Laboratorij za tehnologije znanja (KTLab)
Zavod za elektroniku, mikrolektroniku, računalne i inteligentne sustave
Fakultet elektrotehnike i računarstva
Sveučilište u Zagrebu

Interni dokument

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 292

Automatsko generiranje parafraza izraza i rečenica hrvatskoga jezika

Nikola Šantić

Zagreb, lipanj 2011.

SADRŽAJ

1. Uvod

Jedan od najvećih izazova strojne obrade prirodnog jezika je povezivanje očitanih simbola s njihovim značenjem. Kao dio ovog važnog zadatka, prepoznavanje različitih načina da se iskaže ista informacija iznimno je složen problem. Ukoliko se radi o kratkim izrazima kao što su riječi (primjerice, *stablo* i *drvo*), govorimo o *sinonimiji*. Ukoliko su izrazi duži i složeniji (primjerice "William Shakespeare je napisao 'Hamleta'" i "Autor 'Hamleta' je William Shakespeare") radi se o *parafraziranju*.

Međutim, različite riječi i fraze iz nekog razloga i jesu različite. Rijetko se može reći da dva različita izraza sadrže potpuno identičnu informaciju – najčešće se radi o gotovo identičnoj ili sličnoj informaciji. Ponekad riječi imaju više značenja, a samo se dio tih značenja među njima poklapa. Ponekad su informacije iz više fraza sažete u jednu. Eksplicitne granice između semantičke ekvivalentnosti, sličnosti i nepovezanosti nemoguće je odrediti. Ipak, postoje konvencije prema kojima klasificiramo povezanost izraza.

Sinonime možemo podijeliti na *istoznačnice*, riječi međusobno zamjenjive u svim kontesktima (npr. *ljekarna – apoteka*) i *bliskoznačnice*, koje su zamjenjive samo u određenim kontekstima (npr. *supruga – žena*). Najveći problem pri određivanju sinonima, bilo automatski ili ručno, predstavljaju upravo bliskoznačnice, odnosno određivanje kada su dvije riječi sinonimi, a kada samo srodni pojmovi. Osim sinonima, često se nailazi i na druge semantičke relacije kojima su riječi povezane. One su:

- **Antonimi** odnosno leksemi suprotnoga značenja, npr. *otvoriti – zatvoriti*;
- **Hiperonimi** i **hiponimi** predstavljaju odnos nadređenog i podređenog pojma, npr. *jabuka* je hiponim od hiperonima *voće*. Dva hiponima istog hiperonima su kohiponimi, npr. *plava* i *crvena* kohiponimi su hiperonima *boja*;
- **Meronimi** i **holonimi** predstavljaju odnos dviju riječi od kojih je jedna dio neke cjeline, a druga je ta cjelina, npr. *kora*, *grana* i *list* meronimi su holonima *stablo*;

Relacije između parafraza još su složenije jer, uz sinonimiju, mogu sadržavati i

mnoštvo drugih parafrastičkih metoda. Prema najširoj podjeli, parafraza se s obzirom na diskurz pojavljuje u četiri oblika (Bagić, 2007):

- **Lingvistička parafraza** – variranje ishodišnog iskaza jezičnim mehanizmima. Na razini sintakse može biti sinonimska supstitucija (npr. *Uhićena dvojica pljačkaša* i *Uhapšena dvojica pljačkaša*) ili odnos između pasivnog i aktivnog oblika iste rečenice, (npr. *Plaža je obasjana suncem* i *Sunce je obasjalo plažu*). Gramatička konverzija u kojoj se naglašava jedna perspektiva također se smatra lingvističkom parafrazom, npr. *Ivan je prodao auto Marku* i *Marko je kupio auto od Ivana*. Drugi često korišteni oblici su promjena poretka riječi i zamjena denotacije konotacijom.
- **Komentatorska parafraza** – parafraziranje u funkciji kritike, analize i interpretacije nekog iskaza ili teksta. Ona proizlazi iz materijala o kojem govori dijelom ga preuzimajući, a dijelom mijenjajući. Parafastičar objašnjava stručne pojmove, razrješuje dvosmislenosti i općenito prilagođava jedan iskaz drugim. Ovakav "miješani tekst" može biti proširenje (amplifikacija) ili sažimanje (kondenzacija) izvornog teksta.
- **Literarna parafraza** – oblikovni postupak preuzimanja određenog literarnog elementa ili forme. Može biti proces prevođenja lirskog oblika u prozni, ili prevođenje iz jednog metričkog registra u drugi. Oblicima literarne parafraze mogu se smatrati i različite vrste imitacije poput *pastiša*, koji se temelje na oponašanju stila, teme ili formalnih obilježja. Najčešće mete pastiša su dobro poznati i kanonizirani tekstovi.
- **Ludička parafraza** – stilska figura slična pastišu, ali realizirana na kraćim iskazima kao referenca na prepoznatljiv naslov, frazeologizam ili poslovicu. Osnovni cilj ovakve parafraze je upotrijebiti jezični klišej kako bi se izrazila nova informacija. Pri tome se konstrukcija korištene fraze zadržava, a mijenja se uvijek isti, najčešće semantički najznačajniji dio. Često ovakav tip parafraze obrće značenje originalnog iskaza, npr. *Tko rano rani, il' je pekar il' je budala* ili *Tko tebe kamenom, ti njega tvrdim kruhom!*.

Rad s parafrazama u sklopu obrade prirodnog jezika najčešće podrazumijeva prva dva oblika, s obzirom da oni zadržavaju dozu semantičke ekvivalentnosti. Druga dva oblika se zasnivaju upravo na semantičkom obratu, pa su u jednom aspektu mogu shvatiti kao antonimi parafraze. I jedni i drugi se često pojavljuju u sličnom kontekstu i dodatno otežavaju pronalaženje pozitivnih parafraza i sinonima. Stroža definicija može zahtjevati da su parafraze kao dijelovi rečenice zamjenjive. To znači da se parafrazirani izrazi moraju moći zamijeniti u svojim kontekstima, a da rečenica ostane gramatički

ispravna. Fraze u kurzivu u rečenicama 1.1 i 1.2 nisu stroge parafraze jer bi njihova zamjena stvorila gramatički neispravne rečenice.

Teslin izum izmjenične struje omogućio je elektrifikaciju svijeta i industrijsku revoluciju. (1.1)

Tesla je izumio izmjeničnue struju, omogućivši elektrifikaciju svijeta i industrijsku revoluciju. (1.2)

Stroge parafraze rijetke su u svakodnevnom pisanom jeziku. Stoga ću ovom radu biti razmatrana šira definicija parafraze, pritom tolerirajući eventualnu razliku u informacijama koja postoji između dvaju izraza.

Pojam usko povezan s parafraziranjem je logički slijed (engl. *entailment*). Za izraz prirodnog jezika T kažemo da implicira H ako bi osoba koja pročita i vjeruje u istinitost T zaključila da je i H istinit. Za razliku od parafraziranja, logički slijed nije simetrična relacija – za imenske fraze X i Y, 1.3 implicira 1.4, ali ne i obrnuto (moguće je, na primjer, da je X naslikao Y). Parafraza se pritom može shvatiti kao logički slijed koja vrijedi u oba smjera – u oba se izraza nalazi jednaka informacija. Prepoznavanje logičkog slijeda je utoliko složenije jer zahtjeva semantičku analizu, dok se parafraza može postići i samom obradom sintaksnog stabla.

X je napisao Y. (1.3)

Autor Y-a je X. (1.4)

Parafraze su u obradi prirodnog jezika našle višestruku primjenu. Velik broj metoda za rad s parafrazama osmišljen je u sklopu sustava za odgovaranje na pitanja (engl. question answering). U sustavima koji koriste zbirke dokumenata, problem predstavljaju pitanja postavljena oblikom različitim od onog kojim je zapisan odgovor. Harabagiu i Hickl (2006) pokazali su da uzimanje u obzir varijacija dobivenih parafraziranjem drastično poboljšava izvedbu ovakvog sustava. Sustav za odgovaranje može dohvatiti sve dokumente vezane za ključne riječi pitanja 1.5 te među dobivenim tekstom 1.6 tražiti parafraze pitanja kojem je upitna zamjenica "tko" zamijenjena svim imenovanim entitetima prepoznatim u 1.6.

Tko je napisao "Kralja Leara"? (1.5)

Kralj Lear jedno je od najpoznatijih djela Williama Shakespearea, a zasnovano je na kralju Learu, jednom od mitoloških kraljeva Britanije, o kome prvi zapisi potječu iz 12. stoljeća, koje je ostavio povjesničar Geoffrey od Monmoutha. U 18. i 19. stoljeću zbog kritika tragične i tmurne radnje, drama je romantizirana, tako što je kraj u kojem originalno umire većina glavnih likova, zamijenjen krajem u kojem svi likovi prežive, a Cordelia se udaje za Edgara.[...] (1.6)

Lear/William Shakespeare/Geoffrey od Mnomutha/Cordelia/Edgar je napisao "Kralja Leara". (1.7)

Postavljeno pitanje se također može parafrazirati kako bi se sakupili i drugi relevantni dokumenti. Druga česta upotreba parafraze pitanja je u slučaju posjedovanja liste odgovora na često

postavljana pitanja (FAQ – Frequently asked questions) pri čemu moramo pronaći pitanje koje je najsličnije korisničkom upitu (Tomuro, 2003).

U metodama automatskog sažimanja teksta (engl. *text summarization*), važan dio uključuje ekstrakciju rečenica koje najbolje predstavljaju dani tekst. Ukoliko se radi sažimanje tekstova iz više dokumenata u jedan sažetak (Barzilay i McKeown, 2005), bitno je izbjeći odabir rečenica koje govore o istoj stvari, odnosno sadrže istu informaciju kao i već ekstrahirane rečenice. Ovaj se problem može izbjeći koristeći metode za prepoznavanje parafraza na razini rečenice ili fraze. Sažimanje rečenice (engl. *sentence compression*) često je dio sustava za sažimanje teksta, a može se shvatiti kao poseban slučaj parafraziranja (Zhao et al., 2009). Uvjet je da nastala rečenica mora biti kraća od izvorne, ali i dalje gramatički ispravna. Većina sustava za sažimanje dopušta da se iz izvorne rečenice izbace manje bitne informacije, pri čemu dolazi u pitanje koliko široko definiramo pojam parafraze. Najčešće je potreban dodatni mehanizam koji rangira kandidate prema količini mjesta kojeg uštede i vrijednosti informacije koja je sačuvana.

Sustavi za crpljenje informacija često se oslanjaju na ručno ili automatski stvorene obrasce za pronalaženje dijelova teksta u kojima se opisuje određeni događaj. Obrascima se identificiraju entiteti uključeni u taj događaj, kao što se vidi na primjeru obrazaca 1.8, 1.9 i 1.10. Obrasci mogu biti i kompleksniji, zasnovani na sintaksnim stablima, uključivati oznake vrste riječi. Parafraziranjem obrazaca moguće je pronaći još više dijelova teksta semantički ekvivalentnog sadržaja (Shinyama i Sekine, 2003).

X je raznesen (1.8)

Bomba je eksplodirala blizu X (1.9)

Eksplozija je uništila X (1.10)

Pri automatskoj evaluaciji strojnog prevođenja, prijevod se uspoređuje s ljudskim prijevodom koji može koristiti drugačiji raspored i izbor riječi, stoga se za ispravniju ocjenu koriste metode za prepoznavanje parafraza (Zhou et al., 2006; Kauchak i Barzilay, 2006). Metode parafraziranja koriste se i za obogaćivanje korpusa za učenje (Zhang i Yamamoto, 2005). Ukoliko se sustav susretne s frazom koju do sada nije sreo, generiranjem parafraza može pokušati stvoriti oblik koji poznaje. Pritom je moguće da će se dio informacije izgubiti, ali dobiveni rezultati su bolji onih dobivenih direktnim prevođenjem.

Parafraziranje ima svoju upotrebu i u strojnom generiranju jezika. Koristi se za obogaćivanje jezika, izbjegavanje ponavljanja istih fraza, povećavanja čitljivosti i stila teksta, ili ispunjenje nekih specifičnih ograničenja. Među ostalim primjerima korištenja su pojednostavljivanje teksta (primjerice zamjena stručnih medicinskih izraza jednostavnijim pojmovima) (Deléger i Zweigenbaum, 2009), automatsko ocjenjivanje studentskih odgovora usporedbom s rješenjima (Nielsen et al., 2009) ili čak lingvistička steganografija (Chang i Clark, 2010).

U ovom radu predstavljene su i evaluirane dvije metode generiranja parafraza na hrvatskom jeziku. Jedna će se bazirati na statističkoj zamjeni fraza njihovim sinonimima, pri čemu će svaki n-gram imati određenu vjerojatnost da se parafrazira. Druga metoda će koristiti tehnike prepoznavanja parafrazi te iz tako prepoznatih parova prikupljati pravila parafraziranja. Obje metode će za korpus koristiti usporedivi jednojezični korpus hrvatskih novinskih portala. Obje će metode biti usporedno i zajednički evaluirane na uzorku nasumično odabranih rečenica iz iste domene.

Nastavak rada sljedeće je strukture: u poglavlju 2 predstavljeni su radovi koji se bave sličnom problematikom. Radovi su podijeljeni prema tri najčešća cilja: prepoznavanju, prikupljanju i generiranju parafraza. U poglavlju 3 posebnu pažnju obraćamo na sinonimsku parafrazu kao osnovnom obliku parafraziranja. Ovdje je izvodimo statistički, usporedbom konteksta n-grama da bi pronašli riječi sličnog značenja. Također uspoređujemo različite mjere i parametre i njihov utjecaj na listu sinonima koja se generira. U poglavlju 4 predstavljene su metode zasnovane na pravilima. Za njihovu ekstrakciju potrebno je skupiti veći skup parafraza visoke preciznosti, kako bi se smanjio šum prisutan zbog rijetkosti podataka. Vezano uz taj problem usporedit će se neke od mjera sličnosti za prepoznavanje parafraza u usporedivom korpusu. Za primjer će biti predstavljene tri transformacije čija pravila pravila ćemo izlučiti iz dobivenog skupa. U poglavlju 5 predstavit će se problemi i ideje o evaluaciji ovakvog tipa sustava, te će se provesti evaluacija obje metode. Konačno, u poglavlju 6 će biti iznesen zaključak i predložen smjer daljnjeg rada.

2. Postojeća rješenja

Cjelokupan proces generiranja parafraza je složen zadatak, te se većina radova u ovom području bavi tek jednim dijelom tog procesa. Tako razlikujemo metode koje obavljaju **prepoznavanje**, (engl. *recognition*) **prikupljanje** (engl. *extraction*) te **generiranje** (engl. *generation*) parafraza. Sustav za prepoznavanje parafraza kao ulaz prima par jezičnih izraza za koji mora reći (s eventualnom procjenom vjerojatnosti) jesu li parafraze ili nisu. Pristupi rješavanja najčešće se zasnivaju na izgradnji klasifikatora metodama strojnog učenja, a razlikuju prema značajkama pomoću kojih se klasificira. Sustavi za prikupljanje parafraza primaju cijeli korpus, primjerice paralelan jednojezični korpus nastao od različitih prijevoda nekog romana. Izlaz takvog sustava su parovi, odnosno grupe parafraza rečenica ili izraza. Sustav za generiranje parafraza prima rečenicu, izraz ili uzorak, po mogućnosti u napomenuti da sustav ne mora spadati u isključivo jednu kategoriju ove podjele. Tako se sustav za generiranje može koristiti metodama prepoznavanja kako bi odabrao samo dobre generirane parafraze. Slično, sustav za generiranje može koristiti parove rečenica dobivene metodama prikupljanja kako bi izlučio pravila ili uzorke za oblikovanje novih parafraza.

2.1. Prepoznavanje

Zhang i Patrick (2005) rade na prepoznavanju parafraza na razini rečenice. Dane rečenice transformiraju se u kanonski oblik, s pretpostavkom da slični kanonski oblici dviju rečenica daju veću vjerojatnost da su rečenice parafraze. Ovdje je kanonizacija izvršena samo djelomično kako bi se pokazala valjanost pretpostavke: brojevni entiteti (datumi, postotci i ostale mjere količine) su zamjenjeni generičnim oznakama, pasivni oblici zamjenjeni su aktivnima te su svi glagoli namjere (npr. *planirati*, *namjeravati*) zamjenjeni futurom. Ovakva kanonizacija stoga zahtjeva mogućnost parsiranja i obrade sintaksnog stabla rečenice. Klasifikator se oblikuje nadziranim učenjem stabala odluke, koristeći sljedeće značajke iz kanoniziranih rečenica: duljina najdužeg zajedničkog podniza (slijednog i neslijednog), Levenshteineova udaljenost rečenica te modificirana n-gramska preciznost (ove mjere ćemo posebno dati detaljno u sljedećem poglavlju).

Qiu et al. (2006) također rade prepoznavanje na razini rečenice. Osvrće se na problem zanemarivih informacija u rečenici koje nisu bitne za parafraziranje, ali pogoršavaju rezultate mjere sličnosti. Stoga se fokusira na različitost takvih nesparenih dijelova. Prvo, rečenice dijeli na "informacijske grumene" (engl. *chunks*), pri čemu "grumen" predstavlja n-torka sačinjena od predikata i njegovih argumenata. Zatim se pohlepnim algoritmom uparuju n-torke najsličnije prema sintaksi, a ostatak nesparenih informacija klasificira se kao značajan ili ne. Ukoliko su nesparene n-torke zanemarive, rečenice se klasificiraju kao parafraze. Ovaj postupak zahtjeva sintaksni parser te označivač esmantičkih uloga (engl. *semantic role labeler*).

Connor i Roth (2007) usmjereni su na kontekstom uvjetovano parafraziranje. Dosadašnji radovi parafrazirali su rečenice i fraze neovisno od konteksta u kojem su se one nalazile. Ograničavaju se na prepoznavanje parafraza glagola unutar fraze. Kao kontekst uzimaju se subjekt i objekt vezani uz glagol te pokušava odrediti može li se u zadanom kontekstu glagolska fraza zamjeniti drugom. Pritom se koriste dva pravila za donošenje odluke: konteksto neovisno i ovisno. Kontekstno neovisan prag odluke koristi ocjenu preklapanja skupova konteksta zadanih dviju fraza. Kontekstno ovisan prag odluke koristi ocjenu preklapanja zadanih glagolskih fraza s drugim glagolskim frazama koje se često nalaze u zadanom kontekstu. Oba pravila za svaki kontekst čine jednostavan klasifikator primjenjiv na bilo koji par riječi. Jedini podesiv parametar klasifikatora je njegov prag. Svi jednostavni klasifikatori linearnom kombinacijum se kombiniraju u globalni. Lokalni klasifikatori uče se nenadzirano nad neoznačenim tekstom gdje prepoznaju koje se fraze pojavljuju u kojim kontekstima. Ipak, potreban je sintaksni parser za pronalaženje subjekta i objekta.

2.2. Prikupljanje

Metode prikupljanja u prvom se redu razlikuju prema odabiru mjere sličnosti koja se koristi te prema vrsti korpusa iz kojeg se parafraze prikupljaju. Korpus ne mora biti specifičan, ali mnogi radovi koriste paralelne korpuse (jednakih sadržaja prevedenog na različite jezike, primjerice različiti prijevodi istog romana) ili usporedive jednojezične korpuse (sličnog sadržaja, primjerice novinski korpusi različitih izvora). Ovakvi podatci nam daju dodatnu heuristiku u traženju parafraziranih rečenica.

Barzilay i Lee (2003) koriste korpuse dviju novinskih agencija. Rečenice se prvo grupiraju unutar svakog korpusa, sa svim brojevima, imenima i datumima zamjenjenim tokenima. Metrika koja se koristi je n-gramsko preklapanje za n-grame duljina 1 do 4. Iz grupiranih rečenica metodom višestrukog sravnjivanja rečenica (engl. *multiple sequence alignment*) stvaraju se uzorci u obliku rešetke (engl. *lattice*) koji na kompaktan način predstavljaju strukturnu sličnost n-grama iz rečenica. Varijabilni djelovi rešetaka zamjenjuju se mjestima za argumente (engl. *slot*). Konačno, parovi rešetka uparuju se među korpusima. Uspoređujući argumente rešetaka koji pripadaju člancima objavljenim na isti dan, određuje se koje rešetke

su međusobne parafraze. Nakon izgradnje skupa parova moguće je daljnje generiranje novih parafraza poravnavanjem dane rečenice s postojećim uzorkom i korištenjem njegovih parova.

Dolan et al. (2004) koriste korpus izgrađen od vijesti skupljanih iz tisuća novinskih izvora, grupirajući članke prema tekstu i datumu objave. Među unaprijed grupiranim člancima grupiraju se rečenice prema dva uvjeta. Jedan je Levenshteinova udaljenost manja od 12, čime se prikupljaju slične rečenice. Drugi je heurističko grupiranje prvih dviju rečenica svakog članka – pretpostavka je da će uvod u članak koji govori o istoj temi vjerojatno biti sličan. Na ovaj se način žele obuhvatiti parafraze koje su vrlo različite, ali su opet parafraze. Od ovih rečenica također se zahtjeva sličnost određenog broja riječi i duljina rečenica. Ipak, procjenjeno je da je 40% parova koje ovakva heuristika pronađe zapravo nepovezano.

Sekine (2005) koristi korpuse četiriju novinskih izvora za pronalaženje sinonima fraza. Ograničavaju se na fraze koje se pojavljuju između dva imenovana entiteta, npr. "X je kupio Y" i "X-ova kupovina Y-a". Ideja metode je da bez prethodnog označavanja i pružanja početnog uzorka prikupimo parafraze uz pomoć ključnih riječi i konteksta. Prvo se iz korpusa prikupe imenovani entiteti i sinktaksni grumeni (engl. *chunks*) koji su povezani s njima kao kontekst. Ujedno se za svaki entitet odredi kojoj od unaprijed zadanih kategorija pripada. Za svaku od riječi konteksta računa se mjera ITF-TF, te se ona s najvećom označava kao ključna riječ tog konteksta. Sve fraze s istim ključnim riječima grupiraju se zajedno. Na ovaj se način povećava količina informacija za pojedinu frazu koju želimo upariti. Konačno, grupe fraza s različitim ključnim riječima povezuju se prema parovima imenovanih entiteta koji su se nalazili uz njih.

Wubben et al. (2009) fokusiraju se na pronalaženje parafraziranih naslova novinskih članaka. Korpus se sastoji od četiri mjeseca sakupljanih vijesti na nizozemskom jeziku, grupiranih prema algoritmu k srednjih vrijednosti (engl. *k-means*). Unutar svake grupe naslovi se svrstavaju u podgrupe prema međusobnoj kosinusnoj sličnosti. Koriste se dvije granice – ukoliko je sličnost manja od donje, par se odbacuje, a ukoliko je veća od gornje granice, par se prihvaća. U preostalom slučaju uspoređuje se prvih 150 znakova članka i prema tome donosi sud o uparivanju naslova kao rečeničnih parafraza. U radu se razmatra i uvođenje tranzitivnosti uparivanja, ali je pokazano da takav postupak unosi velik broj pogrešnih parova u korpus.

Metoda koju predstavljaju Bhagat i Ravichandran (2008) ističe se po tome što ne zahtjeva alate za sintaksno parsiranje teksta, već se oslanja na isključivo statističku obradu. Koriste korpus 150 GB-a novinskih tekstova kako bi se nadoknadila rijetkost podataka. Pretprocesiranje uključuje jednostavno označavanje vrsta riječi (engl. *POS tagging*) i primjenu lokalno osjetljivog raspršenja (engl. *LSH - Local sensitivity hashing*) radi smanjenja složenosti izračuna. Ideja je slična prethodnima – fraze sličnog značenja prepoznaju se po kontekstu u kojem se nalaze. U ovom slučaju kontekst se definira kao riječi koje prethode i slijede frazu, s mjerom povezanosti s frazom izračunatom pomoću mjere uzajamne informacije (engl. *PMI – Point mutual information*). Za promatrane fraze uzimaju se n-grami duljine do pet riječi koji uključuju bar jedan glagol i imenicu. Za svaku frazu izgrađuje se vektor sa spomenutim PMI vrijednos-

tima svih riječi konteksta. Usto, svaka riječ ima oznaku je li se nalazila lijevo ili desno od fraze, kako bi se mogle naći parafraze invertiranog konteksta. Zbog velikog broja tekstova, vektori se grupiraju metodom LSH kojom je očuvana mjera kosinusne sličnost među njima.

2.3. Generiranje

Generiranje parafraza podrazumjeva stvaranje skupa semantički ekvivalentnih fraza ili rečenica na osnovu zadanog ulaza. Konvencionalne metode generiranja mogu se klasificirati na sljedeći način:

- **Metode zasnovane na pravilima** – u ovom slučaju, pravila mogu biti ručno napisana, što je skup i zahtjevan posao, ili automatski prikupljena. U oba slučaja, pokrivenost mogućih uzoraka nije velika, pogotovo kada su prikupljeni uzorci dugi i složeni. Ipak, ako je kao rezultat dovoljan specijalizirani podskup parafraza, ovakvim se metodama dobijaju najprecizniji rezultati.
- **Metode zasnovane na rječniku** – ovakve metode zasnivaju se na zamjeni riječi u danoj rečenici njihovim sinonimima. Najčešće se generiraju sve moguće varijante zamjena sinonima, a u drugom se koraku odabire optimalna zamjena za svaku od riječi s obzirom na kontekst u kojem se nalazi. Ovakve metode su jednostavne, ali omogućuju samo parafraziranje sinonimima.
- **Metode zasnovane na generiranju prirodnog jezika** – složene metode u kojima se dana rečenica prvo transformira u morfološku inačicu korištenjem sintaksne, semantičke i morfološke analize. Potom se ovako označena rečenica šalje u sustav za generiranje prirodnog jezika (engl. *natural language generation system*) koji pomoću nje generira nove rečenice. Ovakve metode simuliraju ljudski način parafraziranja, ali je njihova izvedba iznimno složena, kako zbog morfološke analize, tako zbog samog sustava generiranja prirodnog jezika.
- **Metode zasnovane na statističkom strojnom prevođenju** – metode izjednačavaju parafraziranje i prevođenje rečenice na isti jezik. Najčešće se koriste isti modeli kao i u metodama strojnog prevođenja, za čiju je izgradnju potreban velik paralelni korpus. Takvi korpusi trebaju biti istog sadržaja pisanog na različit način – primjerice različitih prijevoda istog romana, ili različitih izvještaja iste vijesti. Najveći problem ovog pristupa je upravo pronalaženje takvih podataka, pa se često koriste kombinacije različitih izvora.

U nastavku su pobliže opisane odabrane metode generiranja parafraza. Barzilay i Lee (2003) generiranje zasnivaju na pronalaženju uzoraka u rečenicama. Iz prikupljenih grupa rečenica uzorci se stvaraju metodom MSA, opisanom u prethodnom poglavlju. Za dobivenu rečenicu, određuje se kojoj grupi pripada usporedbom s izgrađenim rešetkama. Svim par-

alelnim rešetkama argumenti se zamjenjuju onima iz zadane rečenice, pri čemu generiranih parafraza dobijamo koliko je različitih puteva u rešetkama koje koristimo. Nedostatak ove metode je što se ograničava na razinu rečenice, pa će rečenice koje se parafraziraju morati biti vrlo slične rečenicama korpusa koje su korištene za uzorke. Ovime se uvelike smanjuje njena primjenjivost.

Quirk et al. (2004) također koriste korpus uparenih rečenica kako bi zaobišli problem nedostatka dovoljno velikog paralelnog jednojezičnog korpusa. Poravnavanjem fraza unutar rečenica dobija se tablica s parovima fraza, pri čemu se češće uparenim frazama pridružuje veća vjerojatnost. Prvo se izgrađuje rešetka koja predstavlja sve moguće parafraze koje se mogu dobiti zamjenom fraza s njihovim parom u tablici. Za određivanje parafraze koriste se vjerojatnosti iz tablice i jezičnog modela te se kao parafraza vraća izraz s najvećom ukupnom vjerojatnosti.

Zhao et al. (2009) pokazuju kako se kombiniranjem tablica fraza dobijenih iz različitih izvora mogu dobiti bolje generirane parafraze. Također predlažu ocjenjivanje kanidata za parafraze koristeći dodatan model ovisan o primjeni. Nazivaju ga model korištenja (engl. *usability model*); primjerice, pri sažimanju rečenica model korištenja preferira parafraze koje imaju manje riječi od izvorne. Svaki od modela ima težinu kojom se skalira – ove se težine podešavaju maksimizacijom točnih parafraza prema ljudskoj evaluaciji. Evaluacija je rađena na tri različite primjene: sažimanju, pojednostavljanju, te usporedbi sličnosti.

Drugi kriteriji prema kojima možemo kategorizirati radove su razina parafrazirane jedinke (riječ, fraza, rečenica, dulji tekst), način na koji se definira parafraziranje, i alati za obradu jezika koji se koriste u metodi (parseri, plitki parseri (engl. *chunkers*), alati za sravnjivanje (engl. *alligners*), i sl.)

3. Sinonimska parafraza

Svaki oblik parafraze najčešće sadrži sinonimsku supstituciju. Pod sinonimskom supstitucijom smatramo zamjenu bilo koje riječi ili fraze unutar danog teksta njenim semantičkim ekvivalentom. Tako je u slučajevima 3.1 i 3.2 riječ *zadnjoj* zamijenjena sinonimom *posljednjoj*. Zamjena može biti i na višoj razini od riječi, pa je tako u primjeru 3.3 riječ *poveo* zamijenjena frazom *prešao u vodstvo*. Prvotna fraza i njena supstitucija slažu se u gramatičkom vidu.

Jadran je ponovo poveo u zadnjoj četvrtini. (3.1)

Jadran je ponovo poveo u posljednjoj četvrtini. (3.2)

Jadran je ponovo prešao u vodstvo u posljednjoj četvrtini. (3.3)

No, sinonimija nije uvijek jednoznačna. Riječ često ima više značenja i njen sinonim ne mora se slagati sa svakim od njih. Glagol *napravila* u 3.4 se može zamijeniti glagolom *ispekla* u 3.5, ali samo u uskom kontekstu kulinarskih proizvoda koji se peku. *Ispekla* svakako ne može zamijeniti *napravila* u kontekstu rečenice 3.6. Stoga je, osim povezanosti samih riječi čiju relaciju sinonimije određujemo, potrebno obratiti pozornost i na povezanost konteksta u kojem se te riječi nalaze.

Beta je napravila kolače. (3.4)

Beta je ispekla kolače. (3.5)

Beta je napravila domaću zadaću. (3.6)

Osnovno je pitanje kako odrediti relaciju sinonimije. Ovom problemu, premda bliskom određivanju relacije parafraze, posebno se pristupalo u mnoštvu radova s velikim brojem različith pristupa. Metode prikupljanja sinonima mogu se konceptualno podijeliti u dvije skupine, prema izvoru iz kojeg se sinonimi prikupljaju. Metode koje prikupljaju sinonime iz **korpusa** zasnivaju se na ideji da se slične riječi nalaze u sličnim kontekstima, pa se problemu pristupa najčešće distribucijski. Metode koje prikapljaju sinonime iz **rječnika** koriste pretpostavku da su pojmovima srodne riječi nalaze u definiciji. Na taj se način izgrađuju grafovi riječi na kojima se pomoću raznih algoritama pronalaze semantički bliski vrhovi. Nedostatak ovih metoda je što pronalaze sinonime samo na razini riječi. Također, potrebni podatci za obradu (elektronički rječnici) na hrvatskom jeziku ograničeni su količinom.

3.1. Prikupljanje sinonima

U ovom radu za prikupljanje sinonima odabrana je metoda koja koristi korpus. Metoda je slična onoj opisanoj u Bhagat i Ravichandran (2008), i oslanja se na Harrisovu distribucijsku pretpostavku (Harris, 1954) – ideja je da riječi koje se pojavljuju u sličnom kontekstu imaju slično značenje. Metoda primjenjuje ovu pretpostavku na fraze, odnosno n-grame riječi.

Na primjer, fraza *kupio* u izrazu "X je kupio Y" može se naći u kontekstu {*Google*, *Microsoft*, *Milan*, ...} na mjestu X, i {*YouTube*, *Skype*, *Mexesa*, ...} na mjestu Y. Druga fraza, "dogovorio kupnju", opet u izrazu "X je dogovorio kupnju Y", može se naći u kontekstu {*Google*, *Microsoft*, *Kerum*, ...} na mjestu X i {*YouTubea*, *Skypea*, *svega*, ...} na mjestu Y. Iz toga možemo zaključiti da su konteksti ovih fraza slični, pa prema distribucijskoj pretpostavci fraze *kupio* i *dogovorio kupnju* imaju slična značenja. U nastavku rada ćemo ih zbog jednostavnosti nazivati sinonima, iako će se većina složiti da se radi o bliskoznačnicama (*dogovorio kupnju* ostavlja mogućnost da se kupnja izjalovi, dok *kupio* označava obavljenu radnju).

Formalno, neka je *p* fraza (n-gram) u izrazu *X p Y*, gdje X i Y prestavljaju riječi koje se nalaze neposredno lijevo i desno od n-grama *p*. Naš cilj je pronaći skup fraza značenja sličnog frazi *p*. Neka je $P = \{p_1, p_2, p_3, \ldots, p_l\}$ skup svih fraza oblika Xp_iY, gdje je $p_i \in P$. Neka je $S_{i,X}$ skup svih riječi koje se pojavljuju na poziciji X uz p_i, a $S_{i,X}$ skup svih riječi koje se pojavljuju na poziciji Y. Neka je V_i vektor koji predstavlja p_i tako da je $V_i = S_{i,X} \cup S_{i,Y}$. Svaka riječ $f \in V_i$ ima pridruženu vrijednost koja mjeri jačinu povezanosti riječi f i fraze p_i. Ova vrijednost se može izračunati na različite načine, a ovdje ćemo predstaviti samo neke od njih.

Najjednostavniji način označavanja je pridruživanje vrijednosti 1 značajkama koje su prisutne u kontekstu riječi i vrijednosti 0 značajkama koje nisu. Drugi način je pridruživanje frekvencije odnosno broja pojavljivanja svake značajke. Za riječ p i značajku f vrijednost se definira kao:

$$weight_{freq}(p, f) = count(p, f)$$

Korištenjem vjerojatnosti da se značajka pojavi u kontekstu riječi:

$$weight_{prob}(p, f) = P(f|p)$$

Ovakav pristup daje veće vrijednosti riječima koje se općenito češće pojavljuju u tekstu pa se često pruža dojam kako su takve riječi značajnije, što ne mora biti točno. U nastavku su prikazane težinske funkcije koje otklanjaju taj problem.

Mjera uzajamne informacije *PMI* je mjera koja uspoređuje vjerojatnost da se riječ i značajka nađu u istom konteksu s vjerojatnostima da se u tekstu nađu samostalno. Na ovaj način se smanjuje vrijednost značajki koje imaju veliku vjerojatnost samostalnog pojavljivanja.

$$weight_{PMI}(p, f) = \log_2 \frac{P(p, f)}{P(p)P(f)}$$

T-test vrijednost se računa na sličan način, razlikom izmjerene i očekivane srednje vrijednosti, te normaliziranjem varijancom.

$$weight_{t-test}(p, f) = \frac{P(p, f) - P(p)P(f)}{\sqrt{P(w)P(f)}}$$

Dice je još jedna varijanta ovog pristupa. Prema (Curran i Moens, 2002), dodavanje faktora $\log_2(count(w, f) + 1)$ na bilo koju od ovih funkcija može povećati utjecaj značajki veće frekvencije, što u nekim slučajevima dovodi do boljih rezultata.

$$weight_{dice}(p, f) = \frac{2P(p, f)}{P(p) + P(f)}$$

Nakon što se izračunaju vektori za svaku frazu $p_i \in P$, parafraze za svaki p_i tražimo preko njegovih najbližih susjeda. Za usporedbu vektora najčešće kosinusnu mjeru sličnosti. Za fraze $p_i \in P$ i $p_j \in P$ s pripadajućim vektorima V_i i V_j kosinusna sličnost se definira kao:

$$sim_{cosine}(p_i, p_j) = \frac{V_i \cdot V_j}{|V_i| * |V_j|}$$

Svaka riječ u V_i ima i dodatnu zastavicu koja označava nalazi li se riječ u skupu $S_{i,X}$ ili $S_{i,Y}$. Tako za svaku frazu p_i u izrazu Xp_iY imamo odgovarajuću frazu $-p_i$ s pripadajućim izrazom Yp_iX. Ovo je bitno za određenu vrstu parafraziranja, kao što je primjer u sljedećim rečenicama:

Microsoft je kupio Skype. (3.7)

Skype je kupljen od strane Microsofta. (3.8)

Izrazi *je kupio* i *je kupljen od strane* nisu sinonimi, ali posjeduju isto značenje obrnutog konteksta. Na ovaj se način svaka fraza može promatrati na dva načina, dajući nam dvostruko više međusobnih razlikovanja. U hrvatskom jeziku ovakva pojava ipak nije dovoljno česta da bismo je uspjeli bolje analizirati na manjem korpusu poput ovog s kojim radimo. Zbog malog korpusa se također pri selekciji n-grama ne ograničavamo na one koji sadrže određenu vrstu riječi, dok se u u srodnim radovima često specijalizira samo na imenovane entitete ili n-grame koji sadrže glagoli i imenicu.

Problem evaluacije metoda za pronalaženje parafraza, pa tako i sinonima, je nepostojanje generalnog konsenzusa što se prihvaća kao sinonim, a što ne. Automatska evaluacija najčešće se vrši usporedbom s već postojećim tezaurusom (rječnikom sinonima). Ovo je metoda upitne pouzdanosti, pogotovo ako se uzme u obzir da se sinonimi automatski generiraju upravo da bi se tezaurusi nadopunili. Isto vrijedi i za generiranje domenski specifičnih tezaurusa – stvaramo ih upravo radi nedostatka ručno napisanih, pa ih nema smisla uspoređivati s tezaurusima općih ili drugih domena. K tome, tezaurus za hrvatski jezik u digitalnom obliku ne postoji, pa se od ove metode evaluacije odustalo.

Druga, i ovdje korištena metoda za evaluaciju sinonima je direktnim uvidom u sinonime odabranog manjeg uzorka riječi. Na ovaj način može se direktnije pristupiti rezultatima i uočiti sitne promjene koje nastaju promjenom određenih parametara. U našem slučaju to su sljedeći parametri: funkcija značajke vektora, zanemarivanje zaustavnih riječi, odbacivanje rijetkih n-grama, razlikovanje lijevog i desnog konteksta, lematiziranje konteksta te veličina konteksta.

U nastavku će biti prikazani usporedni rezultati deset najbolje rangiranih sinonima za odabrane riječi. Korpus na kojem je izvršena evaluacija sastoji se od 56480 članaka prikupljanih u vremenskom periodu od tri mjeseca sa sljedećih novinskih portala: Nacional[1], Večernji list[2], Index[3], Slobodna Dalmacija[4], Business[5], Jutarnji list[6], Tportal[7], Monitor[8], Net[9], Dalje[10], Dnevnik[11], 24sata[12]. U procesu obrade svi brojevi i imenovani entiteti zamijenjeni su posebnom oznakom. Problem prepoznavanja imenovanih entiteta je složen, pa smo se u ovom radu zadovoljili jednostavnim pristupom: imenovanim entitetima označene su sve riječi koje se u korpusu pojavljuju isključivo s velikim početnim slovom.

3.1.1. Utjecaj funkcije za izračunavanje značajki vektora

U tablicama je prikazano deset najbolje rangiranih sinonima devet odabranih riječi za različite načine izračunavanja vrijednosti u vektorima n-grama. Širina konteksta koji se uzima u obzir je po jedna riječ lijevo i desno. Za ovu evaluaciju birane su riječi koje se mogu pronaći u novinskim člancima, različitih vrsta riječi: tri imenice, tri pridjeva i tri glagola. Osnovni zahtjev prema kojem ćemo ocjenjivati rezultate je da sinonimi budu što sličnijeg značenja izvornoj riječi. Dodatni kriterij je slaganje sinonima po vrsti riječi, rodu, broju, padežu ili vidu.

- *Pobjednik* ima nekoliko bliskoznačnica. Željeni rezultat je izbjeći ostale sportske pojmove vezane uz natjecanja.
- *Automobil* je riječ koja nema mnogo sinonima. Željeni rezultat je što bolje rangirati sinonim *auto* i hipernim *vozilo*.
- *Vođa* ima mnogo sinonima. Željeni rezultat je što bolje rangirati prave sinonime, a što lošije hiponime poput *predsjednik* ili *potpredsjednik*.

[1] www.nacional.hr
[2] www.vecernji.hr
[3] www.index.hr
[4] www.slobodnadalmacija.hr
[5] www.business.hr
[6] www.jutarnji.hr
[7] www.tportal.hr
[8] www.monitor.hr
[9] www.net.hr
[10] www.dalje.hr
[11] www.dnevnik.hr
[12] www.24sata.hr

- *Uspješan* nema pravog sinonima. Željeni rezultat je što bolje rangirati bliske pozitivne atribute, a izbjeći suprotnosti.
- *Ozlijeđeno* je riječ s nekoliko sinonima. Željeni rezultat je izbjeći druge glagole često spominjane u crnoj kronici (*uhićeno*, *ubijeno* i sl.)
- *Velik* ima mnogo sinonima. Željeni rezultat je što bolje rangirati prave sinonime, a izbjeći antonime.
- *Kupio* nema pravog sinonima. Željeni rezultat je što bolje rangiranje bliskoznačnice *platio*, a što lošije antonima poput *prodao*.
- *Porastao* ima nekoliko sinonima. Željeni rezultat je što bolje rangirati bliska značenja, a izbjeći suprotnosti.
- *Smatra* ima mnogo sinonima. Željeni rezultat je što bolje ranigirati prave sinonime koji se slažu u rodu i broju.

U tablici 3.1 prikazani su rezultati metode koja za označavanje koristi funkciju BOOL – samo označavanje je li riječ prisutna u kontekstu ili nije. Rezultati za riječi *vođa*, *velik*, *kazao* i *porastao* su zadovoljavajući – sinonimi se većinom slažu, kako s gramatičke tako i sa semantičke strane. Prisutno je tek nekoliko antonima (npr. *oslabio*, *potonuo* i *pao*), što je shvatljivo jer se riječi poput *porastao* i *pao* često nalaze u sličnom kontekstu, pogotovo u domeni burzovnih i poslovnih izvješća.

Sinonimi za *smatra* su također dobri, no sadrže i određen broj glagolskih oblika u prošlom vremenu za koje se čini da su djelovi bigramskih sinonima (npr. *izjavio* u *izjavio je*). *Ozlijeđeno* sadrži mnogo vezanih riječi koje nisu sinonimi (*poginulo*, *umrlo*, *privedeno*), ali dohvaća i pravi sinonim *ranjeno*. *Pobjednik* je jedina riječ kojoj nije pronađen niti jedan pravi sinonim, već samo semantički bliske riječi vezane za natjecanje.

U tablici 3.2 prikazani su rezultati dobijeni korištenjem funkcije COUNT koja u vektor zapisuje broj pojavljivanja riječi u kontekstu. Odmah je vidljivo da su rezultati znatno lošiji, gdje su najbolje rangirani sinonimi pronađeni jedino za *smatra* i *ozlijeđeno*. Prisutan je i određen broj vlastitih imena koji nisu prepoznati kao imenovani entiteti, pa su zbog malog broja pojavljivanja dosegli umjetnu visoku sličnost s izvornom riječi. Vrijednosti koje mjera sličnosti poprima ukazuju na to da riječi koje su općenito česte u tekstu previše utječu na ukupan rezultat. Zbog toga se neke naizgled nepovezane riječi poput *velik* i *nedefiniran* ocjenjuju sličnima – u ovom slučaju riječ *broj* koja se uz *velik* pojavljuje 410 puta. Potrebna je mjera koja će penalizirati riječi koje se općenito često pojavljuju.

Tablica 3.3 sadrži rezultate dobijene korištenjem funkcije PMI. Ovi rezultati su vrlo slični onima dobivenim funkcijom BOOL, uz male promjene u rangovima i ponekim novim sinonimima. Ukupno nešto bolji sinonimi pronađeni su za riječi *pobjednik* i *uspješan*, ali i dalje ostaje problem potpuno nepovezanih riječi kao *stan* za *automobil* i antonima poput *uspješan* i *loš*.

POBJEDNIK	AUTOMOBIL	VOĐA	USPJEŠAN
finalu 0, 274	auto 0, 267	lider 0, 236	uspješni 0, 284
pobjednici 0, 261	stan 0, 237	čelnik 0, 217	sretan 0, 275
polufinale 0, 246	kamion 0, 231	vođe 0, 216	uvjerljiv 0, 267
pobjednika 0, 245	vozilo 0, 220	šef 0, 186	loš 0, 265
polufinalu 0, 240	brod 0, 217	član 0, 183	dobar 0, 264
osvajač 0, 238	vozač 0, 215	premijer 0, 182	osuđivan 0, 258
finale 0, 235	igrač 0, 215	potpredsjednik 0, 181	ispitan 0, 258
četvrtfinalu 0, 233	klub 0, 210	predsjednik 0, 181	atraktivan 0, 254
nositelj 0, 220	godišnjak 0, 208	dužnosnik 0, 180	težak 0, 253
kolu 0, 218	otkaz 0, 208	lideri 0, 179	egalu 0, 252

OZLIJEĐENO	VELIK	KUPIO	PORASTAO	SMATRA
ranjeno 0, 417	veliki 0, 303	platio 0, 257	ojačao 0, 344	tvrdi 0, 254
uhićeno 0, 359	veći 0, 296	prodao 0, 247	porasla 0, 312	kaže 0, 249
ubijeno 0, 355	najveći 0, 248	kupila 0, 238	oslabio 0, 290	ističe 0, 248
poginulo 0, 352	ogroman 0, 246	kupi 0, 237	iznosio 0, 279	kazao 0, 239
umrlo 0, 347	značajan 0, 246	kupiti 0, 235	potonuo 0, 275	rekao 0, 239
privedeno 0, 321	dobar 0, 238	kupili 0, 229	skočio 0, 272	čini 0, 222
privedenih 0, 305	snažan 0, 237	živio 0, 221	porasti 0, 265	navodi 0, 220
stradalo 0, 302	težak 0, 235	vratio 0, 221	porasle 0, 261	izjavio 0, 217
ozlijeđenih 0, 294	važan 0, 232	napravio 0, 220	pao 0, 256	rekla 0, 214
otkazano 0, 293	ozbiljan 0, 229	učinio 0, 220	dosegnuo 0, 252	piše 0, 213

Tablica 3.1: Sinonimi prikupljeni izgradnjom vektora funkcijom BOOL

POBJEDNIK	AUTOMOBIL	VOĐA	USPJEŠAN
jobs 0, 840	avion 0, 848	diktator 0, 718	dostupan 0, 777
pobijedio 0, 834	klub 0, 831	redatelj 0, 714	pozvan 0, 761
pobijedila 0, 831	kina 0, 830	osnivač 0, 712	neugodan 0, 758
gadafi 0, 829	poledica 0, 828	vlasnik 0, 712	koristan 0, 758
nagrađen 0, 828	auto 0, 824	promrzao 0, 712	oženjen 0, 744
platini 0, 825	izvoz 0, 815	naslijedio 0, 710	aktivan 0, 744
glumac 0, 824	život 0, 812	glumac 0, 708	prisutan 0, 743
bend 0, 821	hajduk 0, 812	balić 0, 708	ponuđen 0, 740
utjelovio 0, 819	otišao 0, 811	marić 0, 708	precizan 0, 720
ukraden 0, 819	odiozno 0, 809	utemeljitelj 0, 707	najbrži 0, 716

OZLIJEĐENO	VELIK	KUPIO	PORASTAO	SMATRA
ranjeno 0, 976	narudžaba 0, 858	pobijedila 0, 959	oslabila 0, 929	tvrdi 0, 989
poginulo 0, 956	uvećavao 0, 858	odigrao 0, 949	glasalo 0, 926	priznaje 0, 950
sudjelovalo 0, 941	nemali 0, 856	postavio 0, 949	ojačala 0, 916	priznavši 0, 950
uhićeno 0, 940	premašivao 0, 818	napustio 0, 947	kliznuo 0, 912	ustvrdivši 0, 946
završilo 0, 937	rekordan 0, 816	ponudio 0, 946	porasla 0, 911	tvrdeći 0, 945
ubijeno 0, 936	popriličan 0, 814	pobijedio 0, 946	protrgovano 0, 910	poručivši 0, 943
poginuo 0, 934	najveći 0, 801	prodao 0, 944	ojačao 0, 906	naglašava 0, 941
umro 0, 930	peteroznamenkasti 0, 792	zabio 0, 943	prodana 0, 900	napominje 0, 938
pogubljen 0, 930	nedefiniran 0, 791	snimila 0, 943	doniran 0, 896	sugerirajući 0, 936
preminulo 0, 929	malen 0, 785	osnovao 0, 941	prikupljeno 0, 895	izjavivši 0, 935

Tablica 3.2: Sinonimi prikupljeni izgradnjom vektora funkcijom COUNT

POBJEDNIK	AUTOMOBIL	VOĐA	USPJEŠAN
pobjednici 0, 218	auto 0, 211	lider 0, 186	uspješni 0, 219
finalu 0, 218	kamion 0, 190	čelnik 0, 176	sretan 0, 216
osvajač 0, 216	stan 0, 165	vođe 0, 173	uvjerljiv 0, 215
finalist 0, 204	vozilo 0, 162	lideri 0, 138	dobar 0, 215
pobjednika 0, 198	automobilom 0, 159	šef 0, 136	loš 0, 200
četvrtfinalu 0, 196	autobus 0, 157	predsjednik 0, 136	nervozan 0, 197
polufinalu 0, 187	kombi 0, 155	član 0, 135	osuđivan 0, 194
polufinale 0, 186	vozač 0, 149	potpredsjednik 0, 132	slavan 0, 192
finale 0, 178	brod 0, 149	dužnosnik 0, 129	ljubazan 0, 191
finala 0, 171	vozio 0, 148	premijer 0, 125	oprezan 0, 191

OZLIJEĐENO	VELIK	KUPIO	PORASTAO	SMATRA
ranjeno 0, 369	veliki 0, 261	prodao 0, 201	ojačao 0, 305	tvrdi 0, 187
uhićeno 0, 322	veći 0, 249	platio 0, 198	oslabio 0, 263	ističe 0, 182
ubijeno 0, 313	ogroman 0, 219	kupi 0, 190	porasla 0, 261	kaže 0, 180
poginulo 0, 312	značajan 0, 213	kupila 0, 188	skočio 0, 238	rekao 0, 165
umrlo 0, 285	najveći 0, 204	kupiti 0, 180	potonuo 0, 236	kazao 0, 164
privedeno 0, 272	snažan 0, 195	kupili 0, 177	iznosio 0, 233	navodi 0, 155
stradalo 0, 266	ozbiljan 0, 189	živio 0, 155	porasti 0, 222	čini 0, 152
ozlijeđenih 0, 259	važan 0, 188	prodati 0, 155	porasle 0, 217	izjavio 0, 150
zaposleno 0, 244	dobar 0, 187	vratio 0, 155	pao 0, 213	upozorava 0, 148
otkazano 0, 243	težak 0, 184	vozio 0, 154	dosegnuo 0, 211	piše 0, 145

Tablica 3.3: Sinonimi prikupljeni izgradnjom vektora funkcijom PMI

POBJEDNIK	AUTOMOBIL	VOĐA	USPJEŠAN
pobjednika 0, 198	auto 0, 207	čelnik 0, 166	dobar 0, 174
finalist 0, 182	automobila 0, 178	lider 0, 166	odličan 0, 156
osvajač 0, 172	automobilom 0, 171	predsjednik 0, 163	najbolji 0, 151
osvojio 0, 166	kamion 0, 168	član 0, 155	kvalitetan 0, 143
pobjednica 0, 162	vozilo 0, 163	potpredsjednik 0, 147	sjajan 0, 142
finale 0, 157	vozač 0, 163	premijer 0, 146	velik 0, 142
prvak 0, 157	vozila 0, 156	šef 0, 146	fantastičan 0, 136
polufinale 0, 155	automobilu 0, 155	vođe 0, 138	veliki 0, 133
prvaka 0, 152	stan 0, 151	stranke 0, 137	loš 0, 132
pobijedio 0, 150	godine 0, 150	trener 0, 135	značajan 0, 132

OZLIJEĐENO	VELIK	KUPIO	PORASTAO	SMATRA
poginulo 0, 291	veliki 0, 251	kupili 0, 174	iznosio 0, 240	kazao 0, 293
ranjeno 0, 280	veći 0, 226	prodao 0, 172	ojačao 0, 236	rekao 0, 287
poginula 0, 242	najveći 0, 214	platio 0, 167	oslabio 0, 214	kaže 0, 287
ubijeno 0, 232	značajan 0, 198	kupiti 0, 165	porasla 0, 214	tvrdi 0, 283
ozlijeđena 0, 229	dobar 0, 193	kupila 0, 157	narastao 0, 211	ističe 0, 270
ozlijeđenih 0, 215	ogroman 0, 190	dobio 0, 154	porasle 0, 203	izjavio 0, 269
stradalo 0, 191	hrvatskoj 0, 172	kupljen 0, 153	dosegnuo 0, 197	dodao 0, 261
privedeno 0, 190	puno 0, 172	napravio 0, 142	dosegnula 0, 183	istaknuo 0, 258
ozlijeđene 0, 180	sigurno 0, 170	prodati 0, 141	skočio 0, 182	može 0, 254
poginule 0, 179	pravi 0, 168	imao 0, 140	porasti 0, 181	treba 0, 247

Tablica 3.4: Sinonimi prikupljeni izgradnjom vektora funkcijom BOOL bez zaustavnih riječi

POBJEDNIK	AUTOMOBIL	VOĐA	USPJEŠAN
finalist 0, 173	auto 0, 182	čelnik 0, 144	dobar 0, 145
pobjednika 0, 172	kamion 0, 156	lider 0, 140	odličan 0, 126
osvajač 0, 160	automobila 0, 153	predsjednik 0, 135	najbolji 0, 122
pobjednica 0, 146	automobilom 0, 152	vođe 0, 130	fantastičan 0, 116
osvojio 0, 145	vozilo 0, 142	član 0, 126	sjajan 0, 115
polufinalist 0, 137	vozač 0, 138	šef 0, 118	kvalitetan 0, 115
prvak 0, 136	automobilu 0, 135	potpredsjednik 0, 115	velik 0, 113
finale 0, 132	mercedes 0, 131	premijer 0, 115	loš 0, 108
superkombinacije 0, 131	vozila 0, 124	čelnika 0, 109	siguran 0, 106
pobjednici 0, 127	oznaka 0, 120	stranke 0, 105	veliki 0, 104

OZLIJEĐENO	VELIK	KUPIO	PORASTAO	SMATRA
poginulo 0, 277	veliki 0, 221	prodao 0, 144	ojačao 0, 228	kazao 0, 225
ranjeno 0, 273	veći 0, 192	kupiti 0, 142	iznosio 0, 218	rekao 0, 220
ozlijeđena 0, 213	najveći 0, 181	kupili 0, 141	oslabio 0, 205	kaže 0, 220
poginula 0, 213	značajan 0, 174	platio 0, 139	porasla 0, 193	tvrdi 0, 219
ozlijeđenih 0, 203	ogroman 0, 174	kupila 0, 137	porasle 0, 183	ističe 0, 209
ubijeno 0, 200	dobar 0, 156	kupljen 0, 124	narastao 0, 179	izjavio 0, 204
ozlijeđene 0, 168	snažan 0, 141	dobio 0, 121	dosegnuo 0, 177	dodao 0, 200
stradalo 0, 166	ozbiljan 0, 140	prodati 0, 113	porasti 0, 163	istaknuo 0, 199
privedeno 0, 159	golemi 0, 139	posjeduje 0, 110	skočio 0, 160	smatraju 0, 186
poginule 0, 158	pravi 0, 134	ima 0, 108	porasli 0, 153	navodi 0, 186

Tablica 3.5: Sinonimi prikupljeni izgradnjom vektora funkcijom PMI bez zaustavnih riječi

POBJEDNIK	AUTOMOBIL	VOĐA	USPJEŠAN
pobjednika 0, 376	automobila 0, 361	vođe 0, 338	uspješni 0, 336
finalu 0, 336	auto 0, 323	čelnik 0, 275	sretan 0, 296
polufinale 0, 334	vozila 0, 316	član 0, 269	dobar 0, 293
pobjednici 0, 313	stan 0, 311	lider 0, 266	atraktivan 0, 270
polufinalu 0, 304	automobilu 0, 306	organizacija 0, 252	loš 0, 269
finala 0, 300	automobilom 0, 282	premijer 0, 252	ispitan 0, 269
finale 0, 296	vozač 0, 264	predsjednik 0, 250	zabrinut 0, 268
turnir 0, 296	sati 0, 263	predstavnici 0, 249	oprezan 0, 267
četvrtfinalu 0, 293	kuće 0, 262	dužnosnika 0, 247	osuđivan 0, 264
četvrtfinale 0, 292	ulici 0, 258	čelnici 0, 247	teškoća 0, 263

OZLIJEĐENO	VELIK	KUPIO	PORASTAO	SMATRA
ranjeno 0, 462	veći 0, 351	prodao 0, 302	ojačao 0, 388	kaže 0, 382
poginulo 0, 391	veliki 0, 347	platio 0, 299	porasla 0, 377	rekao 0, 363
uhićeno 0, 386	najveći 0, 296	kupi 0, 294	iznosio 0, 356	kazao 0, 357
ubijeno 0, 381	značajan 0, 283	kupiti 0, 292	porasli 0, 340	tvrdi 0, 353
umrlo 0, 367	dobar 0, 274	vratio 0, 278	porasle 0, 334	nema 0, 353
stradalo 0, 324	težak 0, 274	prodati 0, 270	oslabio 0, 319	ističe 0, 352
privedeno 0, 323	ogroman 0, 271	napravio 0, 266	dosegnuo 0, 309	ima 0, 347
ozlijeđenih 0, 318	važan 0, 264	živio 0, 264	porasti 0, 296	izjavio 0, 340
privedenih 0, 315	snažan 0, 258	vratiti 0, 264	skočio 0, 295	piše 0, 334
izašlo 0, 308	sigurno 0, 257	učinio 0, 263	ostvareno 0, 295	imaju 0, 333

Tablica 3.6: Sinonimi prikupljeni izgradnjom vektora funkcijom BOOL s lematiziranim kontekstom

3.1.2. Utjecaj izostavljanja zaustavnih riječi

Jedan mogući problem u prepoznavanju sinonima je pojava riječi u kontekstu koje ne doprinose informaciji o značenju riječi. Neke od tih riječi su veznici, pomoćni glagoli i priložne oznake. Ipak, takve riječi mogu dati informaciju o gramatičkom vidu. U nastavku su dani rezultati dobijeni računanjem konteksnih značajki iz korpusa iz kojeg su izbačene zaustavne riječi.

U tablici 3.4 prikazani su rezultati za funkciju BOOL, te u tablici 3.5 za funkciju PMI. Rezultati su precizniji nego u slučaju korištenja zaustavnih riječi, pogotovo u slučaju riječi *uspješan* i *pobjednik*. Antonima je manje, a nema ni nepovezanih riječi poput misterioznog *stana* za *automobil*. Ipak neki su rezultati pogoršani, pa je prvi sinonim riječi *ozlijeđeno* u oba slučaja *poginulo*.

3.1.3. Utjecaj lematizacije

Lematizacija (engl. *lemmatization*) je svođenje pojavnica iz korpusa na njihove natukničke oblike, tj. svođenje različitih pojavnica na zajedničku lemu. Na primjer, pojavnice mačka, mačkama ili mačak bile bi svedene na lemu mačka. Lema je onaj oblik pod kojim bismo tražili neku riječ u rječniku. Pretpostavka je da ćemo, zbog ograničene veličine korpusa, lematizacijom spojiti različite pojavnice u istu lemu i time dati veći značaj riječima iz konteksta koje se češće ponavljaju. U tablici 3.6 prikazani su rezultati prikupljanja sinonima uz funkciju PMI.

Rezultati su vidljivo lošiji od onih dobijenih izostavljanjem zaustavnih riječi. Najveći gubitak dogodio se u gramatičkoj jednoznačnosti. Među liste sinonima dodani su mnogi oblici izvorne riječi koji ne samo da nisu sinonimi, nego se ni ne slažu u gramatičkom vidu (npr. za *automobil* odabrani su *automobila*, *automobilom* i *automobilu*). Možemo zaključiti da lematizacija nije pogodna za ovaj oblik prikupljanja sinonima iz razloga što je velik dio gramatičke informacije sadržan u nelematiziranoj pojavnici.

3.1.4. Utjecaj ostalih parametara

Moguć način za dobivanje veće preciznosti je razlikovanje na kojoj se strani riječi kontekst nalazio. To znači da će se riječi *mora* u "mora preplivati" i "preplivati mora" bilježiti kao dvije različite riječi. Prilikom provjeravanja uspoređuje se vektor, te "invertirani vektor" kojem su strane konteksta zamijenjene. Na ovaj način pronalazimo invertirane sinonime koji su ranije objašnjeni (v. odjeljak 3.1). U tablicama su invertirani sinonimi označeni prefiksom "-".

Rezultati prikazani u tablici 3.7 bliski su onima u tablici 3.5. Invertirani sinonimi se ovdje pojavljuju samo u slučaju kada je sinonim i s jedne i s druge strane dovoljno sličan izvornoj riječi (npr. *ranjeno* i *-ranjeno* odabrani su kao različiti sinonimi).

N-grami u svom neposrednom kontekstu znaju imati riječi koje ne daju mnogo informacije o značenju – pomoćne glagole, prijedloge ili veznike. Stoga je opravdana pretpostavka da će

POBJEDNIK	AUTOMOBIL	VOĐA	USPJEŠAN
osvajač 0, 164	auto 0, 176	čelnik 0, 154	dobar 0, 151
finalist 0, 164	kamion 0, 147	lider 0, 136	odličan 0, 132
pobjednika 0, 163	automobilom 0, 140	predsjednik 0, 134	sjajan 0, 124
pobjednica 0, 154	vozač 0, 139	član 0, 124	fantastičan 0, 122
polufinalist 0, 146	automobilu 0, 133	šef 0, 124	najbolji 0, 121
pobjednici 0, 143	vozilo 0, 129	premijer 0, 117	kvalitetan 0, 117
prvak 0, 135	automobila 0, 127	potpredsjednik 0, 116	zeza 0, 115
finale 0, 131	stan 0, 126	trener 0, 113	zanimljiv 0, 114
kolu 0, 130	autobus 0, 121	dužnosnik 0, 111	bolji 0, 113
polufinale 0, 126	mercedes 0, 118	ministar 0, 107	-pretil 0, 113

OZLIJEĐENO	VELIK	KUPIO	PORASTAO	SMATRA
ranjeno 0, 276	veliki 0, 236	kupila 0, 160	iznosio 0, 231	kaže 0, 242
poginulo 0, 258	veći 0, 202	kupili 0, 150	ojačao 0, 229	rekao 0, 240
-ranjeno 0, 219	najveći 0, 200	kupljen 0, 137	oslabio 0, 206	kazao 0, 239
ozlijeđena 0, 208	ogroman 0, 179	prodao 0, 136	porasla 0, 203	tvrdi 0, 238
stradalo 0, 203	značajan 0, 177	platio 0, 129	porasle 0, 190	ističe 0, 223
ozlijeđene 0, 201	dobar 0, 174	kupiti 0, 129	dosegnuo 0, 185	dodao 0, 219
ozlijeđenih 0, 200	snažan 0, 151	dobio 0, 127	narastao 0, 183	izjavio 0, 218
-ubijeno 0, 198	pravi 0, 151	imao 0, 114	porasti 0, 181	čini 0, 205
poginula 0, 198	novi 0, 149	ima 0, 114	skočio 0, 171	istaknuo 0, 202
-poginulo 0, 194	ozbiljan 0, 146	-prodao 0, 112	rasle 0, 164	navodi 0, 201

Tablica 3.7: Sinonimi prikupljeni izgradnjom vektora funkcijom BOOL s razlikovanjem strana konteksta

POBJEDNIK	AUTOMOBIL	VOĐA	USPJEŠAN
pobjednika 0, 265	auto 0, 290	član 0, 242	sretan 0, 267
prvak 0, 252	stan 0, 244	lider 0, 234	loš 0, 261
polufinale 0, 247	vozilo 0, 242	potpredsjednik 0, 229	odličan 0, 257
favorit 0, 245	vozač 0, 240	šef 0, 225	zanimljiv 0, 253
finalu 0, 244	automobilom 0, 232	čelnik 0, 225	lagan 0, 252
pobjednici 0, 243	automobila 0, 227	premijer 0, 220	dobar 0, 250
finale 0, 241	kamion 0, 224	predsjednik 0, 220	ponosan 0, 248
finala 0, 236	klub 0, 223	bivši 0, 211	aktivan 0, 244
pobjednica 0, 236	osobni 0, 221	vođe 0, 211	gotov 0, 238
osvojio 0, 236	muškarac 0, 219	trener 0, 210	korektan 0, 238

OZLIJEĐENO	VELIK	KUPIO	PORASTAO	SMATRA
ranjeno 0, 427	veliki 0, 310	kupila 0, 275	ojačao 0, 317	tvrdi 0, 307
poginulo 0, 340	veći 0, 304	prodao 0, 270	iznosio 0, 289	kazao 0, 305
ozlijeđenih 0, 331	broj 0, 292	platio 0, 267	oslabio 0, 287	kaže 0, 302
stradalo 0, 331	najveći 0, 273	kupiti 0, 266	porasle 0, 286	ističe 0, 300
privedeno 0, 326	dobar 0, 263	kupili 0, 247	porasla 0, 286	rekao 0, 299
ubijeno 0, 323	značajan 0, 244	kupovao 0, 240	narastao 0, 269	izjavio 0, 285
uhićeno 0, 319	interes 0, 236	uzeo 0, 240	dosegnuo 0, 260	dodao 0, 278
ozlijeđena 0, 294	puno 0, 235	stigao 0, 235	porasli 0, 258	nema 0, 274
smrtno 0, 274	dio 0, 234	napravio 0, 234	porasti 0, 256	želi 0, 272
poginula 0, 271	najviše 0, 234	otišao 0, 232	skočio 0, 255	predsjednik 0, 271

Tablica 3.8: Sinonimi prikupljeni izgradnjom vektora funkcijom BOOL s kontekstom širine 2

POBJEDNIK	AUTOMOBIL	VOĐA	USPJEŠAN
automobila 0, 361	pobjednika 0, 330	vođe 0, 267	poginulo 0, 363
auto 0, 323	pobjednica 0, 300	član 0, 253	ranjeno 0, 351
vozila 0, 316	finalu 0, 280	predsjednik 0, 249	ozlijeđenih 0, 319
vozilo 0, 306	pobjede 0, 278	stranka 0, 245	poginula 0, 316
automobilom 0, 282	osvojio 0, 277	stranke 0, 242	ubijeno 0, 307
automobilu 0, 272	prvaka 0, 267	oporbe 0, 237	ozlijeđena 0, 286
vozač 0, 264	pobijedio 0, 265	vlade 0, 232	poginuo 0, 267
sati 0, 263	medalje 0, 263	premijer 0, 232	poginulih 0, 265
kuće 0, 262	pobjedu 0, 262	čelnik 0, 229	ranjenih 0, 264
ulici 0, 258	finala 0, 256	vlast 0, 229	poginuli 0, 261

se proširenjem prozora konteksta kojeg uzimamo u obzir u n-gramske vektore značajki dodati nove korisne informacije. Međutim, uz njih će se dodati i znatan broj riječi nepovezanih s n-gramom, stvarajući dodatan šum na već ovako rijetke podatke. Kao što vidimo u tablici 3.8, rezultati su ipak nešto lošiji nego oni dobiveni s kontekstom širine jedne riječi.

3.2. Generiranje parafraza

Nakon odabira metode za prikupljanje sinonima, potrebno je dobivene sinonime ugraditi u rečenicu koju parafraziramo. Ideja je jednostavna – za svaki n-gram u rečenici (ovisno do koje veličine n-grama se prikupljaju sinonimi) prikuplja se lista sinonima. Sinonimi koji su bolje rangirani trebali bi imati veću povezanost s izvornim n-gramom, pa je veća vjerojatnost da će stvoriti dobru parafrazu.

No, osim ocjene povezanosti sinonima potreban je još jedan mehanizam koji bi ocjenjivao koliko se dobro sinonim uklapa u kontekst u koji ga stavljamo. Primjerice, *dodati* i *reći* su bliskoznačnice u značenju govora, i ispravno bi bilo parafrazirati 3.9 s 3.10. Međutim, *dodati* se ne može zamijeniti s *reći* u 3.11.

Dodao je kako još nikad nije vidio toliku gužvu. (3.9)

Rekao je kako još nikad nije vidio toliku gužvu. (3.10)

Dodao je šećer u čaj. (3.11)

* Rekao je šećer u čaj. (3.12)

Način na koji možemo provjeriti koliko se dobro sinonim slaže s kontekstom je izračunavanje vjerojatnosti n-grama koji nastaju parafraziranjem. Struktura koja pohranjuje ovakve vjerojatnosti naziva se **jezični model**.

3.2.1. Jezični model

Jezični model je statistički model nizova riječi. Model sadrži informacije o vjerojatnostima pojavljivanja određenih kombinacija riječi u tekstu. Najjednostavniji mogući model bi svakoj

riječi dodijelio jednaku vjerojatnost pojavljivanja iza bilo koje riječi. Ako bi jezik imao 100.000 riječi, vjerojatnost da neka riječ slijedi drugu bila bi uvijek $1{:}100.000$, odnosno $0,00001$. U nešto kompleksnijem modelu, riječ bi i dalje mogla slijediti iza bilo koje druge, ali pojavljivala bi se svojom uobičajenom učestalošću. Na primjer, riječ *u* se u korpusu od $7.800.000$ riječi pojavljuje 246.491 puta (dakle, $3,2\%$ riječi su *u*). Za usporedbu, *povrće* se pojavljuje 120 puta. Ove relativne frekvencije možemo iskoristiti za izračunavanje vjerojatnosti pojave riječi — ako prethodi riječ *dotle* vjerojatnost da slijedi *u* će biti $0,032$, a da slijedi *povrće* $0,00001$. No, pretpostavimo da niz koji prethodi glasi *voće i* – u ovom kontekstu smislenije je da slijedi *povrće* nego *u*. Očito umjesto individualne vjerojatnosti riječi treba računati uvjetnu vjerojatnost u odnosu na prethodeću riječ. Vjerojatnost da *povrće* slijedi nakon riječi *i* (što zapisujemo kao $P(\textit{povrće}|i)$) veća je od vjerojatnosti samog povrća.

Slijedeći tu logiku, vjerojatnost niza riječi (koji zapisujemo kao $w_1 w_2 ... w_n$ ili w_1^n) možemo zapisati kao

$$P(w_1, w_2 \cdots, w_n).$$

Pravilom ulančavanja vjerojatnosti dobije se izraz:

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2)\cdots P(w_n|w_1^{n-1})$$

Problem ove formule jest što ne postoji jednostavan način izračuna $P(w_n|w_1^{n-1})$ za velike nizove prethodećih riječi. Ne možemo brojati koliko se puta riječ pojavi iza svakog dugačkog niza — za sve kombinacije trebao bi nam prevelik korpus. Zato se služimo vrlo korisnom aproksimacijom – računat ćemo vjerojatnost riječi samo u odnosu na prethodnu riječ. Bigramski model aproksimira vjerojatnost riječi s prethodećim nizom $P(w_n|w_1^{n-1})$ vjerojatnošću riječi s prethodećom rječju $P(w_n|w_{n-1})$. Drugim riječima, umjesto računanja vjerojatnosti $P(povre|S\ placa\ sam\ donio\ voe\ i)$, ona se aproksimira na vjerojatnost $P(povre|i)$. Uvrštavanjem aproksimacije u gornji izraz dobivamo da je vjerojatnost jednaka:

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_1^{k-1}).$$

Bigramska aproksimacija može se generalizirati i na n-gramsku, ali zbog jednostavnosti i memorijske uštede, model koji ovaj program koristi upravo je bigramski (Jurafsky et al., 2000).

Bigramski jezični model izgrađuje se brojanjem i normaliziranjem bigrama u nekom korpusu. Kroz korpus prolazimo prozorom veličine 2, prebrajamo sve bigrame, te ukupan broj pojavljivanja svakog dijelimo s brojem bigrama koji imaju istu početnu riječ. Vjerojatnost, dakle, glasi:

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)}$$

Ovaj izraz se može pojednostavniti pošto vrijedi da je broj pojavljivanja bigrama koji počinju zadanom riječju w_{n-1} zapravo jednak broju pojavljivanja same riječi w_{n-1}:

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

Osnovni problem ovakvog standardnog modela jest upravo njegova izgradnja iz korpusa. Točnije, problem je u korpusu – koliko god on bio velik, konačan je, i uvijek će nedostajati neki n-grami koji su inače savršeno prihvatljivi i pojavljuju se u jeziku. Ti n-grami, tzv. *nul n-grami*, bit će u modelu pohranjeni s vjerojatnošću 0, što bi značilo da se nikada neće pojaviti, a to očito nije istina. Ovaj problem je dijelom endemičan za n-grame — pošto n-gram ne može koristiti udaljeniji kontekst, sklon je potcijenjivaju vjerojatnosti pojave riječi koja mu nije bila blizu u korpusu iz kojeg je model učio.

Problem se rješava metodom koju nazivamo zaglađivanje (engl. *smoothing*). Ona nanovo vrednuje bigrame vjerojatnosti 0 i pridružuje im pozitivne vrijednosti. Jednostavna inačica zaglađivanja bila bi pribrajanje jedinice svakom bigramu koji se nije pojavio u korpusu, no pokazuje se da takva metoda znatno narušava točnost modela. U jezičnom modelu ovog programa implementirana je inačica nešto kompleksnijeg, Witten-Bellovog zaglađivanja (Jurafsky et al., 2000), čiji se opis nalazi u daljnjem tekstu.

Witten-Bellov postupak zaglađivanja temelji se na jednostavnoj ideji – riječ ili n-gram frekvencije 0 zapravo je događaj koji se još nije dogodio. Kada se dogodi, bit će to prvi put da vidimo taj n-gram, pa zapravo njegovu vjerojatnost možemo modelirati pomoću vjerojatnosti da vidimo bilo koji n-gram po prvi puta. Ovakav koncept korištenja broja pojava koje smo vidjeli jednom da modeliramo broj pojava koje nismo nikada vidjeli i inače se vrlo često koristi u statističkoj obradi jezika. Tako se vjerojatnost da prvi put vidimo n-gram računa brojeći koliko se puta u korpusu po prvi put pojavi neki n-gram. To je vrlo jednostavno jer je broj prvih pojavljivanja n-grama zapravo broj različitih n-grama koji su se pojavili u tekstu.

Shvatimo li korpus kao slijed pronalaženja n-grama (N) i pronalaženja novog tipa n-grama (T), vjerojatnost pronalaženja novog tipa je omjer pronađenih tipova i svih događaja koji su se dogodili:

$$\sum_{i:c_i=0} p_i^* = \frac{T}{N+T}$$

To je ukupna vjerojatnost za sve nul n-grame u modelu, koju treba razdijeliti među njima. Jedna od mogućnosti jest da je podijelimo na jednake djelove. Ako je Z ukupan broj nul n-grama, onda će redistribuirana vjerojatnost svakog od njih biti:

$$p_i^* = \frac{T}{Z(N+T)}$$

Ta nova "količina" vjerojatnosti mora od nekud doći — oduzima se od vjerojatnosti viđenih n-grama na sljedeći način:

$$p_i^* = \frac{c_i}{Z(N+T)} \; ako\; je\; (c_i > 0)$$

Premda za unigrame ima sličan učinak kao metoda dodavanja jedinice, Witten-Bellova metoda pokazuje svoju posebnost kada je raspišemo za slučaj bigrama. Da bismo izračunali vjerojatnost nul bigrama $w_{n-1}w_n$, računamo vjerojatnost pojavljivanja novog bigrama koji počinje s

w_{n-1}. Na taj način je vjerojatnost pojavljivanja nove riječi ovisna o njenom kontekstu. Riječi koje se pojavljuju u malom broju bigrama trebale bi davati manju vjerojatnost pojave nove riječi od onih "druželjubivijih". Ovo se postiže specificiranjem broja tipova T i broja bigrama N na prethodnu riječ w_x:

$$\sum_{i:c(w_x w_i)=0} p^*(w_i|w_x) = \frac{T(w_x)}{N(w_x) + T(w_x)},$$

odnosno u raspodijeljenom obliku:

$$p^*(w_i|w_{i-1}) = \frac{T(w_{i-1})}{Z(w_{i-1})(N(w_{i-1}) + T(w_{i-1}))}$$

Ponovno, tu vjerojatnost treba oduzeti bigramima koji se pojavljuju, pa će izraz za njihovu vjerojatnost biti:

$$\sum_{i:c(w_x w_i)>0} p^*(w_i|w_x) = \frac{c(w_x w_i)}{c(w_x) + T(w_x)}$$

Zaglađivanje je relativno jednostavan, ali jako bitan element u modeliranju jezične statistike. U slučajevima malog korpusa ili korpusa neodgovarajućeg tipa, zaglađivanje će uvelike popraviti dobivene rezultate. Velik broj bigrama koji bi se po osnovnom jezičnom modelu eliminirali zbog vjerojatnosti 0, sada će se uključiti u evaluaciju.

3.2.2. Vjerojatnosni model

Osnovna ideja ovakvog generiranja parafraza temelji se na metodama statističkog strojnog prevođenja. Kao i tamo, i ovdje se pokušava pronaći "prijevod" s najvećom vjerojatnosti. Pretpostavimo da želimo parafrazirati rečenicu F koja se sastoji od fraza $f_1, f_2, \ldots, f_{|F|}$. Neka je N bilo koji kandidat za parafrazu. Najbolja parafraza, označena s N^* je N s najvećom vjerojatnosti da bude parafraza od F, tj:

$$N^* = \underset{N}{\operatorname{argmax}} P(N|F) = \underset{N}{\operatorname{argmax}} \frac{P(N)P(F|N)}{P(F)} = \underset{N}{\operatorname{argmax}} P(N)P(F|N)$$

$P(N)$ predstavlja jezični model, a $P(N|F)$ model prevođenja, odnosno model zamjene parafrazom. Parafraziranje se, međutim, razlikuje od strojnog prevođenja jer ne zahtjeva zamjenu svih riječi. Zato je pogodno uvesti operator identiteta, koji zadanu frazu ostavlja nepromijenjenom. Ako se ovom operatoru pridruži visoka vjerojatnost, parafraziranje će biti konzervativnije, s manje zamjena. Suprotno, niža vjerojatnost poticat će više zamjena.

4. Parafraziranje po pravilima

Problem s prethodnom metodom je što direktno ovisi o kvaliteti i veličini korpusa iz kojeg se računaju modeli. Što je korpus veći, to će baza n-grama biti raznolikija, n-grami će imati veće korpuse za usporedbu, a vjerojatnosti će biti bolje procijenjene. Korpus novinskih članaka spomenut u prethodnom poglavlju jedini je jednojezični usporedivi korpus za hrvatski jezik na raspolaganju.

Prva metoda pristupala je korpusu na najopćenitiji mogući način – sumarno, dajući svakoj frazi jednaku važnost i vrednujući svaki kontekst jednako. Sljedeća metoda koju ćemo predstaviti koristit će specifičnost korpusa – njegovu usporedivost. Korpus novinskih članaka nije paralelan u smislu da postoji potpuna ekvivalencija u sadržajima različitih djelova, ali određena razina preklapanja postoji. Različiti izvori će pisati o istim događajima na slične načine. Među člancima koji pišu o istom događaju veća je vjerojatnost naći parafraziranu rečenicu. Zato će prvi korak naše metode biti grupiranje članaka prema događaju o kojem pišu. Nakon grupiranja, unutar grupa se pronalaze parafrazirane rečenice. Ovo se temelji na sljedećoj pretpostavci: veća je vjerojatnost da su dvije slične rečenice parafraze ako se nalaze u tekstovima koji govore o istom događaju. Treba istaknuti da se ovdje radi o rečeničnim parafrazama, dok se u prvoj metodi radilo o n-gramskim (u slučaju 1-grama, sinonimskim) parafrazama. Nakon što su parafrazirane rečenice uparene, pomoću njih se prikupe pravila koja će se kasnije koristiti za stvaranje novih parafraza.

4.1. Grupiranje događaja

Za proces grupiranja koristi se jednostavna metoda: Za svaki članak izgradi se vektor značajki, pri čemu se za svaku riječ označava koliko se puta pojavila. Pri tome se iz riječi izbace svi interpunkcijski znakovi i brojevi, te se sva slova postave na mala. Također, u značajke se dodaju i riječi naslova skalirane s određenom vrijednosti, u ovom radu s vrijednošću tri. Članci se smještaju u grupe, čiji se centroidi računaju prema vektorima članova grupe. Članak se smješta u grupu čijem je centroidu najsličniji, pod uvjetom da su ispunjen dva uvjeta: da su članci objavljeni u intervalu od najviše četiri dana, te da im je kosinusna sličnost vektora značajki barem $0,4$. Ukoliko nema takve grupe, članak stvara novu grupu u koju se dodaje kao

prvi član.

Na korpusu od 56481 članka, stvoreno je 16901 grupa, što znači da se u grupi nalazi prosječno $3,3$ članaka. Od toga se čak 10633 grupa sastoji samo od jednog člana, pa se neće razmatrati u nastavku metode.

4.2. Prikupljanje parafraza

Nakon što se grupiranjem članaka prema temi smanjila veličina skupa pretraživanja, prikupljanju parafraza može se pristupiti usporedbom rečenica unutar grupe. Za usporedbu odabrano je nekoliko mjera sličnosti, odnosno različitosti:

- **Levenshteinova udaljenost** je funkcija udaljenosti dvaju nizova riječi dana kao najmanji broj promjena koji se treba učiniti kako bi prvi niz postao jednak drugome. Vrste promjena koje se mogu napraviti su kopiranje riječi iz prvog niza u drugi (cijena 0), brisanje riječi iz prvog niza (cijena 1), dodavanje riječi u drugi niz (cijena 1) i zamjena dviju riječi (cijena 1).
- **n-gramsko preklapanje** (engl. *n-gram overlap*) za dani par rečenica računa broj 1-grama, 2-grama, 3-grama, ..., N-grama koji se preklapaju. N je najčešće 4 ili manje. Neka je funkcija koja broji preklapanja $Count_{match}(ngram)$. Za dani $N \geq 1$, normalizirana metrika sličnosti dviju rečenica S_a i S_b koja daje jednaku težinu svakom uparivanju n-grama dana je sljedećom formulom:

$$NGsim(Sa, Sb) = \frac{1}{N} * \sum_{n=1}^{N} \frac{Count_{match}(ngram)}{Count(ngram)}$$

  pri čemu funkcija $Count(ngram)$ označava maksimalni broj n-grama koji postoji u kraćoj rečenici, što je ujedno i najveći mogući broj preklapanja.
- **Jaccardova sličnost** temelji se na sličnosti skupova riječi. Pri tome se kažnjavaju skupovi s malo zajedničkih elemenata, normalizirano s ukupnim brojem članova, odnosno, zapisano pomoću operatora nad skupovima:

$$sim_{setJaccard}(x, y) = \frac{|x \cap y|}{|x \cup y|},$$

- **StrikeMatch sličnost** (White) napravljena je da riješi neke od čestih problema standardnih metrika sličnosti. Algoritam je izgrađen počevši od dva preduvjeta: prepoznavanje sličnosti nizova s malim razlikama u znakovima i robustnost na promjenu redoslijeda riječi. Ideja je pronaći koliko se susjednih parova znakova nalazi u oba niza. Sličnost skupova parova se ocjenjuju se Diceovom metrikom:

$$sim_{setDice}(s1, s2) = \frac{2 * |pairs(s1) \cap pairs(s2)|}{|pairs(s1)| + |pairs(s2)|}$$

Kako bi se evaluirala primjena ovih metrika na rečenicama koje su parafraze, odabrana je rečenica 4.1 iz jednog od novinskih izvora korištenih u korpusu. Iz drugih izvora ručno je prikupljeno pet parafraza te rečenice (4.2-4.6), te pet rečenica koje su povezane, ali nisu parafraze (4.7-4.11).

Američki predsjednik Barack Obama objavio je da su vođu al-Qaide, Osamu bin Ladena, ubili američki vojnici u Abotabadu, nedaleko pakistanskog glavnog grada Islamabada. (4.1)

Predsjednik SAD Barack Obama potvrdio je na vanrednoj press konferenciji u Washingtonu da je vojska SAD izvela napad na vilu u blizini Islamabada i da je Bin Laden ubijen. (4.2)

Vođa terorističke organizacije Al-Qaide Osama bin Laden ubijen je noćas u akciji američke vojske u Abbottabadu 150 kilometara od pakistanskog glavnog grada Islamabada, objavio je u 4 sata ujutro u posebnoj izjavi i obraćanju američki predsjednik Barack Obama. (4.3)

Barack Obama objavio je kako je Bin Laden ubijen u operaciji američkih vojnika nedaleko od Islamabada. (4.4)

Vođa najzloglasnije terorističke organizacije al-Qaide ubijen je u Pakistanu, objavio je američki predsjednik Barack Obama. (4.5)

Američki predsjednik Barack Obama objavio je da je Osama bin Laden, osnivač i vođa al-Qaide, ubijen u Pakistanu. (4.6)

Američki je predsjednik Obama o mogućoj lokaciji bin Ladena bio obaviješten još tijekom prošlog kolovoza. (4.7)

Američki vojnici nakon su okršaja, tvrdi Obama, uspjeli zarobiti i tijelo Osame bin Ladena. (4.8)

Predsjednik Obama istaknuo je i kako niti jedan američki vojnik nije stradao tijekom spomenute operacije. (4.9)

Pakistanski talibani povezani s Al-Qaidom zakleli su se da će osvetiti Osamu bin Ladena, ako se potvrdi da je ubijen u današnjem napadu. (4.10)

Barack Obama je definitivno odlučio ne objaviti fotografije mrtvog tijela Osame bin Ladena. (4.11)

U tablici 4.1 dani su rezultati usporedbe mjera sličnosti rečenica. S $_{LEM}$ su označene mjere

	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	4.10	4.11
$Levenshtein$	25	35	17	22	16	20	21	20	21	21
$Levenshtein_{LEM}$	25	34	16	22	14	20	20	19	21	20
$Ngram$	0,094	0,302	0,230	0,279	0,378	0,118	0,146	0,067	0,116	0,138
$Ngram_{LEM}$	0,115	0,434	0,310	0,295	0,489	0,118	0,204	0,101	0,148	0,220
$Jaccard$	0,211	0,318	0,267	0,310	0,367	0,188	0,194	0,118	0,216	0,161
$Jaccard_{LEM}$	0,243	0,425	0,407	0,357	0,519	0,188	0,233	0,152	0,286	0,241
$Strike$	0,508	0,686	0,604	0,595	0,646	0,455	0,422	0,455	0,441	0,428
$Strike_{LEM}$	0,515	0,708	0,615	0,650	0,710	0,440	0,444	0,519	0,450	0,455

Tablica 4.1: Usporedba funkcija sličnosti rečenica

u kojima su se ulazne rečenice lematizirale prije provođenja kroz algoritam. Levenshteinova mjera označava različitost rečenica, pa je cilj da ispravne parafraze imaju što manju vrijednost, a neispravne što veću. Sve ostale mjere su mjere sličnosti, pa vrijednost ispravnih parafraza trebaju biti što bliže jedinici, a neispravnih što bliže nuli.

Levenshteinova udaljenost daje najgore rezultate – u ovom primjeru neispravne parafraze čak su bolje ocjenjene od ispravnih. To je zbog povezanosti ove mjere s duljinom rečenica koje se uspoređuju. Što je dulja rečenica, više je operacija potrebno obaviti da se rečenica promijeni, pa će i udaljenost biti veća. Lematizacija gotovo uopće ne utječe na rezultate, minimalno smanjujući vrijednosti i na jednoj i na drugoj strani.

Mjera n-gramskih preklapanja daje u globalu solidne rezultate, osim u slučaju 4.2 koji ima neobično malu sličnost s izvornom rečenicom. Ovo je posljedica n-gramskog načina uspoređivanja riječi u nizu, pri čemu svaka permutacija šteti sličnosti (npr. *predsjednik Barrack Obama* i *predsjednik SAD Barrack Obama*). Lematizacija neznatno ublažava problem, ali 4.2 ostaje lošije rangiran od gotovo svih neispravnih parafraza.

Jaccardova mjera ima rezultate slične prethodnima, s tim da razlika s ot1 nije toliko drastična. Zahvaljujući pristupu temeljenom na vreći riječi (engl. *bag-of-words*), problem permutacije riječi je otklonjen i sličnost s ot1 rangira bolje od svih neispravnih parafraza osim jedne. Lematizacija očekivano povećava sličnosti i na jednoj i na drugoj strani, u nekim slučajevima znatno (ot3, ot5)

StrikeMatch mjera pokazuje se najboljom prema točnosti klasifikacije – sve ispravne parafraze ocijenjene su sličnošću iznad $0,5$, a sve neispravne nižom od te granice. Ovi se rezultati slažu s postavkama na kojima je ova mjera konstruirana – otpornost na male promjene (npr. padeže) i na razmještaj riječi. Nedostatak ove mjere je relativno mala razlika između pozitivno i negativno klasificiranih primjera, što ukazuje na to da bi bilo dobro koristiti još jednu mjeru uz ovu. Lematizacija znatno ne mijenja rezultate, što ima i smisla s obzirom da StrikeMatch djeluje na razini znakova, a ne riječi.

Za ekstrakciju parafraza u ovom radu korištene su mjere Jaccard i StrikeMatch. Unutar svake grupe uspoređuju se rečenice i svrstavaju u grupe parafraza. Rečenica će biti uvrštena u grupu ako joj je Jaccardova sličnost sa svim članovima grupe veća od $0,2$ te StrikeMatch mjera veća od $0,5$. Na ovaj način se prikupilo 138683 rečenica u 56122 grupe.

Zbog prirode korpusa, dio rečenica uparenih kao jako slične zapravo nisu parafraze, nego identične rečenice od kojih jedna ima tipkarsku pogrešku. Također je čest slučaj da se kao parafraze klasificira par poput 4.12 i 4.13.

Da je to na otvorenom, u redu, ali dvoranskih se mitinga nisam baš zaželjela. (4.12)

Da je to na otvorenom, u redu, ali dvoranskih se mitinga nisam baš zaželjela – zaključila je Blanka. (4.13)

Ove rečenice, premda bi se u najširoj definiciji mogle shvatiti kao specifična komentatorska parafraza, ne donose nikakvu iskoristivu informaciju o parafraziranju. Zato se iz prikupljenih rečenica izbacuju one koje se razlikuju za manje od pet znakova od svoje parafraze, te one koje se mogu naći kao podniz unutar parafraze. Nakon ovog filtriranja, u skupu ostaje 88373 rečenica u 32187 grupa. Kao procjenu kvalitete ovog skupa, neovisni ocjenjivač je od 50 nasumično odabranih parova parafraza označio 37 kao ispravne parafraze. Najšešće pogreške odnose se na greške u pisanju koje su naknadno ispravljane ili dodatne informacije (sustav takve članke ponekad bilježi kao nove, ovisno je li adresa na kojoj se članak nalazi također promijenjena).

4.3. Stvaranje pravila

Do sada smo pokazali da se uz usporediv korpus s dovoljno različitih izvora mogu prikupiti parafraze visoke preciznosti. Takve parafraze možemo iskoristiti za ekstrakciju uzoraka, odnosno pravila kojima možemo dalje generirati nove parafraze. Zbog nedostatka alata za parsiranje, nemoguće je potpuno egzaktno odrediti strukturu rečenice, i samim time baratanje transformacijama strukture je znatno otežano. U ovom radu se kao zamjena za stablo parsiranja koristi struktura dobivena n-gramskim preklapanjem. Pritom se u dvije rečenice pronalaze najdulji zajednički podnizovi koji se označavaju kao osnovne građevne jedinice rečenica. Cilj je pronaći transformacije ovih jedinica na razini n-grama koje se dovoljno često pojavljuju, a mogu se poopćeniti izvan svog konteksta.

(Kad su)$_1$ (ušli u)$_2$ nju (istražitelji)$_3$ (su)$_4$ (pronašli)$_5$ zbirku od (tridesetak)$_6$ slika i skulptura te nekoliko kutija sa zasad nepoznatom (dokumentacijom)$_7$ (4.14)

(Kad su)$_1$ (istražitelji)$_3$ napokon (ušli u)$_2$ sobicu (pronašli)$_5$ (su)$_4$ još (tridesetak)$_6$ vrijednih umjetnina i gomilu (dokumentacije)$_7$ (4.15)

U nastavku žemo prikazati upotrebu ove metode na prepoznavanje tri različite transformacije, te pravila koja se prikupe njenim korištenjem. Transformacije koje ćemo analizirati su

umetanje/izostavljanje riječi, **promjena redoslijeda** riječi u rečenici, te **promjena oblika** riječi .

4.3.1. Transformacija umetanja/izostavljanja riječi

Transformacija umetanja riječi odnosi se na parafraziranje u kojem u rečenicu dodajemo dodatne riječi koje ne sadrže novu informaciju, ali utječu na sveukupnu čitkost i fluentnost rečenice. Transformacija izostavljanja je inverzna umetanju, a koristi se za sažimanje i pojednostavljenje rečenice. Kao u primjeru 4.16 i 4.17, podložno je interpretaciji jesu li parafraze potpuno informacijski ekvivalentne, ali osnovni zahtjev ove transformacije je da nakon umetanja ili izostavljanja rečenica ostane gramatički ispravna.

Danas, **dakle**, **još uvijek** ne znamo što **točno** izaziva AS i savantizam. (4.16)

Danas ne znamo što izaziva AS i savantizam. (4.17)

Algoritam za pronalaženje ove vrste transformacija koristi parove rečenica za koje smo prethodno utvrdili da su parafraze. Nakon povezivanja parova identičnih n-grama, analiziramo preostale, jedinstvene n-grame. Za svaki od njih provjeravamo postoji li u drugoj parafrazi slučaj gdje se lijevi i desni kontekst n-grama nalaze neposredno jedan uz drugog. Drugim riječima, tražimo identičan niz s izostavljenim jedinstvenim n-gramom. Problem su mali konteksti koji daju lažni dojam umetanja kada se zapravo radi o permutaciji rečenice. Na primjeru vidimo kako se n-gram *je* može činiti umetnutim ako promatramo u kontekstu širine jedne riječi.

Redatelj "Društvene mreže" David Fincher proglašen je najboljim redateljem, a film *je* dobio nagrade i za najbolji scenarij te montažu. (4.18)

Taj je film dobio nagrade i za najbolji adaptirani scenarij i montažu. (4.19)

Kontekst širine 2 smanjuje broj pronađenih umetanja za preko 60%, ali njihova preciznost se znatno poveća. Iznimka su konteksti početka i kraja rečenice koje također prihvaćamo jer nisu podložni permutaciji rečenice. Za svaki n-gram zapisuju se konteksti u kojima se umetanje dogodilo, te se za pravilo prihvaća svaka strana konteksta koja se pojavila dva ili više puta. U tablici 4.2 prikazani su n-grami najčešće korišteni u ovoj transformaciji, zajedno s kontekstima u kojima se najčešće nalaze. Oznake <*s*> i <*e*> predstavljaju kontekst početka i kraja rečenice, dok *L:* i *R:* označuju o kontekstu koje strane se radi.

Među često umetnutim n-gramima našli su se i pojedini specifični za novinski korpus. Na primjer, riječ *Zagreb* se našla umetnuta na početak rečenice 49 puta – uobičajeno je na da se na početak članka stavi mjesto događaja. Slično, n-gram *piše večernji list* našao se umetnut na kraj rečenice 36 puta. Ovakvi primjeri općenito nisu poželjni među pravilima, ali u specifičnom slučaju obrade novinskog teksta bilo bi korisno znati da se Zagreb može izbaciti s početka rečenice.

i (110)	godine (73)	naime (70)	a (63)	danas (63)	no (57)	se (37)
L:<s> (9)	L:2010 (13)	L:<s> (63)	L:<s> (12)	L:je (33)	L:<s> (47)	R:u (6)
L:je (5)	L:2009 (11)	R:u (4)	L:godine (3)	L:su (8)	R:u (5)	L:da (6)
L:će (3)	L:2008 (6)	R:nakon (2)	R:na (5)	L:se (5)	R:za (3)	L:su (5)
R:u (5)	L:1991 (6)		R:što (3)	R:u (6)	R:nisam (2)	L:će (5)
R:za (4)	L:2005 (5)		R:koji (3)	R:da (5)		L:bi (3)
R:po (2)	R:<e> (30)		R:i (3)	R:je (5)		

Tablica 4.2: Frekvencije najčešće umetnutih n-grama i konteksta u kojima se pojavljuju.

Pri generiranju parafraza, ovu transformaciju treba koristiti oprezno. Često pojavljivanje umetnute riječi u kontekstu znači da se može bez znatnog gubitka izostaviti, ali ne i da se može uvijek umetnuti u isti kontekst. Primjerice, riječ *godine* se gotovo uvijek smije izostaviti ako slijedi nakon rednog broja, ali to ne znači da se ista riječ iza svakog rednog broja smije umetnuti. Za umetanje je potrebno uzeti u obzir i vjerojatnost pojavljivanja konteksta bez mogućnosti umetnute riječi. Pritom je najjednostavnije koristiti jezični model.

4.3.2. Transformacija promjene redoslijeda riječi

Transformacija promjene redoslijeda riječima (permutacija), odnosi se na parafraziranje u kojem riječi ili grupe riječi mijenjaju mjesta u rečenici, ali pritom ostaju istog oblika. Prema definiciji, parafraze nastale ovom transformacijom sadrže identičnu informaciju. Međutim, ova se vrsta transformacije rijetko koristi samostalno. Štoviše, najčešće se koristi kako bi prilagodila konstrukciju rečenice uvođenju drugih transformacija, primjerice sinonimije ili promjeni oblika. Na primjerima (1) i (2) prikazana je "čista" permutacija u kojoj se kontekst ne mijenja.

Horoskop za danas: većina **ljudi će se** osjećati dobro. (4.20)

Horoskop za danas: većina **će se ljudi** osjećati dobro. (4.21)

Algoritam za pronalaženje ovog tipa transformacije kao i u prethodnom slučaju grupira par rečenica u podudarne n-grame. U ovom slučaju ne zanimaju nas jedinstveni n-grami, već n-grami zajednički objema rečenicama – elementi permutacije moraju biti prisutni i u jednoj i u drugoj rečenici. Za svaki se podniz takvih uzastopnih n-grama traži se podniz s istim n-gramima u drugoj rečenici. Zgodna prednost ovakvog pristupa je što pronađeni podniz nikad neće biti jednak podnizu iz prve rečenice – kada bi rečenice sadržavale dva jednaka niza uzastopnih zajedničkih ngrama, oni bi bili prepoznati kao jedinstveni zajednički n-gram. Parovi $(niz, permutacija)$ spremaju se zajedno s kontekstima obje strane.

Pokazalo se da, kao i u prethodnom slučaju, bez strožeg uvjeta konteksta preciznost permutacija opada. Na primjerima 4.22 i 4.23 prikazano je pogrešno prepoznavanje permutacije

je (1224)	se (348)	su (332)	i (161)	će (81)	bi (50)
bio (20)	stječe (8)	bili (8)	prostornog	moći (4)	bilo (4)
rekao (14)	zabrana (7)	sudjelovali (6)	uređenja (13)	prilike (4)	to (3)
dobio (14)	nalazi (6)	dobili (5)		tvrtke (4)	
riječ (10)	brojke (6)	upropastili (4)		gospodarski rast (3)	
policija (8)	ne može (5)	banke (4)		biti (3)	
bilo (8)	sastao (30)	podignute (4)			

Tablica 4.3: Frekvencije n-grama najčešće korištenih u permutacijama i dijelova s kojim se permutiraju.

zbog neraspoznavanja sintaksnih cjelina.

U 18 sati **[na središnjem zadarskom trgu nošeni transparenti]** na kojima... (4.22)

U 18 sati na **[središnjem zadarskom trgu nošeni transparenti na]** kojima... (4.23)

Stoga se kao prave permutacije prikupljaju samo nizovi n-grama koji imaju jednaki kontekst s barem jedne strane. U tablici 4.3 prikazani su n-grami koji su se najčešće pojavljivali u permutacijama, te njihovi najčešći parovi. U obzir su uzeti n-gramski nizovi od dva člana. Može se primjetiti kako se kao najčešći članovi permutacija pojavljuju pomoćni glagoli, bilo u službi glagola u određenom obliku (*je rekao* i *rekao je*) bilo u službi imenice s oblikom glagola biti (*je riječ* i *riječ je*).

Korištenje ove transformacije pri generiranju parafraza ne donosi veliku površinsku raznolikost s obzirom da se riječi u rečenici ne mijenjaju. Ipak, ona može biti korisna u kombinaciji s ostalim transformacijama. Permutiranje n-grama može stvoriti oblik na koji se mogu primjeniti transformacije neprimjenjive u originalnom obliku. Primjer takve složene transformacije dan je u nastavku:

Zato se u policijskim krugovima priča o tome. (4.24)

Zato u policijskim se krugovima priča o tome. (4.25)

U policijskim se krugovima priča o tome. (4.26)

4.3.3. Transformacija promjene oblika riječi

Transformacija promjene oblika riječi (infleksija), odnosi se na parafraziranje u kojem riječi mijenjaju gramatičke oblike, najčešće padež ili glagolski vid. Ova transformacija se najčešće događa uz promjenu konteksta, mada to nije nužno. Također, promjena konteksta može biti na razini cijele rečenice, pa se paralelno može dogoditi više infleksija, kao u sljedećem primjeru:

Doris Košta se tereti za to što nije u cijelosti platila porez za 2004. godinu. (4.27)

Doris Koštu terete da nije u cijelosti platila porez za 2004. godinu. (4.28)

zakona	zakonu	30
predsjednika	predsjednik	29
zakon	zakona	24
udruga	udruge	23
temelju	temeljem	23
godina	godine	22
doznaje	doznajemo	21
zavod	zavoda	20
županijskom	županijskome	20
put	puta	19
mogu	može	19

Tablica 4.4: Frekvencije parova riječi najčešće korištenih u infleksiji.

Metoda za pronalaženje infleksija u tekstu prolazi kroz sve parove parafraziranih rečenica, te u svakoj nalazi uparene n-grame. Unutar njih traže se riječi koje su različite, ali lematizacijom daju isti oblik. Lematizacija ne daje uvijek jednoznačno rješenje, već skup mogućih lema. Stoga se za riječi istog lematiziranog oblika uzimaju one s bar jednom zajedničkom lemom.

I ovdje bez strožeg uvjeta imamo nisku preciznost prikupljenih parova. Uglavnom se radi o riječima koje uopće nisu transformacije nego slučajno posjeduju istu lemu kao i original. Zato je uveden uvjet da konteksti bar jedne strane originala i infleksije mora biti jednaki. Time se broj pronađenih infleksija s 40407 smanjuje na 9359 takvih parova. U tablici 4.4 prikazana je lista najčešćih pronađenih infleksijama s uvjetom istog konteksta. Ovakav uzorak je premalen da bi se donosile sigurnije odluke, ali uz upotrebu većeg korpusa rezultati bi bili relevantniji.

Analiza infleksije bila bi mnogo korisnija u slučaju posjedovanja alata za sintaksnu obradu rečenice ili bar za određivanje vrste riječi. Pritom bi se moglo dublje ući u povezanost između konteksta i načina na koji se oblik riječi mijenja. Nažalost, zbog neposjedovanja takvih alata u ovom radu smo ograničeni na povezivanje pojavnih oblika bez sintaksnog i semantičkog razumijevanja.

5. Evaluacija

Pri evaluaciji sustava za generiranje parafraza nije nam u interesu stvoriti što točniju parafrazu, već i stvoriti što veći broj raznovrsnih parafraza. Ova dva cilja odgovaraju zahtjevima za visoku preciznost i visok odziv. Za određeni ulaz s_i, preciznost p_i i odziv r_i sustava za generiranje mogu se definirati kao:

$$p_i = \frac{TP_i}{TP_i + FP_i}, \tag{5.1}$$

$$r_i = \frac{TP_i}{TP_i + FN_i}, \tag{5.2}$$

gdje TP_i predstavlja broj dobro generiranih rezultata za unos s_i, FP_i broj netočnih izlaza, a FN_i broj dobrih izlaza koji se nisu generirali. Međutim, odziv se u ovoj vrsti problema ne može izračunati, jer je broj svih mogućih parafrazi izraza nepoznat, a prema široj definiciji parafraze i beskonačan. Umjesto korištenja odziva, često se koristi mjera prinosa (engl. *yield*) koja predstavlja prosječan broj generiranih izraza, odnosno:

$$y_i = \frac{1}{N} \sum_{i=1}^{n} (TP_i + FP_i) \tag{5.3}$$

Drugi problem predstavlja samo određivanje kvalitete parafraza. Ne postoji šire prihvaćeni skup podataka za usporedbu rezultata sustava, a uspoređivanje rezultata dobivenih na različitim skupovima najčešće nema smisla. Nepostojanje univerzalnog skupa vrlo je vjerojatno uzrokovano činjenicom da je čak i za ograničen skup izraza nemoguće unaprijed odrediti sve moguće parafraze koje se mogu generirati iz njega. Teoretski, ocjena sustava za generiranje mogla bi se donijeti korištenjem sustava za prepoznavanje parafraza koji bi za svaki par rečenica dao pozitivan ili negativan odgovor. Ipak, današnji sustavi za prepoznavanje nisu dovoljno precizni za valjanu uporabu, a mjere sličnosti korištene za ocjenu strojnog prevođenja su prejednostavne da bi prepoznavale parafraze. Štoviše, u našem se slučaju metoda prepoznavanja parafraza koristi kao dio sustava za generiranje, pa bi korištenje iste metode za ocjenu generiranja bilo besmisleno. Drugi i skuplji način je korištenje ljudskih ocjenjivača, što omogućava i kvalitativniju ocjenu sustava. Zhou et al. (2006) koriste od svojih ocjenjivača traže ocjene fluentnosti, ispravnosti i primjenjivosti. Prije svega, bitno je ocjenjivačima predočiti definiciju parafraze

koja se koristi u danom sustavu s obzirom da neslaganja ocjena znaju biti poprilična. Rad sustava se može mjeriti i posredno, promatrajući njegov utjecaj na veći sustav obrade prirodnog jezika čiji je dio (npr. sustava za odgovaranje na pitanja, sažimanje teksta i sl.)

Teško je odrediti i osnovicu za usporedbu rezultata. Barzilay i Lee (2003); Quirk et al. (2004) za usporedbu koriste rečenice s nasumičnom zamjenom riječi njihovim sinonimima prikupljenih iz baze WordNet. Bhagat i Ravichandran (2008) za svaki od izraza čije parafraze generira uzima prvih 1000 rezultata s Interneta, te koristi njih kao korpus za prikupljanje.

U ovom radu evaluaciju smo vršili ručnim ocjenjivanjem. S novinskih portala nasumično je odabrano 50 rečenica čije će se parafraze generirati. Parafraze se ocjenjuju u dvije kategorije: gramatičkoj ispravnosti, i očuvanosti značenja. Ocjena može biti samo pozitivna ili negativna – ovime se pokušala umanjiti moguća razlika u interpretaciji različitih razina valjanosti. Također se vodi računa o tome koliko je promjena napravljano na rečenici i kojom transformacijom. Za svaku transformaciju korištena su pravila prikupljena prema postupcima opisanim u prethodnom poglavlju. Da bi se pravilo upotrijebilo zahtjevalo se poklapanje barem jedne strane konteksta s kontekstom pravila. U tablici 5.1 prikazani su rezultati evaluacije parafraza dobivenih korištenjem transformacija baziranih na pravilima.

Transformaciju umetanja i izostavljanja ovdje smo evaluirali samo kao izostavljanje. Umetanje se može bez pogreške uvesti u većinu rečenica, ali na način koji ne pridonosi kvaliteti parafraze (primjerice, dodavanje riječi *naime* na početak rečenice). Kao tehnika parafraziranja, izostavljanje je mnogo zanimljivije, pogotovo uzevši u obzir primjenu u sažimanju i pojednostavljivanju teksta.

U našoj evaluaciji transformacija izostavljanja se provodi na cijeloj rečenici onoliko puta koliko je moguće. To je moguće relativno rijetko, u prosjeku manje od jednom po rečenici, ali je većini slučajeva ispravno. Pogreške se većinom svode na neispravno eliminiranje *se* ili *je* iz fraza *da se* i *da je*. U nastavku su dani slučajevi ispravnog i neispravnog izostavljanja ovih riječi:

Izgleda da se od parcele treba oduzeti 1316 metara četvornih za pomorsko dobro. (5.4)

Izgleda da od parcele treba oduzeti 1316 metara četvornih za pomorsko dobro. (5.5)

Napomenuo je da se u demokraciji vlast mijenja na izborima, koji će biti nakon potpisa pristupnog ugovora Europskoj uniji. (5.6)

Napomenuo je da u demokraciji vlast mijenja na izborima koji će biti nakon potpisa pristupnog ugovora Europskoj uniji. (5.7)

Posljednji par je i jedini slučaj u kojem je ocijenjeno da parafraza ima izmijenjen sadržaj. Očigledno je ovo rijedak slučaj kada izostavljanje pomoćnog glagola u specifičnom kontek-

Transformacija	Gramatička ispravnost	Očuvanje sadržaja	Broj transformacija po rečenici
Izostavljanje	80%	98%	0,70
Permutacija	62%	100%′	1,06
Infleksija	26%	78%	2,28
Stroga Infleksija	90%	98%	0,22

Tablica 5.1: Ocjene rečenica dobivenih parafraziranjem pomoću pravila

stu može dati novo značenje. Transformacija izostavljanja inače je pouzdana što se očuvanja značenja tiče.

Transformacija permutacije jednako malo mijenja značenja, što je razumljivo uzevši u obzir da riječi u rečenici permutacijom ostaju nepromijenjene. I nju smo provodili maksimalan broj puta na rečenici, analizirajući konačan rezultat. Transformirani dijelovi rečenica su uglavnom glagoli, ali ima i nekih zanimljivih fraza, poput *metara četvornih* i *četvornih metara*. Preciznost ove transformacije je ipak nešto niža – većina pogrešaka nastaju zbog neprimjerenog inverza glagola u perfektu, npr:

Dodao je kako je poznato da ne mogu svi biti zadovoljni. (5.8)

Je dodao kako je poznato da ne mogu svi biti zadovoljni. (5.9)

Inverzi glagola ne bi trebali spadati u jednostavne permutacije jednakog konteksta koje analiziramo, jer se njihovi konteksti obrnu s njima (npr. *X je dodao* i *dodao je X*). Ovakve pogreške su vjerojatno slučaj lažnog konteksta prethodno prikazanog u 4.22 i 4.23. Potencijalno poboljšanje ove transformacije uključivalo bi proširenje na permutaciju konteksta, ali to bi vrlo vjerojatno zahtjevalo prepoznavanje sintaksnih cjelina.

Transformacija infleksije, kako je i predviđeno, daje znatno lošije rezultate. Premda je primjenjiva dvostruko češće nego permutacija, transformacije koje se dobiju su vrlo često gramatički neispravne, kao u sljedećem primjeru:

Dodao je kako je poznato da ne mogu svi biti zadovoljni te predložio nezadovoljnima da ulože prigovor. (5.10)

Dodao je kako je poznato da ne može svi biti zadovoljno te predložio nezadovoljnima da ulože prigovor. (5.11)

Problem je što širina konteksta od jedne riječi nije dovoljna za određivanje točnog gramatičkog oblika riječi. Još jednom, za takve odluke potreban je sintaksni parser. Pogreške u značenju nisu toliko česte, ali se znaju dogoditi kada infleksija slučajno pogodi ispravan gramatički oblik, promjenivši smisao rečenice:

Sinonimska parafraza	Gramatička ispravnost	Očuvanje sadržaja	Broj transformacija po rečenici
$FIRST$	37%	55%	$13,75$
$CONTEXTMAX$	42%	$52\%'$	$13,75$
$CONTEXTMAXMIN$	64%	59%	$6,38$

Tablica 5.2: Ocjene rečenica dobivenih korištenjem sinonimske parafraze

Međutim, ona nije prestajala pjevati iako su joj Randy, Jennifer i Steven govorili da prestane. (5.12)

Međutim, ona nije prestajala pjevati iako su joj Randy, Jennifer i Steven govorili da prestanem. (5.13)

Drugi čest slučaj promjene značenja je zamjena veznika *da* oblikom glagola *dati*. Ipak, infleksija ponekad zna stvoriti i ispravnu zamjenu, većinom u slučajevima određenih i neodređenih pridjeva:

Muškarac je kazneno prijavljen Općinskom državnom odvjetništvu u Ogulinu. (5.14)

Muškarac je kazneno prijavljen općinskome državnom odvjetništvu u Ogulinu (5.15)

Zbog male preciznosti, dodatno smo testirali i "strogu" infleksiju, koja zahtjeva da se obje strane konteksta slažu s pravilom. Kako je vidljivo iz rezultata, preciznost se drastično povećala, ali se u 50 rečenica pronašlo svega 11 mjesta na kojima se transformacija može primijeniti.

U tablici 5.2 prikazani su rezultati evaluacije istih rečenica parafraziranih sinonimskom metodom. Za pronalaženje sinonima korištena je kosinusna sličnost vektora sa značajkama izračunatim uzajamnom mjerom informacije. Zaustavne riječi su izbačene iz korpusa, a širina konteksta je jedna riječ lijevo i desno. Za svaku riječ u rečenici dohvaćene su tri riječi njoj najsličnije. Uspoređujemo tri različita načina odabira zamjenske riječi: $FIRST$ bira riječ najsličniju izvornoj, neovisno o kontekstu; $CONTEXT-MAX$ odabire riječ koja ima najveću vjerojatnost pojavljivanja bilo uz lijevi ili desni kontekst; $CONTEXT-MAXMIN$ odabire riječ s najvećom lošijom vjerojatnosti pojavljivanja uz bilo koju stranu konteksta.

Iz rezultata se da uočiti kako je $FIRST$ najmanje restriktivan po pitanju gramatičke ispravnosti, pa zato i daje najviše potencijalnih parafraza (u prosjeku preko 14 po rečenici). S davanjem većeg značaja kontekstu raste gramatička ispravnost, ali se i smanjuje broj stvorenih fraza. Uzrok tom smanjenju je eliminacija onih za koje jezični model kaže da nisu dovoljno vjerojatne.

Najbitnija razlika između parafraza stvorenih metodom sinonimske supstitucije i onih stvorenih pravilima nije samo u broju, već i u kvaliteti. Sinonimske parafraze imaju puno veću raznolikost izraza, kao što je i vidljivo u sljedećem primjeru:

Moncef Marzouki najavio je 17. siječnja kandidaturu za predsjedničke izbore koji bi se trebali organizirati za šest mjeseci u Tunisu, ocjenjujući "maškaradom" vladu nacionalnog jedinstva u kojoj su na ključna mjesta imenovani brojni ministri iz vlade svrgnutog predsjednika Bena Alija. (5.16)

Moncef Marzouki *izjavio* je 17. siječnja kandidaturu za predsjedničke izbore koji bi se *mogli održati* za šest dana u Tunisu, ocjenjujući "maškaradom" vladu nacionalnog jedinstva u kojoj su na *bitna* mjesta *novoimenovani ostali* ministri iz vlade svrgnutog predsjednika Bena Alija. (5.17)

Premda sve zamijenjene riječi nisu pravi sinonimi niti je sadržaj rečenice ostao potpuno nepromijenjen, očigledna je velika razlika između ove vrste parafraziranja i parafraziranja po pravilima koja smo odabrali.

6. Zaključak

U ovom radu opisan je sustav za automatsko generiranje parafraza. Pokazali smo da je generiranje parafraza složen postupak, ali i sve potrebniji kao dio mnogih složenijih sustava. Strojno parafraziranje općenito je u zadnjih desetak godina postalo iznimno zanimljivo području obrade prirodnog jezika. Rad s parafrazama tako može imati tri cilja: prepoznavanje parafraza, prikupljanje korpusa parafraza i generiranje parafraza. Sustav koji smo razvili koristi sva tri postupka u svrhu što kvalitetnijeg parafraziranja.

U sklopu ovog rada koristili smo dva različita pristupa. Prvi, sinonimsko parafraziranje, ostvaren je uspoređivanjem konteksta parova riječi prikazanih u vektorskom obliku. Pokazalo se da najbolje rezultate daje usporedba kosinusnom mjerom gdje su značajke vektora izračunate mjerom uzajamne informacije, sa zaustavnim rječima izbačenim iz korpusa. Kao potporu sinonimskoj zamjeni koristimo bigramski jezični model s Witten-Bellovim zaglađivanjem. Drugi pristup je koristio ekstrakciju pravila, pri čemu smo zbog veličine korpusa prvo grupirali parafrazirane rečenice. Takve grupe su pružale veću sigurnost da su jednake riječi doista povezane pa se prema njima vršilo prepoznavanje potencijalnih transformacija. U radu smo obradili tri takve transformacije: dodavanje i uklanjanje riječi, zamjenu mjesta riječima, i mijenjanje gramatičkog oblika riječi. Evaluacija je pokazala da parafraziranje pravilima dobro čuva značenje izvorne rečenice, ali ne radi velik broj promjena i zna izazvati gramatičke pogreške. S druge strane, sinonimsko parafraziranje radi mnogo više pogrešaka u značenju, ali i unosi mnogo više raznolikosti u parafraze.

U daljnjem radu trebalo bi pokušati primjeniti iste tehnike na mnogo veći korpus. Premda se pokazao dovoljnim, drugi relevanti radovi koriste korpuse nekoliko tisuća puta veće od našeg. Time bi dobili mnogo više informacija o manje učestalim riječima koje smo morali zanemarivati. Od koristi bio bi i bilo kakav alat za sintaksnu obradu rečenice, pogotovo kod metoda temeljenih na pravilima. Primjerice, permutacija bi za kontekst imala sintaksno stablo čija bi se pravila transformiranja učila. Konačno, trebalo bi razmotriti način kombiniranja različitih metoda parafraziranja, pogotovo pitanja redoslijeda i odabira metoda. Ukoliko bi se pronašla prikladan način za ocijenjivanje dobrote parafraze, ovo bi bio vrlo zanimljiv problem za područje evolucijskog računarstva.

LITERATURA

Krešimir Bagić. *Postoji li jezik fikcije*, stranice 37–49. Zbornik Zagrebačke slavističke škole. Zagrebačka slavistička škola, Zagreb, 2007.

Regina Barzilay i Lillian Lee. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. U *North American Chapter of the Association for Computational Linguistics*, 2003.

Regina Barzilay i Kathleen R. McKeown. Sentence fusion for multidocument news summarization. *Comput. Linguist.*, 31:297–328, September 2005. ISSN 0891-2017.

Rahul Bhagat i Deepak Ravichandran. Large scale acquisition of paraphrases for learning surface patterns. U *Meeting of the Association for Computational Linguistics*, stranice 674–682, 2008.

Ching-Yun Chang i Stephen Clark. Linguistic steganography using automatically generated paraphrases. U *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, stranice 591–599, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5.

M. Connor i D. Roth. Context sensitive paraphrasing with a single unsupervised classifier. U *Proc. of the European Conference on Machine Learning (ECML)*, 9 2007.

James R. Curran i Marc Moens. Improvements in automatic thesaurus extraction. U *In Proceedings of the Workshop on Unsupervised Lexical Acquisition*, stranice 59–66, 2002.

Louise Deléger i Pierre Zweigenbaum. Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. U *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, BUCC '09, stranice 2–10, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-53-4.

Bill Dolan, Chris Quirk, i Chris Brockett. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. U *Proceedings of the 20th international*

conference on Computational Linguistics, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

Sanda M. Harabagiu i Andrew Hickl. Methods for using textual entailment in open-domain question answering. U *ACL*. The Association for Computer Linguistics, 2006.

Zellig Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

D. Jurafsky, J.H. Martin, A. Kehler, K. Vander Linden, i N. Ward. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. MIT Press, 2000.

David Kauchak i Regina Barzilay. Paraphrasing for automatic evaluation. U *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, stranice 455–462, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Rodney d. Nielsen, Wayne Ward, i James h. Martin. Recognizing entailment in intelligent tutoring systems*. *Nat. Lang. Eng.*, 15:479–501, October 2009. ISSN 1351-3249.

Long Qiu, Min-Yen Kan, i Tat-Seng Chua. Paraphrase recognition via dissimilarity significance classification. U *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, stranice 18–26, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-73-6.

Chris Quirk, Chris Brockett, i William Dolan. Monolingual machine translation for paraphrase generation. U *In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, stranice 142–149, 2004.

Satoshi Sekine. Automatic paraphrase discovery based on context and keywords between ne pairs. U *Proceedings of International Workshop on Paraphrase*, Proceedings of International Workshop on Paraphrase, 2005.

Yusuke Shinyama i Satoshi Sekine. Paraphrase acquisition for information extraction. U *Proceedings of the second international workshop on Paraphrasing - Volume 16*, PARAPHRASE '03, stranice 65–71, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

Noriko Tomuro. Interrogative reformulation patterns and acquisition of question paraphrases. U *Proc. of the 2nd Int. Workshop on Paraphrasing*, stranice 33–40, Sapporo, Japan, 2003.

Simon White. How to strike a match. URL `http://www.devarticles.com/c/a/Development-Cycles/How-to-Strike-a-Match`. Pristupio 20-6-2011.

Sander Wubben, Antal van den Bosch, Emiel Krahmer, i Erwin Marsi. Clustering and matching headlines for automatic paraphrase acquisition. U *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, stranice 122–125, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

Yitao Zhang i Jon Patrick. Paraphrase identification by text canonicalization. U *Proceedings of the Australasian Language Technology Workshop 2005*, stranice 160–166, Sydney, Australia, December 2005.

Yujie Zhang i Kazuhide Yamamoto. Paraphrasing spoken chinese using a paraphrase corpus. *Nat. Lang. Eng.*, 11:417–434, December 2005. ISSN 1351-3249.

Shiqi Zhao, Xiang Lan, Ting Liu, i Sheng Li. Application-driven statistical paraphrase generation. U *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, stranice 834–842, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-46-6.

Liang Zhou, Chin-Yew Lin, i Eduard Hovy. Re-evaluating machine translation results with paraphrase support. U *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, stranice 77–84, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-73-6.

Automatsko generiranje parafraza izraza i rečenica hrvatskoga jezika

Sažetak

Parafraze su alternativni način iskazivanja jednog te istog značenja. Automatsko generiranje parafraza važan je zadatak u okviru obrade prirodnog jezika koji svoju primjenu nalazi kod sustava za strojno prevođenje, sustava za odgovaranje na pitanja, kod pojednostavljenja ili generiranja teksta, određivanja sličnosti rečenica i sl. U okviru ovog rada proučeni su statistički postupci generiranja parafraza te postupci temeljeni na pravilima. Generiranje se obavlja na razini rečenice, fraze i riječi. Također je razvijen i podsustav za prepoznavanje rečeničnih parafraza. Razrađen je postupak temeljen na usporedivom jednojezičnom korpusu novinskih članaka te je programski implementiran i eksperimentalo vrednovan nad uzorcima iste domene.

Ključne riječi: generiranje parafraza, prepoznavanje parafraza, parafraza, sinonimi

Automatic Paraphrasing of Croatian Expressions and Sentences

Abstract

Paraphrasing is a method of differently expressing the same meaning. Automatic paraphrase generation is an important task in the area of natural language processing, finding its use in machine translation systems, question answering, text simplification or generation, sentence similarity measurement, etc. In the scope of this paper we describe statistical methods of paraphrase generation in addition to the rule based methods. The system works on both sentence, phrase and word level. A subsystem for sentence level paraphrase identification is also developed. A procedure based on a comparable monolingual corpus of news articles is analyzed, implemented and experimentally evaluated using the samples from the same domain.

Keywords: paraphrase generation, paraphrase identification, paraphrases, synonyms