpostaviti da će semantički kontekst višerječnog izraza biti uži i jednoličniji od konteksta proizvoljnog niza riječi iste forme koji u cjelini ne nosi posebno (preneseno) značenje. Po svojoj definiciji, višerječni izrazi također su rjeđe višeznačni od individualnih riječi, zbog čega im je semantički kontekst daljnje ograničen.

Pri razmatranju konteksta višerječnog izraza, u obzir se uzimaju tri sheme vrednovanja koje daju tri vrijednosti mjere:

1. prva najbliža riječ koja slijedi pojavu višerječnog izraza (po uzoru na (Tsvetkov i Wintner, 2014)),
2. zadnja najbliža riječ koja prethodi pojavi višerječnog izraza i
3. okolina od najbližih pet riječi koje prethode i pet koje slijede pojavu višerječnog izraza, uz dodatak svih riječi koje se pojavljuju unutar višerječnog izraza (u slučaju kada dolazi do umetanja riječi i rastavljanja sastavnica).

Radi efikasnijeg izračuna, pri usporedbi konteksta ne koriste se sami leksemi koji sačinjavaju kontekst kandidata, već se izračunava frekvencijski histogram različitih

lema koje se kroz korpus pojavljuju u kontekstu kandidata. Time se svakako gubi dio informacije o kontekstu izraza, no ostaje sačuvan podatak o frekvenciji i distribuciji jedinstvenih lema. Gradi se sortirani vektor frekvencija za kontekst kandidata, uz odstranjivanje svih vrijednosti koje predstavljaju riječi s frekvencijom nižom od 0.1% (negativni primjeri imat će znatno veći broj riječi s vrlo niskom frekvencijom pojavljivanja).

Na 0.1%-tnom podskupu korpusa izračunati su prosječni vektori distribucije svih triju shema konteksta, za uzorak pozitivnih i negativnih primjera. Pri izračunu značajke `context` (engl.), vektor kandidata uspoređuje se s referentnim vektorima koristeći kosinus kuta kao mjeru udaljenosti vektora, a značajka vraća 3-dimenzionalni rezultantni vektor s vrijednošću 1 na i-toj poziciji akko je vektor kandidata za pojedinu shemu vrednovanja bliži referentnom vektoru za pozitivni primjer.

U tablicama 3.2 i 3.3 prikazan je izračun kontekstnog vektora okoline i pripadne oznake značajke za pozitivni (*Olimpijske igre*) i negativni (*isti dan*) primjer iz korpusa, izračunat na uzorku od 50 000 rečenica.

Tablica 3.2: Izračun semantičkog konteksta za pozitivni primjer

Olimpijske igre

Riječ prije	Vektor	Riječ poslije	Vektor	5+5 okolina	Vektor
1 dan	6.25	5 london	31.25	4 velik	3.38
1 igrati	6.25	3 gluh	18.75	3 nositi	2.02
1 ljetni	6.25	1 francuzi	6.25	3 osvojiti	2.02
1 nastup	6.25	1 moskva	6.25	2 olimpijada	2.02
1 otvaranje	6.25	1 nadmetanje	6.25	2 zastava	2.02
(...)	(...)	1 velik	6.25	2 svjetski	2.02
1 zimski	6.25	1 svjetski	6.25	2 sportski	2.02
1 prvenstvo	6.25	1 mlad	6.25	1 zbor	1.35
1 povijest	6.25	1 hrvat	6.25	1 dospjeti	1.35
				(...)	(...)
				1 škola	1.35
				1 adrenalin	1.35
0 - -		- 1 -		- - 1	

rezultat: `[0, 1, 1]`

Tablica 3.3: Izračun semantičkog konteksta za negativni primjer

isti dan

Riječ prije	Vektor	Riječ poslije	Vektor	5+5 okolina	Vektor
1 dogovoriti	4.76	1 centar	4.76	4 godina	1.93
1 godina	4.76	1 dospijevati	4.76	4 mjesec	1.93
1 poslati	4.76	1 izuzetan	4.76	2 dospijevati	0.96
1 račun	4.76	1 mjesto	4.76	2 kreditan	0.96
(...)	(...)	(...)	(...)	2 lopata	0.96
1 uhititi	4.76	1 velik	4.76	2 nalog	0.96
1 susret	4.76	1 vrsta	4.76	2 klub	0.96
1 predavanje	4.76	1 operacija	4.76	2 oglas	0.96
1 pravilo	4.76	1 naknada	4.76	(...)	(...)
				1 prijaviti	0.48
				1 zapremina	0.48
				1 pravilo	0.48
0 - -		- 0 -		- - 0	

rezultat: `[0, 0, 0]`

Semantička neprozirnost Semantička prozirnost – mjera u kojoj se svojstva i značenje višerječnog izraza mogu predvidjeti iz svojstava i značenja sastavnih leksema – odlika je po kojoj se uvelike razlikuju različite skupine višerječnih izraza, no također je značajan indikator pri binarnoj klasifikaciji kandidata za višerječne izraze. Među varirajućim stupnjevima semantičke prozirnosti, potpuna neprozirnost karakteristična je za idiome, a ujedno i njihova definirajuća odlika.

Primjerice, u jednome od dva oblika uporabe, fraza *slijepa ulica* označava ulicu zatvorenu na jednome kraju i u takvom je obliku samo djelomično neprozirna – iako je pridjev *slijepa* prisutan u prenesenom značenju, svakako je riječ o *ulici*, stoga bi se kontekst izraza djelomično trebao preklapati s kontekstom riječi *ulica* u zasebnim pojavljivanjima. S druge strane, u primjerima frazema poput *labuđeg pjeva* ili *gordijskog čvora*, stranom govorniku koji nije intuitivno i iskustveno upućen u specifični leksikon jezika ili govorniku bez potrebnog pozadinskog znanja o zapadnoeuropskoj kulturi te su, potpuno semantički neprozirne fraze, sasvim neshvatljive samo iz promatranja zasebnih značenja i konteksata sastavnih riječi. Za očekivati je da semantički konteksti pridjeva *labuđi* ili imenice *pjev* s kontekstom frazema u cijelosti imaju manje presjeka

i sličnosti od semantičkih konteksata negativnog primjera *pasji lavež* i njegovih sastavnica.

S obzirom na to da semantička neprozirnost ne čini okosnicu opisanog sustava, već jedan njegov ravnopravni dio, radi brzine i efikasnosti izračuna semantika kandidata ne određuje se modelima distribucijske semantike (poput metode opisane u (Šnajder i Almić, 2015)), već se ponovno koristi metoda distribucijskog histograma. Grade se zasebni vektori učestalosti riječi u kontekstu svih pojava fraze u cjelini te sastavnica zasebno, za koje se potom određuje udaljenost (koristeći kosinus kuta između vektora), a za oba konteksta određuje se udio zajedničkih leksema (tj. preklapanje konteksta).

Na 0.1%-tnom podskupu korpusa utvrđene su prosječne vrijednosti za sličnost distribucijskih vektora, kako za pozitivne, tako i za negativne primjere, a za iste primjere utvrđene su i prosječne vrijednosti preklapanja. Kako bi izlazne vrijednosti značajki bile pogodne za korištenje u modelu Bayesovog klasifikatora kojim se kasnije provodi klasifikacija primjere na pozitivne i negativne, sve kontinuirane vrijednosti preslikavaju se u diskretne. U tu svrhu određuju se granični intervali unutar kojih se smještaju tri oznake: (1) niska, (2) srednja i (3) visoka sličnost vektora, odnosno postotak preklapanja. Novouvedene značajke `comp` (engl. *compositionality*) i `overlap` (engl.) vraćaju odgovarajuću vrijednost kodnog vektora ovisno o tome unutar kojeg se od eksperimentalno utvrđenih intervala nalaze izračunate vrijednosti kandidata za sličnost kontekstnih vektora, odnosno postotak preklapanja. Pritom se primjenjuju načela neizrazite logike nad referentnim intervalima, te funkcije mogu poprimiti po jednu od pet mogućih vrijednosti semantičke neprozirnosti: visoka ($[1, 0, 0]$), srednje-visoka ($[1, 1, 0]$), srednja ($[0, 1, 0]$), srednje-niska ($[0, 1, 1]$) i niska ($[0, 0, 1]$).

Asocijativnost Jedna od značajki koja razlikuje višerječne izraze od proizvoljnih nizova riječi bez proširenog značenja jest visina udjela pojave sastavnica u kontekstu višerječnog izraza u odnosu na sveukupnu pojavnost sastavnica u korpusu. Drugim riječima, kada je riječ o pravim višerječnim izrazima, njegove sastavnice imaju veću vjerojatnost pojave u kontekstu izraza (u "društvu" ostalih sastavnica) nego u kontekstu drugih riječi iz korpusa. Snagu veze između dvije ili više riječi može se izraziti mjerama leksičke asocijacije. Počevši od (Manning i Schütze, 1999), razvijene su i definirane raznovrsne mjere asocijativnosti kolokacija, koje se temelje na statističkim, vjerojatnosnim, heurističkim i mjerama iz teorije informacije, a bez proširenja primjenjive su samo na 2-gramske fraze.

Za određivanje mjere asocijativnosti sastavnica kandidata odabran je Sørensen–Diceov koeficijent sličnosti (u daljnjem tekstu: Diceov koeficijent), intuitivna mjera frekven-

cije pojavljivanja odabrana zbog jednostavnosti izračuna i utvrđene pogodnosti za korištenje na manjim korpusima (Och i Ney, 2003). S obzirom na to da je sustav razvijen s namjerom da bude primjenjiv na višerječne izraze proizvoljne duljine (unutar razumnih ograničenja semantičkih mogućnosti jezika), Diceov koeficijent također je pogodan zbog jednostavnog proširenja na n-grame.

Diceov koeficijent izražava se kao:

$$Dice(a, b) = \frac{2f(ab)}{f(a) + f(b)},$$

gdje je $f(ab)$ frekvencija pojavljivanja kandidata 2-grama u korpusu, a $f(a)$ i $f(b)$ frekvencije zasebnih pojavljivanja sastavnih leksema a i b.

Po uzoru na (Petrović et al., 2010), Diceov koeficijent proširuje se na višerječne izraze duljine n temeljnim proširenjem funkcije:

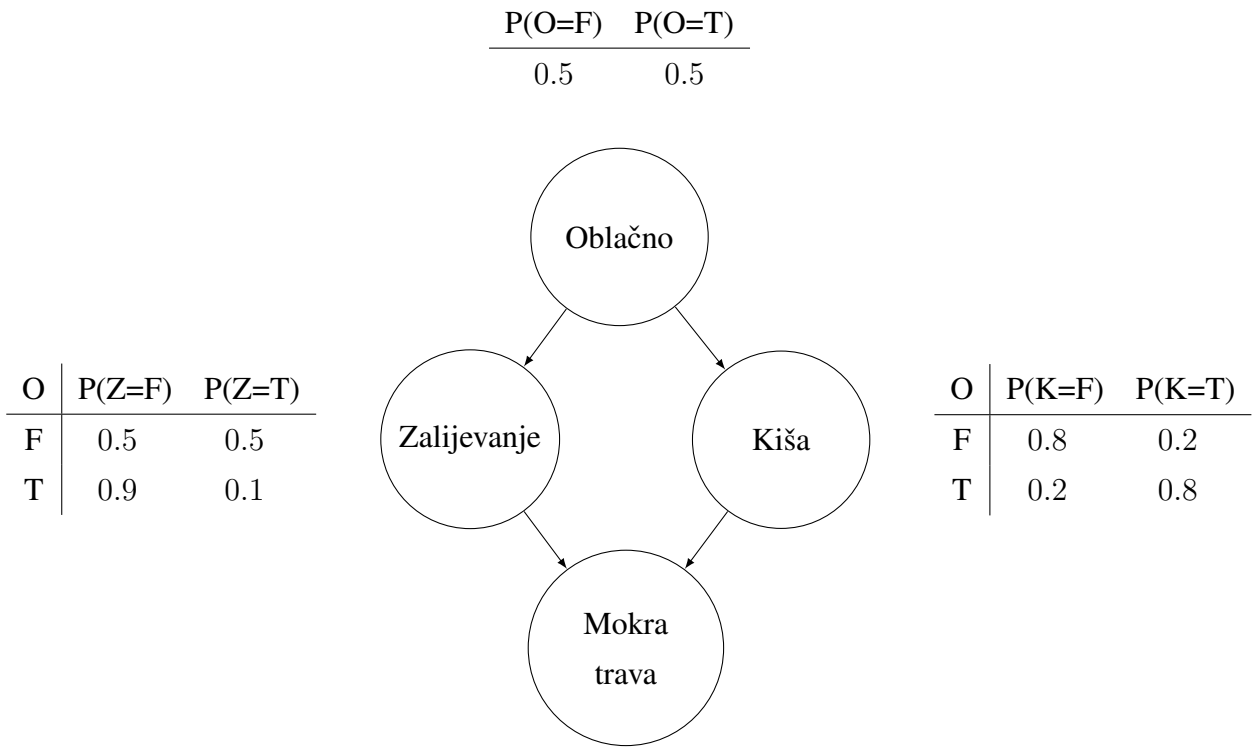
$$Dice_n(w_1, ..., w_n) = \frac{nf(w_1...w_n)}{\sum_{i=1}^{n} f(w_i)}.$$

Na podskupu korpusa računa se prosječna vrijednost Diceovog koeficijenta za uzorak pozitivnih i negativnih primjera. Temeljem tih dvaju vrijednosti, a unutar granica određenih nulom i eksperimentalno utvrđenom nedostižnom maksimalnom vrijednosti (100) određuju se intervali asocijativnosti oko referentnih točaka. Kao i pri izračunu semantičke neprozirnosti, vrijednost asocijativnosti izražava se djelomičnom primjenom neizrazite logike nad referentnim intervalima, a može poprimiti pet mogućih vrijednosti: niska ($[1, 0, 0]$), srednje-niska ($[1, 1, 0]$), srednja ($[0, 1, 0]$), srednje-visoka ($[0, 1, 1]$) i visoka ($[0, 0, 1]$). Značajka `assoc` (engl. *association measure*) kao rezultat vraća pripadni kodni vektor naveden u zagradama.

3.2.2. Bayesova mreža

Bayesova mreža (Jensen, 1996) usmjereni je aciklički graf čiji su čvorovi varijable, a bridovi predstavljaju zavisnosti među tim varijablama. Zavisnosti koje proizlaze iz znanja o području interesa izražavaju se kroz strukturna svojstva mreže, tako da bridovi usmjereni prema određenoj varijabli imaju ishodište u svakom od njenih izravnih uzroka. Takva mreža naziva se kauzalnom mrežom.

Svaki čvor Bayesove mreže povezan je s vjerojatnosnom funkcijom koja prima skup vrijednosti iz njegovih čvorova-roditelja, a vraća vjerojatnost (ili distribuciju vje-



Slika 3.1: Primjer Bayesove mreže

rojatnosti) varijable koju čvor predstavlja. Učenje Bayesove mreže svodi se na izračun zajedničke distribucije vjerojatnosti skupa za učenje, a klasifikacija maksimizira aposteriornu vjerojatnost čvora (tj. varijable) koji se traži – u ovome slučaju, centralni čvor za varijablu MWE (višerječni izraz).

Slika 3.1 prikazuje jednostavan primjer Bayesove mreže, s četiri binarne varijable i uzročnim vezama među njima. Za svaku varijablu definirana je distribucija uvjetnih vjerojatnosti (engl. *Conditional Probability Distribution*, CPD) tj., u slučaju diskretnih varijabli, kao ovdje, tablica uvjetnih vjerojatnosti (engl. *Conditional Probability Table*, CPT), u kojoj su iznesene vjerojatnosti s kojima varijabla može poprimiti određene vrijednosti, u ovisnosti o vrijednostima koje poprimaju roditeljski čvorovi te varijable (ako postoje).

Iz tablice je vidljivo da varijabla *Mokra trava* za vrijednost istine ($M = T$) ima dva moguća uzroka – ili pada kiša ($K = T$), ili je upaljen sustav za zalijevanje ($Z = T$) (ili oboje). U slučaju varijable *Oblačno*, koja nema roditelja, tablica vjerojatnosti određuje apriornu vjerojatnost da jest ili nije oblačno (što slijedno utječe na vjerojatnost da pada kiša ili da je upaljen sustav za zalijevanje).

Primjenom pravila lanca i uzimajući u obzir nezavisnosti među varijablama, zajednička vjerojatnost svih čvorova u mreži izražava se kao:

$$P(O, Z, K, M) = P(O) \cdot P(Z|O) \cdot P(K|O) \cdot P(M|Z, K).$$

Poznavajući vrijednosti čvorova-roditelja, moguće je odrediti vjerojatnost da će varijabla *Mokra trava* poprimiti vrijednost T. Takvo zaključivanje naziva se kauzalnim, ili zaključivanjem odozgora prema dolje (engl. *top-down reasoning*). S druge strane, poznavajući vrijednosti čvorova-djece, moguće je odrediti uzrok takvome stanju – vjerojatnost da je varijabla *Oblačno* poprimila vrijednost T. Takvo zaključivanje naziva se dijagnostičkim, ili odozdo prema gore (engl. *bottom-up reasoning*) i uobičajen je zadatak ekspertnih sustava. Zaključivanje se provodi primjenom Bayesovog pravila:

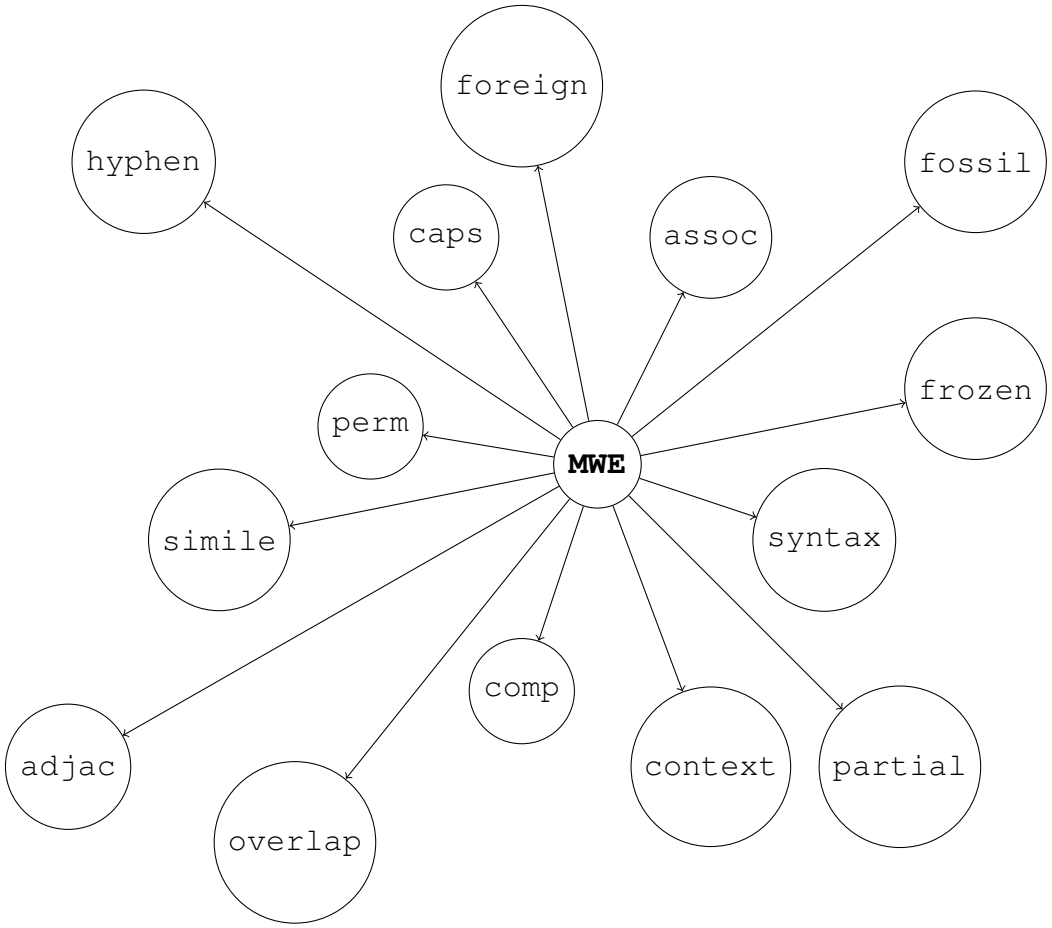
$$P(O_T|X) = \frac{P(X|O_T) \cdot P(X)}{P(X)},$$

gdje je O_T hipoteza da varijabla O poprima vrijednost T, ako je poznato da vrijedi X (gdje je $P(X)$, primjerice, $P(M = T, Z = F, K = T)$). Pri tome se $P(H|X)$ naziva aposteriornom vjerojatnošću (vjerojatnost da hipoteza vrijedi uz poznate činjenice), a $P(X|H)$ izglednošću (koliko su činjenice sukladne s hipotezom). Primjenom MAP-hipoteze (*maximum a posteriori*):

$$h(X) = \underset{O_k}{\operatorname{argmax}}\, p(X|O_k)P(O_k),$$

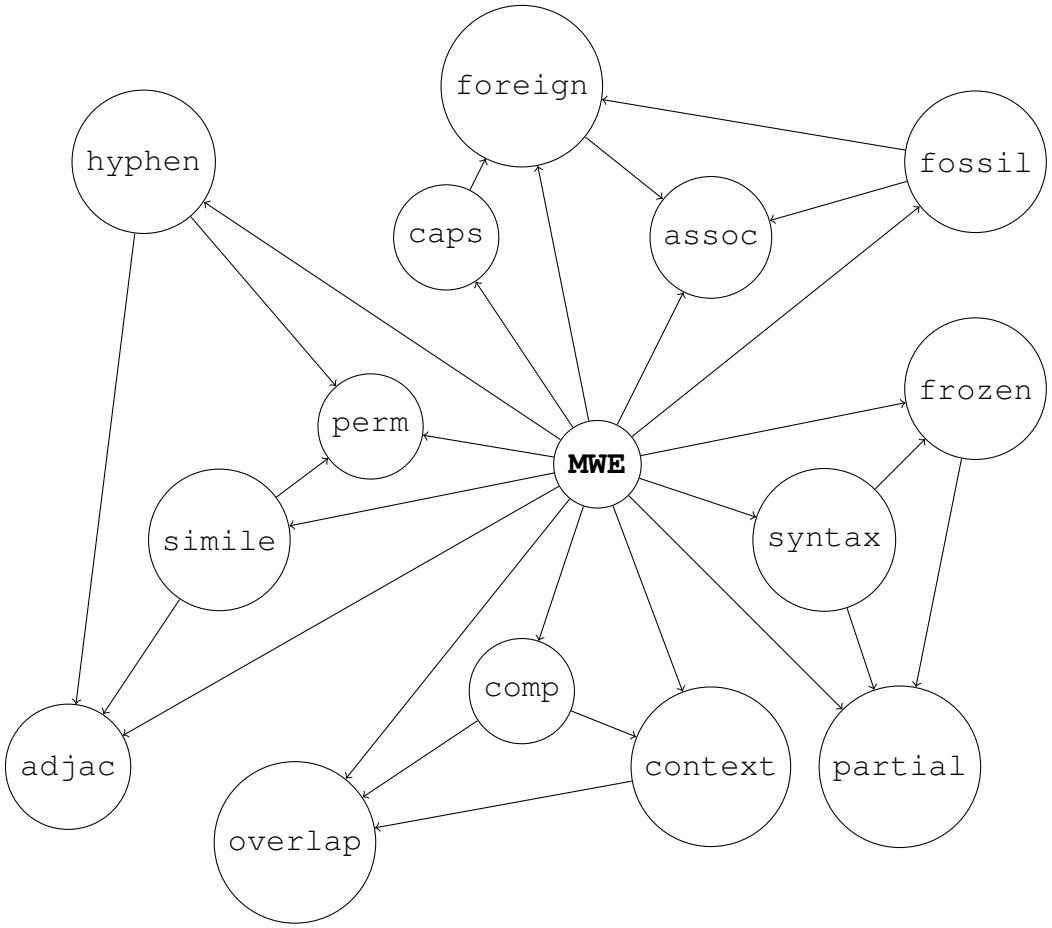
određuje se ona vrijednost k (od mogućih K) za koju tražena varijabla ima najveću vjerojatnost poprimanja s obzirom na poznate vrijednosti ostalih varijabli. Upravo na taj način vrši se binarna klasifikacija u sustavu za identifikaciju višerječnih izraza.

Na slici 3.2 prikazana je shema Bayesove mreže koja objedinjuje opisane jezične značajke. Svaka značajka predstavlja jedan čvor u mreži, uz dodatak čvora `MWE` koji u cijelosti označava zadatak klasifikacije višerječnih izraza. Definiranjem zavisnosti među značajkama gradi se struktura Bayesove mreže koju sustav koristi za klasifikaciju, prikazana na slici 3.3. Zavisnosti među značajkama proizlaze iz intuitivnih



Slika 3.2: Naivni Bayesov klasifikator

zaključaka o zavisnostima između pojedinih opisanih svojstava višerječnih izraza, temeljene na jezičnim pravilima. Konačna shema zavisnosti utvrđena je eksperimentalno. Ručno je izgrađen manji skup uzoraka za učenje i vrednovanje (33 pozitivna i 33 negativna primjera), za koji su vrijednosti značajki izračunate nad podskupom korpusa od 50 000 rečenica. Počevši od jednostavne Bayesove mreže, u kojoj je jedina definirana zavisnost da su svi čvorovi-značajke djeca čvora `MWE`, broj zavisnih veza iterativno je proširivan prema heuristici temeljenoj na utvrđenim pretpostavkama o vezama. Svaka nova instanca mreže vrednovana je na skupu uzoraka dok nije postignut lokalni optimum rezultata vrednovanja. Rezultat ručne izgradnje mreže jest mreža na slici 3.3.

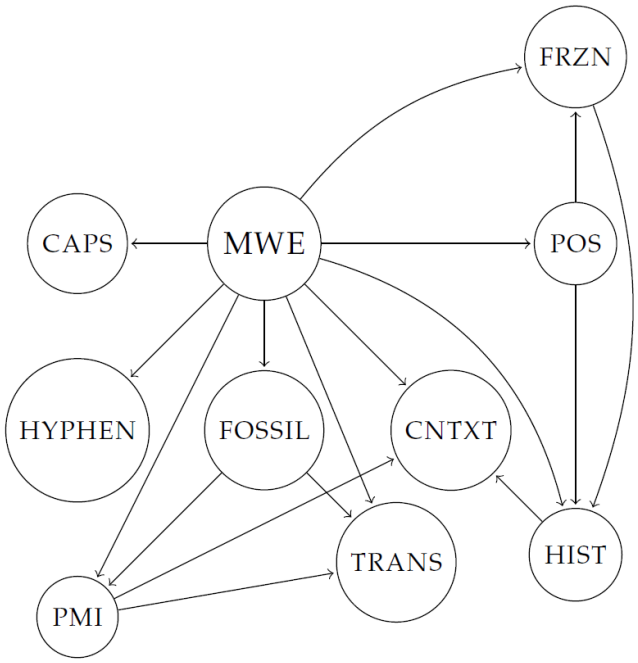


Slika 3.3: Bayesova mreža za klasifikaciju MWE-ova

Za usporedbu, na slici 3.4 prikazana je mreža zavisnosti među značajkama na način definiran i vrednovan u (Tsvetkov i Wintner, 2014). Veze primjenjive na mrežu izgrađenu za potrebe razvijenog sustava preuzete su iz navedene sheme, dok su ostale dodane samostalno, temeljem saznanja o pravilima i ponašanju višerječnih izraza u hrvatskome jeziku.

Svi čvorovi zavise o čvoru `MWE`, s obzirom na to da sve značajke ovise o tome je li primjer višerječni izraz ili nije. Sintaksa izraza utječe na broj morfoloških varijacija koje izraz može poprimiti, stoga su čvorovi `frozen` i `partial` zavisni od čvora `syntax`. Djelomična morfološka transformacija, također zavisi o tome je li koja od sastavnica morfološki nepromjenjiva, stoga postoji i veza od `frozen` prema `partial`.

Kako asocijativnost ovisi o preferenciji riječi za određenim kolokacijama, a takvo je ponašanje izraženo kod stranih i riječi-fosila, koje imaju strogo ograničen skup mogućih kolokacija, `assoc` je zavisan od `fossil` i `foreign`. Nadalje, s obzirom na



Slika 3.4: Mreža zavisnosti (Tsvetkov i Wintner, 2014)

to da se pojam strane riječi i riječi-fosila na ovdje definiran način u pojedinim slučajevima preklapaju, postoji zavisnost od `fossil` prema `foreign`. S obzirom na to da su u hrvatskom jeziku strane riječi najčešće u funkciji imena, `foreign` također zavisi od značajke za velika početna slova, `caps`.

Kako postojanje povlake u izrazu implicira strogi poredak riječi unutar fraze, susjednost (`adjac`) i permutacija (`perm`) u takvim slučajevima ovise o `hyphen`. Izraz u obliku poredbe također utječe na poredak riječi, stoga `perm` i `adjac` zavise i od `simile`.

Na kraju, semantička neprozirnost izraza utječe na općeniti kontekst izraza, stoga se definira zavisnost od `comp` prema `context`, a udio preklapanja riječi u kontekstu zavisi o obje navedene značajke, stoga se određuje i zavisnost od `comp` i `context` prema `overlap`.

Kao što je ranije navedeno, zavisnosti među značajkama određene su heuristički i nisu rezultat temeljitog kombinatoričkog ispitivanja mogućih shema. U postupku izgradnje mreže stvoreno je nekoliko varijanti slične uspješnosti, međusobno različitih u pojedinim vezama ili orijentacijama veza, za koje postoje istovrijedna opravdanja s lingvističkog aspekta. Usprkos određenom nedeterminizmu pri odabiru konačne sheme, opisani korak postupka (osim poboljšanja točnosti klasifikacije uzoraka) koristan je i radi jasnijeg ilustriranja utjecaja jezičnih značajki na određivanje vrijednosti kandidata

za klasifikaciju.

Nakon utvrđivanja strukture mreže, početne tablice zavisnih vjerojatnosti preuzete su iz opisanog manjeg skupa primjera za učenje (66 pozitivnih i negativnih primjera). Opisana mreža dalje je učena i vrednovana na potpunom skupu za učenje i vrednovanje, opisanom u sljedećem poglavlju, uz prikaz rezultata vrednovanja.

4. Eksperimentalno vrednovanje

Nakon izgradnje sustava, za potrebe njegovog učenja i vrednovanja bilo je nužno osigurati korpus tekstova za provedbu analize jezičnih značajki te leksikone potvrđeno pozitivnih i negativnih primjera za učenje sustava. U prvom potpoglavlju opisani su korišteni jezični resursi i uzorci za vrednovanje te metode njihove obrade. Drugo potpoglavlje iznosi rezultate označavanja primjera za učenje. U trećem potpoglavlju izneseni su rezultati vrednovanja sustava, uz diskusiju utjecajnih faktora, promjenjivih parametara i mogućih proširenja.

4.1. Izrada skupova za učenje i vrednovanje

U svim pokusima koristi se fHrWaC (Šnajder et al., 2013), pročišćena verzija hrWaC korpusa (Ljubešić i Erjavec, 2011) – korpusa tekstova dobivenog prikupljanjem tekstualnih podataka s web-stranica koje su dio *.hr* domene. Pri tome se koristi podskup od 500 000 rečenica, tj. 10 067 638 riječi. Korpus je tokeniziran i lematiziran, a svakoj riječi pridružena je oznaka vrste riječi, uz pripadne podatke (npr. rod, broj i padež).

Pozitivni primjeri za učenje i vrednovanje dobiveni su iz "Hrvatskog pravopisa" (Jozić, 2013) i "Školskog rječnika hrvatskoga jezika" (Birtić et al., 2012) Instituta za hrvatski jezik i jezikoslovlje. Iz oba resursa ekstrahirane su sve natuknice predstavljene riječima s bjelinama, a zatim u potpunosti lematizirane (s obzirom na to da sustav kao kandidate prima izraze čija svaka sastavnica mora biti lema) i proširene pridruženim nizom oznaka vrsta riječi sastavnica. Nakon lematizacije, ručno su uklonjeni primjeri koji ne zadovoljavaju radnu definiciju višerječnog izraza (npr. *bilo što*), kao i primjeri koji, iako zadovoljavaju definiciju višerječnog izraza, nisu prikladni jer bi pri automatskom pretraživanju korpusa čak i s maksimalnim pooštravanjem kriterija za oblik i kontekst pojave rezultirali visokim brojem lažno pozitivnih identifikacija (npr. *biti na mjesto – on **je** čovjek **na mjestu***; *automobil je **bio** parkiran **na mjestu** zabranjenog zaustavljanja*).

Nakon uređivanja i pročišćavanja liste izraza, određena je frekvencija pojavljiva-

nja svakog izraza na odabranom podskupu korpusa kako se vrijeme pretraživanja ne bi trošilo na primjere s manje od 10 pojavljivanja. Od 4672 primjera iz leksikografskih rječnika, 3100 primjera odbačeno je jer na čitavom korpusu (50 940 598 rečenica) broje 10 ili manje pojavljivanja, a na radnom uzorku korpusa (pola milijuna (500 000) rečenica) ne pojavljuju se niti jednom. Daljnjih 1194 primjera odbačeno je jer se na istom uzorku pojavljuju manje od 10 puta. Preostalih 138 najfrekventnijih uzoraka prihvaća se, a njihova frekvencija pojavljivanja iznosi između 10 i 156 pojavljivanja. Primjeri su sortirani silazno prema učestalosti pojavljivanja. Sastav primjera prema klasifikaciji iz potpoglavlja 2.1.3. dân je u tablici 4.1.

Tablica 4.1: Sastav pozitivnih primjera

Vrsta izraza	Broj primjera	Udio (%)
Idiomi	39	28.26
Ustaljeni izrazi	30	21.74
Stručni izrazi	19	13.77
Strani izrazi	43	31.16
Vlastita imena	7	5.07

Negativni primjeri za učenje i vrednovanje, uz dodatne pozitivne primjere, dobiveni su iz opisanog hrMWELex leksikona. S obzirom na to da je hrMWELex skup neoznačenih pozitivnih i negativnih primjera, podatke je bilo potrebno ručno označiti, stoga je sastavljen skup primjera za označavanje.

Za izraze iz hrMWELex-a unaprijed je izračunata mjera asocijativnosti na hrWaC korpusu. Izrazi su po vrijednosti asocijativnosti podijeljeni u četiri kategorije (visoka, umjereno visoka, umjereno niska i niska), te je iz svake kategorije odabrano 2000 na radnom korpusu najfrekventnijih primjera. Jednolikim nasumičnim odabirom iz svake kategorije sastavljen je uzorak od 4098 primjera za označavanje. Uzorak je nadopunjen kontrolnim skupom od 127 nasumično odabranih pozitivnih primjera iz ranije sastavljenog uzorka pozitivnih primjera, čija je svrha provjera dosljednosti u radu označivača.

S obzirom na to da su svi primjeri iz hrMWELex-a u potpunosti lematizirani, za potrebe lakše čitljivosti primjera bilo je potrebno primjerima pridružiti morfološki transformiran oblik izraza kakav bi se pojavio u tekstu ili govoru. Rekonstrukcija izraza izvršena je pretraživanjem korpusa za najfrekventnijim oblikom izraza, uz dodatak preferiranog odabira pri brojanju pojava transformacija (za imenice i pridjeve: očekivani padež (ako je DepMWEx oznaka za padež različita od "bilo koji"), dok su glagoli

ostajali u tzv. rječničkom infinitivu). U tablici 4.2 navedeni su primjeri automatske rekonstrukcije.

Tablica 4.2: Primjer rekonstrukcije morfoloških transformacija

lisan uš	lisna uš
rad škola	rad škole
broj bod	broj bodova
sav odluka	sve odluke
isti mjera	istom mjerom
biti u soba	biti u sobi
ugledan član	uglednog člana
pravo građanin	prava građana
poklopiti sebe uš	poklopiti se ušima

Postupak obrade rezultirao je skupom od 4225 uzoraka za označavanje, danih u lematiziranom i gramatički ispravno morfološki transformiranom obliku. Uz pretpostavku da skup sačinjava uzorak od negativnih i pozitivnih primjera, u nepoznatom omjeru u korist negativnih, skup je predan na ručno označavanje binarnim oznakama.

4.2. Rezultati označavanja podataka

Uzorak za učenje i vrednovanje označavalo je troje studenata: *A*, *B* i *C*. Svakom je studentu dan skup od 4225 uzoraka, sačinjen od 4098 nepoznatih primjera i 127 nasumično umetnutih kontrolnih pozitivnih primjera (u daljnjem tekstu: pozitivna kontrola), bez znanja studenta o takvoj raspodjeli. Radi minimizacije utjecaja okolnih primjera na ocjenu pojedinog primjera, kao i efekta zamora, svaki student na označavanje je dobio primjere u različitom redoslijedu. Svaki uzorak bilo je potrebno označiti oznakom `0` (nije višerječni izraz) ili `1` (jest višerječni izraz). Detaljne upute za označavanje koje je dobio svaki student prikazane su u Dodatku A. U prosjeku, svaki student na označavanju je radio 3 sata, a ukupno vrijeme rada iznosi 9 sati.

Za utvrđivanje kvalitete ručne klasifikacije i suglasnosti među označivačima, koristi se Cohenov kappa-koeficijent (Cohen, 1968), mjera suglasnosti koja umanjuje utjecaj slučajnog slaganja meču označivačima, tj. u obzir uzima tendenciju pojedinog označivača da u graničnim slučajevima preferira određenu oznaku.

Kappa-koeficijent mjeri slaganje između dva označivača koja klasificiraju N primjera u C kategorija, a izražava se kao:

Tablica 4.3: Parametri kappa-koeficijenta

	B_+	B_-	
A_+	p_{11}	p_{12}	$p_{1\cdot}$
A_-	p_{21}	p_{22}	$p_{2\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	n

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

gdje je p_o omjer slaganja među označivačima, a p_e vjerojatnost slučajnog slaganja. Slučajno slaganje računa se prema izrazu:

$$p_e = p_{1\cdot}p_{\cdot 1} + p_{2\cdot}p_{\cdot 2},$$

gdje vrijednosti $p_{1\cdot}$, $p_{\cdot 1}$, $p_{2\cdot}$ i $p_{\cdot 2}$ predstavljaju vjerojatnost da pojedini označivač svrsta primjer u određenu kategoriju, na način prikazan u tablici 4.3. Tablica ilustrira slučaj u kojem dva označivača, A i B, klasificiraju $N = n$ primjera u $C = 2$ kategorije – pozitivni primjeri $(+)$ i negativni primjeri $(-)$.

Definirana su tri intervala vrijednosti kappa-koeficijenta koja opisuju razinu suglasnosti označivača (Green, 1997):

1. $\kappa < 0.4$ – niska razina suglasnosti,
2. $0.4 \leqslant \kappa \leqslant 0.7$ – srednja razina suglasnosti i
3. $\kappa > 0.7$ – visoka razina suglasnosti.

Po završetku označavanja, utvrđena je mjera suglasnosti među označivačima. U tablici 4.4 navedeni su kappa-koeficijenti za ukupni skup uzoraka za označavanje (nepoznati uzorci i kontrolni skup), a tablica 4.5 prikazuje kappa-koeficijente za označavanje skupa nepoznatih uzoraka, ne računajući kontrolne primjere.

Najviša je postignuta suglasnost između označivača A i B, koja iznosi $\kappa = 0.51$ nad čitavim skupom, a $\kappa = 0.43$ u slučaju nepoznatih uzoraka bez kontrolne skupine. Prema ranije navedenim intervalima, vrijednost se nalazi blizu donje granice srednje razine suglasnosti, iz čega se može zaključiti da je zadatak raspoznavanja višerječnih izraza izrazito težak za neprofesionalne označivače, a čak i uz definicije nekoliko vrsta

Tablica 4.4: Kappa-koeficijent za skup s kontrolom

$\kappa(x,y)$	A	B	C
A	-	**0.51**	0.19
B	0.51	-	0.15
C	0.19	0.15	-

Tablica 4.5: Kappa-koeficijent za skup bez kontrole

$\kappa(x,y)$	A	B	C
A	-	**0.43**	0.15
B	0.43	-	0.12
C	0.15	0.12	-

višerječnih izraza, u skupu uzoraka još uvijek velik broj primjera spada u "granične slučajeve".

Detaljniji pregled rezultata označavanja dân je u tablici 4.6. Tablica prikazuje ukupan broj primjera oko čije su oznake sva tri označivača bila suglasna i brojčanu raspodjelu primjera u kojima je došlo do neslaganja među označivačima. Radi kompaktnijeg izričaja, vrsta neslaganja izražena je binarnim kôdom koji ukratko ilustrira oznaku po označivaču – primjerice, stavak "Odstupanje `011`" označava sve primjere koje je prvi označivač (A) označio negativnima, dok su ih ostali označivači (B i C) označili pozitivnima.

Iz tablice suglasnosti uočljivo je da broj većinski pozitivnih primjera (tj. primjera oko čije se pozitivne klasifikacije jedan označivač ne slaže) nadilazi ukupan broj suglasno pozitivnih primjera, dok je suglasno negativnih primjera više od većinski negativnih primjera (s jednim odstupanjem). Izvor ovakve raspodjele djelomično leži u sastavu hrMWELex leksikona (koji po eksperimentima na manjim uzorcima većinski sadrži negativne primjere), a djelomično u činjenici da je označivačima lakše bilo prepoznati negativan primjer, tj. jednostavnije je argumentirati zašto je neki izraz negativni primjer višerječnog izraza, nego raspoznati granični slučaj pozitivnog primjera.

Analizom rezultata označavanja po duljini izraza, nije uočena korelacija između broja riječi u izrazima i broja nesuglasnih oznaka, a slično vrijedi i za vrste riječi sastavnica. Jedina značajnija korelacija uočena je kod općesuglasno negativnih primjera, među kojima prevladavaju nizovi od četiri i više riječi, no to je očekivano, s obzirom na metodu sastavljanja hrMWELex leksikona i predviđen omjer 4-gramskih višerječnih izraza u odnosu na slobodno sastavljene, gramatički ispravne fraze iste duljine.

Tablica 4.6: Rezultati označavanja uzoraka

			Ukupno	**Ukupno prihvaćeno**
Pozitivno			192	192
	Odstupanje `011`	146		
	Odstupanje `101`	200	380	188
	Odstupanje `110`	34		
			572	**342**
Negativno			1906	1906
	Odstupanje `100`	52		
	Odstupanje `010`	71	1620	1636
	Odstupanje `001`	1497		
			3526	**3542**
Pozitivna kontrola			81	-
	Odstupanje `011`	3		
	Odstupanje `101`	22	46	-
	Odstupanje `110`	21		
			127	-

Usprkos visokom neslaganju među označivačima, oznake kod primjera nesuglasja svejedno su uzete u obzir pri mogućem razrješavanju nesuglasnih oznaka. U najdesnijem stupcu tablice navedeno je koliko je graničnih primjera ipak prihvaćeno kao pozitivni ili negativni, autorskim nadjačavanjem oznaka označivača gdje je došlo do nesuglasnosti. U slučaju pozitivnih primjera, prihvaćeni su primjeri koji po definiciji zadovoljavaju kriterije višerječnog izraza, a odstupanje u oznakama je najizvjesnije zamor ili pad koncentracije označivača – takvih je primjera ukupno 188. Od preostalih primjera, 48 je identificirano kao negativni te su pridruženi skupu pouzdano negativnih primjera. Ostali dvosmisleni primjeri izbačeni su iz uzorka kako ne bi unosili šum u postupak učenja i vrednovanja.

Među negativnim i pretežno negativnim primjerima, primjeri kod kojih je postojalo nesuglasje u označavanju većinom su prihvaćeni kao negativni te je, uz dodatak 48 pogrešno definiranih dvosmisleno pozitivnih primjera, skup negativnih primjera proširen s ukupno 1636 primjera. Preostali primjeri odbačeni su zbog dvosmislenosti.

Konačno, postupak označavanja rezultirao je skupom uzoraka za učenje i vrednovanje od ukupno 3884 primjera, od čega su 342 pozitivna (8.81%), a 3542 negativna (91.19%). Taj će se uzorak, uz proširenje pozitivnim primjerima iz leksikografskih

rječnika, koristiti u postupku učenja i vrednovanja sustava.

Kako postotak jednoglasno označenih primjera na čitavom uzorku, a time i maksimalno slaganje među označivačima, iznosi 51.6%, zadatak je potvrđeno težak za ljudskog označivača, a pretpostavljivo i za računalni sustav. Ipak, s obzirom na to da je takva točnost u klasifikaciji na razini sa slučajnim dodjeljivanjem oznaka, te uzevši u obzir stručnu neprilagođenost označivača i ponaosobno opterećenje količinom uzoraka za označavanje, pretpostavlja se da je ovakav rezultat ipak nešto pesimističniji od očekivane točnosti za sustav klasificiranja.

4.3. Postupak vrednovanja

U ovom potpoglavlju opisan je čitav postupak vrednovanja sustava. Prvi odjeljak opisuje korištene resurse i uzorak primjera za vrednovanje. U drugom odjeljku prikazan je izračun mjera koje se koriste za opisivanje učinkovitosti sustava. Rezultati učenja i vrednovanja prikazani su u trećem odjeljku, a u posljednjem odjeljku provodi se diskusija mogućih izmjena i proširenja sustava.

4.3.1. Korpus i skup za vrednovanje

Nakon provođenja postupka prikupljanja, obrade i označavanja pozitivnih i negativnih primjera višerječnih izraza, dobiven je skup od 342 pozitivna primjera dobivena označavanjem, 3542 negativna primjera dobivena označavanjem i 138 pozitivna primjera odabrana iz leksikografskih rječnika.

Uklanjanjem duplikata pozitivnih izraza presjekom skupa primjera iz rječnika i skupa pozitivnih primjera dobivenih označavanjem, gube se 94 primjera, te konačan skup pozitivnih primjera iznosi 386 primjera sa zadovoljavajućom zastupljenošću na korpusu. Demonstrativni uzorak priložen je u Dodatku B.

Od dobivenih pozitivnih primjera, odabran je uzorak od 79 najfrekventnijih primjera, a istom metodom odabran je simetričan uzorak iz skupa negativnih primjera. Dobiveni skup pozitivnih i negativnih primjera korišten je za učenje i vrednovanje sustava.

4.3.2. Mjere učinkovitosti sustava

Osnovne mjere empirijskog utvrđivanja učinkovitosti sustava jesu preciznost, odziv, točnost i F-mjera (Makhoul et al., 1999). Izračun navedenih mjera izražen je kao od-

nos broja ispravno i neispravno klasificiranih primjera; točnije, ispravno klasificiranih pozitivnih primjera (engl. *true positive*, tp), ispravno klasificiranih negativnih primjera (engl. *true negative*, tn), neispravno klasificiranih pozitivnih primjera (engl. *false negative*, fn) i neispravno klasificiranih negativnih primjera (engl. *false positive*, fp).

Točnost Točnost (engl. *accuracy*, A) statistička je mjera koja opisuje koliko vjerno sustav za binarnu klasifikaciju klasificira primjere, a izražava se kao udio točno klasificiranih primjera u skupu svih primjera:

$$A = \frac{tp + tn}{tp + tn + fp + fn}.$$

Preciznost i odziv Preciznost (engl. *precision*, P) i odziv (engl. *recall*, R) mjere su kvalitete i kvantitete rezultata klasifikatora, koje ukazuju na to koliko je klasifikator koristan, odnosno koliko je potpun skup njegovih rezultata. Preciznost se definira kao udio točno klasificiranih pozitivnih primjera u skupu svih pozitivno klasificiranih primjera:

$$P = \frac{tp}{tp + fp}.$$

Preciznost služi kao mjerilo za pogrešku zamjene i umetanja (proširivanja skupa pozitivno klasificiranih uzoraka netočno klasificiranim uzorcima) te ilustrira koliko su mjerodavne oznake klasifikatora. S druge strane, odziv se definira kao udio točno klasificiranih pozitivnih primjera u skupu svih pozitivnih primjera:

$$R = \frac{tp}{tp + fn},$$

a služi kao mjerilo za pogrešku zamjene i uklanjanja (izostavljanja pozitivnih uzoraka iz skupa pozitivno klasificiranih uzoraka). Drugim riječima, odziv ilustrira koliko je potpun skup pozitivno klasificiranih dokumenata.

U sustavima za identifikaciju višerječnih izraza često postoji obrnuto proporcionalna veza između preciznosti i odziva, tj. porastom preciznosti pada odziv i obrnuto.

F-mjera F-mjera (engl. F-*score*, F_β) srednja je vrijednost preciznosti i odziva, a općenito se izražava kao:

$$F_\beta = (1 + \beta) \cdot \frac{P \cdot R}{(\beta^2 \cdot P) + R},$$

pri čemu je $\beta \in \mathbb{R}_{>0}$. Parametar β utječe na mjeru efektivnosti sustava s gledišta prema kojem je odziv β puta važniji od preciznosti. Kao uravnoteženi pokazatelj učinkovitosti, u vrednovanju sustava korištena je F-mjera kao harmonijska sredina preciznosti i odziva, uz parametar $\beta = 1$:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}.$$

4.3.3. Rezultati vrednovanja

Vrednovanje sustava provedeno je s uzorcima opisanima u potpoglavlju 4.1. Nakon izračuna vrijednosti svih parametara koji predstavljaju značajke za pojedini uzorak, stvorena je i vrednovana Bayesova mreža pomoću programske okoline za strojno učenje Weka (Hall et al., 2009). Za vrednovanje je korištena metoda 10-struke unakrsne provjere, pri čemu se skup primjera dijeli na 10 podskupova. Postupak učenja i vrednovanja ponavlja se 10 puta, na način da se u svakoj iteraciji 9 podskupova koristi za učenje mreže (u ovom slučaju, podešavanje tablica uvjetnih vjerojatnosti), a preostali podskup za vrednovanje, uz slijedno pomicanje ispitnog podskupa kroz iteracije. U tablici 4.7 izneseni su rezultati vrednovanja.

Sustav je najprije vrednovan na osnovnoj shemi Bayesove mreže, tj. naivnom Bayesovom klasifikatoru, koji na mjestima čvorova sadrži iste značajke kao ciljni sustav, no struktura mreže ograničena je na roditelj-dijete odnos čvora `MWE` sa svim ostalim čvorovima. Zatim je za klasifikaciju korištena shema Bayesove mreže s ručno modeliranim zavisnostima među značajkama, kao što je ilustrirano na slici 3.3. Nad obje mreže provedeni su eksperimenti s apriornim vjerojatnostima za čvor `MWE`, s varijacijama u intervalu od 0.35 do 0.55, no nisu uočene značajnije razlike u F-mjeri, stoga je zadržana apriorna vjerojatnost $P(X_{MWE}) = 0.5$, proizašla iz skupa za učenje s jednakim brojem pozitivnih i negativnih primjera.

Kao referentni model binarne klasifikacije višerječnih izraza, na istom skupu podataka vrednovan je model koji za donošenje odluke razmatra samo mjeru leksičke asocijacije (Diceov koeficijent). Pri tome je korišten algoritam grupiranja k srednjih vrijednosti (engl. *k-means*).

Tablica 4.7: Rezultati vrednovanja sustava (10-struka unakrsna provjera)

	Točnost (%)	Preciznost	Odziv	F_1-mjera
k-means	70.25	0.721	0.703	0.696
NBK	80.52	0.761	**0.897**	0.824
BM	**83.54**	**0.827**	0.848	**0.837**

Tablica 4.8: Rezultati vrednovanja usporedivih sustava (10-struka unakrsna provjera)

Jezik	Točnost (%)	F_1-mjera
Hebrejski	76.82	0.77
Francuski	79.04	0.778
Engleski	83.52	0.835
Hrvatski	83.54	0.837

Iz rezultata je vidljivo poboljšanje točnosti sustava nakon uvođenja zavisnosti među čvorovima. S porastom točnosti uočen je i porast u preciznosti, no uz to dolazi do manjeg pada odziva. F_1-mjera usporediva je s rezultatima koje su istom metodom postigli Tsvetkov i Wintner (2014), izneseni u tablici 4.8.

Analizom pogrešaka klasifikacije pozitivnih primjera po definiranim vrstama višerječnih izraza, ustanovljeno je da sustav najčešće griješi u slučaju ustaljenih izraza (51.85% lažno negativnih klasifikacija); primjerice, kod izraza *održivi razvoj* ili *bala sijena*. Nešto je rjeđa pogrešna klasifikacija za stručne izraze (*žuta pjega*) i nazive (*Hrvatske vode*). Svi primjeri stranih izraza točno su klasificirani. Udio svake vrste višerječnih izraza u skupu pogrešno klasificiranih izraza dân je u tablici 4.9.

S obzirom na rezultate dobivene u usporedivim radovima te na indikaciju težine zadatka danu pri ručnom označavanju podataka, rezultati vrednovanja sustava nadilaze očekivanja. Moguć utjecaj na rezultate vrednovanja imao je manji skup pomno oda-

Tablica 4.9: Lažno negativni primjeri po vrstama izraza)

Vrsta izraza	Udio (%)
Idiomi	7.4
Ustaljeni izrazi	**51.85**
Stručni izrazi	18.51
Strani izrazi	0.0
Vlastita imena	22.22

branih uzoraka za vrednovanje (uzorci s apsolutnim slaganjem označivača i frekventni na radnom uzorku korpusa), stoga bi vrijedilo vrednovanje ponoviti na većem skupu podataka radi provjere utjecaja šuma u uzorcima za učenje i vrednovanje.

4.3.4. Moguća proširenja

Postojeći sustav ostavlja prostora za razne izmjene i proširenja čiji bi utjecaj na efektivnost sustava bilo zanimljivo vrednovati. Primjerice, postojeći sustav za akviziciju kandidata ekstenzivan je i daje leksikon kandidata s visokim odzivom, no s obzirom na nisku preciznost, iziskuje značajan ljudski napor u označavanju pozitivnih i negativnih primjera. Alternativni pristup ili dodatak istog u postupku, primjerice akvizicijom kandidata iz paralelnih korpusa, kao u (Tsvetkov i Wintner, 2012) ili (Fišer et al., 2011), mogao bi pojednostaviti postupak ili poboljšati uzorak za učenje i vrednovanje.

U kontekstu morfosintaktičkog parsera, preciznost u identifikaciji pojava kandidata pri izračunu vrijednosti mjera jezičnih značajki moglo bi poboljšati proširenje sintaktičkih uzoraka informacijom o morfologiji sastavnica; točnije, zadržavanje informacije o preferiranom padežu i sl., na način na koji je to primijenjeno u parseru DepMWEx. Naravno, u slučaju izgradnje skupa za učenje iz leksikografskih resurasa, obrada takvih podataka zahtijevala bi nešto intenzivniji posao od jednostavnog pripisivanja oznake vrste riječi svakoj sastavnici.

Slično, pri identifikaciji pojava kandidata, u trenutnoj inačici sustava koristi se sintaktički širok izraz, koji pri pojavi kandidata tolerira promjenu redoslijeda riječi u rečenici i umetanje riječi koje nisu dio izraza, u toj mjeri da ukupan broj riječi između prve i zadnje pronađene sastavnice ne prelazi dvostruku duljinu samoga izraza. Permutacije i umetanje nisu univerzalno dozvoljene kod višerječnih izraza, no prihvaćene su pri pretraživanju pojava uz pretpostavku da će pripadne značajke za permutaciju i umetanje donekle korigirati šum nastao takvim ponašanjem kod lažnih primjera pojava. Sustav bi se, stoga, mogao proširiti u vidu definiranja sintaktičkih uzoraka kod kojih jest ili nije prihvatljiva permutacija, odnosno umetanje, kada je riječ o pozitivnom primjeru višerječnog izraza, te sukladno tome koristiti odgovarajuća pravila pri identificiranju pojava kandidata u korpusu. Takvo proširenje iziskivalo bi preliminarno istraživanje jezičnih pravila i ponašanja izraza u korpusu.

Od nedostataka u postupku učenja i vrednovanja sustava koji su mogli imati značajniji utjecaj na konačne rezultate vrednovanja, svakako se ističe problem označavanja pozitivnih i negativnih primjera iz leksikona kandidata. Pokazano je da je to zadatak koji iziskuje stručno znanje s područja višerječnih izraza, te bi vrednovanje sustava

valjalo ponoviti s većim skupom uzoraka oko kojih je postignuta značajnija suglasnost označivača.

Na broj uzoraka također je utjecala veličina radnog korpusa i njihova distribucija na istom, koja je bila ograničena vremenom i količinom izračuna za jezične značajke. Osim veličine korpusa, na broj pojava kandidata i kvalitetu pronađenih pojava utjecao je i sastav samog korpusa, koji je s jezičnog aspekta ograničen. Iako novinski članci, komercijalne objave i neformalni govor sadrže bogatstvo višerječnih izraza, velik broj kvalitetnih primjera izgubljen je zbog toga što su karakteristični za stručni ili književni govor i rjeđe se pojavljuju van toga konteksta. Osim proširenja opsega radnog korpusa, pri ponovljenim pokusima poželjno bi bilo proširiti korpus dokumentima iz stručne literature i književnosti. Korpus stručnih i znanstvenih tekstova za hrvatski jezik postoji,[1] no trenutno među resursima za hrvatski jezik ne postoji korpus književnih tekstova u javnome vlasništvu na kojem bi bilo moguće slobodno eksperimentirati.

[1] `http://hrcak.srce.hr`

5. Zaključak

Cilj ovoga rada bio je razviti i vrednovati sustav za identifikaciju višerječnih izraza u hrvatskome jeziku. Temeljem istraživanja dosadašnjih radova na području identifikacije višerječnih izraza, sustav se temelji na kombinaciji jezičnih značajki koje objedinjuju statističke mjere i lingvistička pravila, a u izgradnju i rad sustava implementirani su i već postojeći resursi i alati za obradu hrvatskog jezika – lematizator i označivač vrsta riječi, morfosintaktički parser i automatski generirani leksikoni.

U opisu razvoja sustava prikazan je problem identifikacije višerječnih izraza i dana radna definicija višerječnih izraza po karakterističnim skupinama. Predstavljene su uvedene jezične značajke, uz pojašnjenje motivacije svake značajke, primjenu i način izračunavanja rezultantnih vrijednosti. Opisan je postupak obrade podataka i izgradnje skupa uzoraka za vrednovanje, od obrade jezikoslovnih izvora do postupka ručnog označavanja podataka i konsolidacije rezultata. Na kraju, prikazana je primjena Bayesove mreže na rješavanje problema objedinjavanja značajki i uspješnost takvog pristupa pri identifikaciji višerječnih izraza.

Vrednovanjem je ispitan utjecaj odabira apriorne vjerojatnosti klasifikacijske oznake, proširivanja skupa značajki i intuitivnog određivanja zavisnosti među njima. Dobiveni rezultati zadovoljavajući su s obzirom na kompleksnost problema, pretpostavljenu i utvrđenu analizom rezultata ručnog označavanja, te uzevši u obzir oskudnost uzoraka za učenje i vrednovanje.

Poboljšanja su moguća ponajprije na problemu izgradnje većeg i kvalitetnijeg skupa uzoraka za učenje i vrednovanje, što bi podrazumijevalo stručnije ručno označavanje i veći opseg izračuna vrijednosti parametara jezičnih značajki za svaki uzorak. Također, uzevši u obzir karakteristike pojedinih vrsta višerječnih izraza, veću frekvenciju pojavljivanja pojedinih skupina izraza, a stoga i vjerodostojnije rezultate, moglo bi se očekivati proširenjem korpusa tekstova znanstvenom literaturom i književnim tekstovima.

LITERATURA

Timothy Baldwin. Deep lexical acquisition of verb–particle constructions. *Computer Speech & Language*, 19(4):398–414, 2005.

Timothy Baldwin i Su Nam Kim. Multiword expressions. *Handbook of natural language processing*, 2:267–292, 2010.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, i Dominic Widdows. An empirical model of multiword expression decomposability. U *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, stranice 89–96. Association for Computational Linguistics, 2003.

Goranka Blagus Bartolec. Kolokacijske sveze prema drugim leksičkim svezama u hrvatskom jeziku. *Fluminensia*, 24(2), 2012.

Laurie Bauer. *English word-formation*. Cambridge university press, 1983.

Matea Birtić, Goranka Blagus Bartolec, Lana Hudeček, Ljiljana Jojić, Barbara Kovačević, Kristian Lewis, Ivana Matas Ivanković, Milica Mihaljević, Irena Miloš, Ermina Ramadanović, i Domagoj Vidović. *Školski rječnik hrvatskoga jezika*. Institut za hrvatski jezik i jezikoslovlje, 2012.

Kenneth Ward Church i Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

Paul Cook, Afsaneh Fazly, i Suzanne Stevenson. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. U *Proceedings of the workshop on a broader perspective on multiword expressions*, stranice 41–48. Association for Computational Linguistics, 2007.

Davor Delač. Postupci ekstrakcije kolokacija iz zbirki tekstova. 2009.

Darja Fišer, Špela Vintar, Nikola Ljubešić, i Senja Pollak. Building and using comparable corpora for domain-specific bilingual lexicon extraction. U *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, stranice 19–26. Association for Computational Linguistics, 2011.

Annette M Green. Kappa statistics for multiple raters using categorical classifications. U *Proceedings of the 22nd annual SAS User Group International conference*, svezak 2, stranica 4. San Diego, 1997.

Spence Green, Marie-Catherine De Marneffe, John Bauer, i Christopher D Manning. Multiword expression identification with tree substitution grammars: A parsing tour de force with french. U *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, stranice 725–735. Association for Computational Linguistics, 2011.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, i Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

Ray Jackendoff. *The architecture of the language faculty*. Broj 28. MIT Press, 1997.

Finn V Jensen. *An introduction to Bayesian networks*, svezak 210. UCL press London, 1996.

Željko Jozić. *Hrvatski pravopis*. Institut za hrvatski jezik i jezikoslovlje, 2013.

Su Nam Kim i Timothy Baldwin. Automatic identification of english verb particle constructions using linguistic features. U *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*, stranice 65–72. Association for Computational Linguistics, 2006.

Dekang Lin. Automatic identification of non-compositional phrases. U *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, stranice 317–324. Association for Computational Linguistics, 1999.

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

Nikola Ljubešić, Darja Fišer, Špela Vintar, i Senja Pollak. Bilingual lexicon extraction from comparable corpora: A comparative study. U *First International Workshop on Lexical Resources*, stranica 48, 2011.

Nikola Ljubešić, Kaja Dobrovoljc, i Darja Fišer. Mwelex - mwe lexica of croatian, slovene and serbian extracted from parsed corpora. *Informatica*, 39(3):293, 2015.

John Makhoul, Francis Kubala, Richard Schwartz, Ralph Weischedel, et al. Performance measures for information extraction. U *Proceedings of DARPA broadcast news workshop*, stranice 249–252, 1999.

Christopher D Manning i Hinrich Schütze. *Foundations of statistical natural language processing*, svezak 999. MIT Press, 1999.

Antica Menac. *Hrvatska frazeologija*. Knjigra, 2010.

Franz Josef Och i Hermann Ney. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51, 2003.

Pavel Pecina. Lexical association measures and collocation extraction. *Language resources and evaluation*, 44(1-2):137–158, 2010.

Saša Petrović, Jan Šnajder, i Bojana Dalbelo Bašić. Extending lexical association measures for collocation extraction. *Computer Speech & Language*, 24(2):383–394, 2010.

Scott Songlin Piao, Paul Rayson, Dawn Archer, i Tony McEnery. Comparing and combining a semantic tagger and a statistical tool for mwe extraction. *Computer Speech & Language*, 19(4):378–397, 2005.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, i Dan Flickinger. Multiword expressions: A pain in the neck for nlp. U *Computational Linguistics and Intelligent Text Processing*, stranice 1–15. Springer, 2002.

Frank Smadja. Retrieving collocations from text: Xtract. *Computational linguistics*, 19(1):143–177, 1993.

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.

Yulia Tsvetkov i Shuly Wintner. Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, 18(04):549–573, 2012.

Yulia Tsvetkov i Shuly Wintner. Identification of multiword expressions by combining multiple linguistic information sources. *Computational Linguistics*, 40(2):449–468, 2014.

Dodatak A
Upute za označavanje

Dokument sadrži pozitivne i negativne primjere višerječnih izraza u hrvatskome jeziku, odabrane prema učestalosti supojavljivanja. Svaki redak sadrži jedan izraz, najprije u potpuno lematiziranom obliku, a potom u rječničkom infinitivu s deklinirani imenicama. Potrebno je označiti koji primjeri jesu višerječni izrazi (oznaka **1** na početku retka), a koji nisu (oznaka **0** na početku retka).

Više je vrsta i definicija višerječnih izraza; u ovom zadatku, pozitivnim primjerima smatraju se sve fraze koje imaju barem jednu od sljedećih značajki:

- **neprozirnost (idiom)** – Pojam koji fraza označava ne tumači se doslovno iz riječi sastavnica, nego u prenesenom značenju. Primjerice: *sjediti na ušima* jest višerječni izraz; *sjediti na klupi* nije.
- **ustaljeni izraz** – Uobičajena fraza koja označava određeni pojam. Značenje je jasno iz riječi sastavnica, no nije uobičajeno da se koja od njih zamijeni sinonimom. Primjerice: umjesto *zračne luke* ne može se reći *avionska luka*; ali umjesto *bivšeg vlasnika* može se reći *prijašnji vlasnik*.
- **stručni izraz** – Pojmovi iz stručne terminologije. Primjerice: *ugljična kiselina*, *plućno krilo*, *parna turbina*.
- **strani izraz** – Fraza preuzeta iz drugog jezika. Riječi sastavnice ne pojavljuju se same u hrvatskom jeziku. Primjerice: *lingua franca*, *ad hoc*.
- **vlastito ime** – Imena osoba, geografskih pojmova, institucija i sl. jesu višerječni izrazi. Primjerice: *Miroslav Krleža*, *Jadransko more*, *Hrvatski zavod za javno zdravstvo*. Takve primjere treba označiti kao pozitivne, neovisno o nedostatku velikih početnih slova u dokumentu.

Dokument sadrži primjere višerječnih izraza u kojima redoslijed riječi sastavnica može biti drugačiji od uobičajenog (npr. "*vlastite bojati se sjene*"). Gdje je moguće, takve primjere valja označiti kao pozitivne.

Dokument također sadrži primjere koji pojedine višerječne izraze zahvaćaju samo djelomično, ili u širem kontekstu (npr. "*zavod za zaštitu*" [*na radu*], ili "pregovori s *europskom unijom*"). Takve primjere valja označiti kao negativne.

Ukupno, dokument sadrži 4225 primjera fraza, a predviđeno je vrijeme označavanja dva sata. U nastavku je dano nekoliko primjera s oznakama:

```
1 carski rez              carski rez
0 obrazovan građanin      obrazovani građani
0 redovito koristiti sebe     redovito koristiti se
1 boriti sebe s vjetrenjača   boriti se s vjetrenjačama
```

Zadatak nije jednostavan, ali pomaže voditi se mišlju o duhu jezika – bi li strani govornik mogao upotrijebiti pojedinu frazu bez poznavanja njenog uobičajenog konteksta? Je li neku frazu moguće prevesti bez korištenja frazeološkog leksikona? Za pomoć je također dozvoljeno služiti se rječnicima, internetom i ostalim izvorima znanja. Primjerice, ako Hrvatski jezični portal[1] pojedinu frazu navodi pod sintagmom ili frazeologijom riječi sastavnice, to je pouzdan indikator da je riječ o višerječnom izrazu.

[1] `http://hjp.znanje.hr`

Dodatak B
Pozitivni primjeri

U ovome dodatku dân je uzorak od 133 pozitivna primjera višerječnih izraza iz kojih je odabran podskup za učenje i vrednovanje sustava. Izrazi su dobiveni iz leksikografskih izvora i ručnog označavanja automatski ekstrahiranih kandidata za višerječne izraze, a odabrani su po frekvenciji pojavljivanja na radnom podskupu korpusa.

Primjeri su navedeni u obliku u kojem se predaju programu: niz lematiziranih sastavnica izraza, za kojima slijede oznake vrsta riječi.

```
čovjek na mjesto     N S N
čvrst forma      A N
širok brijeg      A N
štitast uš      A N
što prije      P S
žeđati duša teba      N V N
žut pjega      A N
bala slama      N N
biograd na more      N S N
biti na cijena      V S N
biti na put      V S N
biti na raspolaganje      V S N
biti na teret      V S N
biti od riječ      V S N
biti pri sebe      V S P
biti pri svijest      V S N
biti svoj čovjek      V P N
biti u drugi plan      V S A N
biti u forma      V S N
```

biti u funkcija V S N
biti u krug V S N
biti u prvi plan V S P N
bolonjski proces A N
bosna i hercegovina N C N
cetinski krajina A N
civilan društvo A N
crven križ A N
crven vrag A N
dan branitelj N N
dati do znanje V S N
dati sav od sebe V A S P
dionički društvo A N
doći na ideja V S N
doći na svoj V S P
dobar volja A N
dobro proći R V
domovinski rat A N
društvo književnik N N
drugi svjetski rat A A N
dug otok A N
europski komisija A N
europski unija A N
filipinski pojas A N
fiziološki fimoza A N
gorski kotar A N
grad država N N
grad zagreb N N
groban mrak A N
gustoća naseljenost N N
hrvatski branitelj A N
hrvatski sabor A N
hrvatski voda A N
hrvatski vojska A N
igran film A N
imati as u rukav V N S N

imati smisao V N
imati veza V N
istjerati vrag V N
izazivati zbrka u glava V N S N
južan koreja A N
južnoafrički republika A N
jutarnji list A N
katolički crkva A N
klasičan glazba A N
ključan riječ A N
križni put A N
kupaći kostim A N
lisan uš A N
los angeles N N
matičan stanica A N
ministarstvo branitelj N N
na čelo S N
na trag S N
nadmorski visina A N
nastavan plan A N
nba liga X N
new york X X
nov era A N
nov godina A N
nov list A N
nov zagreb A N
nov zeland A N
održiv razvoj A N
olimpijski igra A N
otac domovina N N
paški čipka A N
paški sir A N
paklen ponor A N
paran valjak A N
park priroda N N
petar krešimir N N

planinarski društvo A N
početi tući srce V V N
pravi put A N
prijediplomski studij A N
pun podrška A N
pun pogodak A N
pun ruka A N
put u život N S N
ravan kotar A N
real madrid X N
registar branitelj N N
republika hrvatska N N
s obzir na S N S
sav svijet P N
sjedinjen američki država A A N
slavonski brod A N
slobodan dalmacija A N
sovjetski savez A N
stanica iz pupkovina N S N
star grad A N
stati pred oltar V S N
stjepan radić N N
svet otac A N
svet rok A N
svoj čovjek P N
trgovački društvo A N
u životan opasnost S A N
u drugi plan S A N
u prvi plan S P N
u sjena S N
ujedinjen narod A N
uzlazan vod A N
varaždinski županija A N
varaždinski biskupija A N
večernji list A N
velik britanija A N

velik gorica A N
zabavan program A N
zagrebački županija A N
zaviriti iza zavjesa V S N
značajan krajobraz A N
zračan luka A N

Identifikacija višerječnih izraza zasnovana na kombinaciji jezičnih značajki

Sažetak

Višerječni izrazi čine značajan udio vokabulara prirodnoga jezika, no zbog specifičnosti i nepredvidivosti obilježja, iziskuju posebnu pažnju pri razvoju sustava za automatsku identifikaciju izraza u sklopu računalne obrade prirodnog jezika.

U sklopu ovoga rada, razvijen je i predstavljen sustav za identifikaciju višerječnih izraza u hrvatskome jeziku koji iz korpusa tekstova ekstrahira i klasificira potencijalne višerječne izraze kombinacijom statističkih mjera i lingvističkih značajki specifičnih za višerječne izraze. Opisan je izračun vrijednosti značajki, struktura Bayesove mreže za klasifikaciju i rezultati vrednovanja u ovisnosti o određivanju zavisnosti među značajkama.

Ključne riječi: Obrada prirodnog jezika, višerječni izrazi, Bayesova mreža, hrvatski jezik.

Multiword Identification Based on the Combination of Linguistic Features

Abstract

Multiword expressions constitute a significant portion of any natural language vocabulary, but due to their characteristic idiosyncrasy, MWEs call for particular dedication in the development of applications for automatic identification, within the scope of automated natural language processing.

In this thesis, we develop and present a multiword expression identification system that extracts and classifies potential MWEs from a corpus of Croatian text documents through a combination of statistical measures and linguistic features specific to MWEs. We describe the computation of feature values and the structure of the Bayesian network used in classification, and present evaluation results relative to different dependency relations between features.

Keywords: Natural language processing, multiword expressions, Bayesian networks.