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. 1814

Postupci ekstrakcije kolokacija iz zbirki tekstova

Davor Delač

Zagreb, rujan 2009.

Sadržaj

Popis slika

Popis tablica

1. Uvod

Ekstrakcija jezičnih fraza iz zbirnih tekstova jedna je od zadaća obrade prirodnog jezika (engl. *NLP – Natural language processing*) kao discipline umjetne inteligencije. Kolokacije su bitan dio svakog terminološkog leksikona te su, zbog činjenice da su svojstvo svakog jezika, nužne u zadaćama poput automatskog prevođenja, određivanje značenja riječi i sinteze prirodnog jezika. Izdvajanje ovakvih pojmova iz velikih zbirnih tekstova prevelik je posao da bi se radio ručno te je potrebno razviti metode koje leksikografima i označivačima olakšavaju ovu zadaću.

U sklopu istraživanja razmatraju se statističke mjere leksičke asocijacije (engl. *lexical association measures*) ili, kraće, asocijacijske mjere (AM) kao alat za ekstrakciju jezičnih fraza iz zbirki tekstova. Razmatraju se dva pristupa dobivanja asocijacijskih mjera: proširenje dvogramskih mjera (Petrović et al., 2008) i evolucija mjera genetskim programiranjem (Šnajder et al., 2009). Učinkovitost dobivenih mjera uspoređuje se pri ekstrakciji kolokacija od dvije, tri i četiri riječi (dvogrami, trigrami i četverogrami). Cilj istraživanja je naći što bolju asocijacijsku mjeru te opravdati opisani postupak ekstrakcije kolokacija. Pritom se predlaže podjela kolokacija koja omogućava da se asocijacijske mjere vrednuju zasebno za pojedine vrste kolokacija. Ocjenjuje se i utjecaj lematizacije na postupak ekstrakcije te razrađuje postupak POS–filtriranja.

U nastavku poglavlja dan je kratak pregled područja ekstrakcije kolokacija te definicija kolokacije kao višerječne imenske skupine. U drugom poglavlju opisan je postupak obrade zbirnih tekstova. Uvodi se definicija zbirnih tekstova te formalni opis tokenizacije, lematizacije i POS–filtriranja. U trećem poglavlju opisane su asocijacijske mjere vrednovane u sklopu istraživanja. Opisane su osnovne dvogramske mjere te metode proširivanja tih mjera na trigrame i četverograme. U četvrtom poglavlju opisan je pokušaj izvođenja novih asocijacijskih mjera genetskim programiranjem. U petom poglavlju opisani su pokusi vezani uz ekstrakciju kolokacija, metode prikupljanja podataka i vrednovanja rezultata pokusa. Rezultati istraživanja su dani u šestom poglavlju. Zadnje poglavlje sastoji se od

diskusije rezultata i izvedenih zaključaka.

1.1. Definicija višerječne imenske skupine

Brojni su načini definiranja kolokacija. Postoje lingvističke definicije poput: "Izrazi od dvije ili više riječi koje odgovaraju standardnom načinu iskazivanja nekog pojma" (Manning i Schütze, 1999). Isto tako brojni autori odabiru definiciju kolokacija kao: "kombinaciju više riječi koje se često supojavljuju" (Benson, 1990) U sklopu ovog rada koristiti će se definicija višerječnih izraza prema Sag et al. (2002), a prilagođena hrvatskom jeziku. U pravom smislu riječi, evaluirat ćemo višerječne imenske skupine s imenicom kao glavom (engl. *multi word expressions, MWEs*). Takvi se pojmovi mogu opisati kao *jezične fraze*, tj. kombinacije riječi koje imaju neko svojstveno značenje u duhu jezika. U svrhu boljeg razumijevanja višerječnih izraza u hrvatskom jeziku uvodi se podjela izraza u pet razreda.

1. *Frazemi (idiomi):* Fraze čije značenje ne proizlazi doslovno iz značenja riječi od kojih se sastoje, tj. fraze kod kojih su jedna ili više riječi upotrjebljene u prenesenom, nedoslovno značenju. *Npr. "rad na crno", "cvijet mladosti", "bačva bez dna", i sl.*

2. *Vlastita imena:* Označavaju ime osobe, institucije, udruge, mjesta, tvrtke, manifestacija, itd. Prva se riječ uvijek piše velikim početnim slovom. *Npr. "Hrvatska poštanska banka", "Stjepan Mesić", i sl.*

3. *Terminološki izrazi:* Izrazi koji označavaju stručne koncepte i objekte iz bilo koje domene, te bi se tipično nalazili u nekom stručnom rječniku. *Npr. "energetska učinkovitost", "koronarna arterija", i sl.*

4. *Ustaljene fraze:* Ustaljeni nazivi u jeziku za koje bi se mogao koristiti i neki drugi izraz, ali se, u principu, ne koristi. *Npr. "oštar pas", "sloboda medija", i sl.*

5. *Ostale učestale fraze:* Fraze koje ne pripadaju vrstama 1-4, ali su učestale u jeziku zbog izvanjezičnih uzroka ili su relevantne u smislu opisa sadržaja teksta. *Npr. "afera premijerovih satova", "zaštita privatnih podataka", i sl.*

Ovakva definicija višerječnih izraza u hrvatskom jeziku potrebna je kako bi se u fazi prikupljanja podataka točno znalo kakve se vrste izraza traže. U postupku

vrednovanja pokušati će se odrediti najbolje mjere za pojedine vrste višerječnih izraza iz grupa 1 - 4. Grupa 5 predstavlja statistički česte izraze koji sami po sebi nisu svojstveni jeziku. Ovakvi izrazi zanimljivi su za probleme klasifikacije dokumenata. Ta činjenica dovoljan je povod da se takvi izrazi uključe u definiciju višerječnih izraza, ali će u većini slučajeva biti uklonjeni iz evaluacijskog skupa. Radi jednostavnosti, u daljnjem tekstu koristiti se izraz *kolokacija* za ovako definirane višerječne imenske skupine. Dodatno se definira i pojam n-gram (engl. *word n-gram*). N-gram je bilo koji niz od n riječi. U sklopu ovog rada koriste se i pojmovi dvogram, trigram i četverogram koji predstavljaju nizove od dvije, tri, odnosno četiri.

1.2. Pregled područja

Automatska ekstrakcija terminologije počinje se razvijati početkom 90ih godina. Među prvim radovima sa područja računalne lingvistike i ekstrakcije terminologije jest rad Church i Hanks (1990). U tom se radu prvi puta razmatraju statističke mjere asocijacije kao opis brojnih lingvističkih pojava poput semantičkih odnosa i leksiko-sintaktičkih pravila slaganja riječi. Uvodi se i mjera uzajamne informacije (engl. *pointwise mutual information, PMI*) kao najučinkovitija mjera.

O primjeni kolokacija u obradi prirodnog jezika provedena su brojna istraživanja. Yarowsky (1993) te Mihalcea i Faruque (2004) razmatraju primjenu kolokacija za razrješavanje višeznačnosti riječi, McCardell Doerr (1995) bavi se sintezom prirodnog jezika, a Orilac i Dillinger (2003) strojnim prevođenjem. Kroz sve radove provlači se problem definicije kolokacija. Najvažniji rad na tu temu je ranije spomenuti rad Sag et al. (2002) koji raščlanjuje višerječne izraze u engleskom jeziku u niz grupa te za svaku grupu daje definiciju sa lingvističkog gledišta.

Jedan od prvih programa za automatsku ekstrakciju terminologije temeljenu na statistici jest *Xtract* (Smadja, 1991, 1993). U navedenim se radovima razmatra i vrednovanje postupaka ekstrakcije kolokacija. Početkom 21. stoljeća ekstrakcija terminologije postaje bitan faktor u računalnoj lingvistici. Razvijaju se računalni programi za automatsku ekstrakciju osposobljeni za obradu velike količine dokumenata u malom vremenu. Najpoznatiji su *Collocate* (Barlow, 2004), *Collocation Extract* (Aroonmanakun, 2000), *Terminology Extractor* (Chamblon Systems, Inc., 2004). Navedeni alati mogu obrađivati zbirke tekstova različitih formata te imaju ugrađenu široku paletu različitih mjera asocijacija.

Većina istraživanja vezanih uz ekstrakciju terminologije posvećena je usavršavanju mjera asocijacije. Tako se u radovima (Petrović et al., 2006) i (Petrović et al., 2008) predlažu prirodna i heuristička proširenja dvogramskih mjera (PMI, DICE i Log-likelihood) za 3-grame i 4-grame. Od alternativnih pristupa poboljšanja mjera asocijacije valja naglasiti rad (Šnajder et al., 2009) u kojem se evolucijskim algoritmom traži optimalna asocijacijska mjera. Prilagodba ovog pristupa koristi se i u sklopu ovog diplomskog rada.

2. Obrada zbirki tekstova

Prilikom postupka ekstrakcije kolokacija stvaraju se liste n-grama rangirane u odnosu na pojedinu mjeru asocijacije. Mjere asocijacije zasnivaju se na frekvencijama pojavljivanja pojedinih n-grama i njihovih dijelova. Postupak određivanja navedenih frekvencija dio je obrade zbirnih tekstova. Obrada zbirnih tekstova sastoji se od četiri faze: tokenizacije, lematizacije, POS filtriranja i brojenja pojavljivanja. U nastavku poglavlja opisan je postupak obrade zbirnih tekstova. Uvodi se definicija zbirnih tekstova i formalizam preuzet iz (Petrović et al., 2006).

2.1. Prikupljanje podataka o n-gramima

Prikupljanje podataka o n-gramima sastoji se od četiri faze: *tokenizacije*, *lematizacije*, *POS filtriranja* i *brojenja pojavljivanja n-grama*. Kako bismo lakše opisali te četiri faze, definirajmo pojmove zbirnog teksta i n-grama.

Definicija 2.1. *Neka je W skup riječi a P skup svih interpunkcijskih znakova. Vrijedi da je $W \bigcap P = \oslash$. Zbir tekstova, u daljnjem tekstu* korpus *(engl.* corpus*), C prikazuje se kao niz od konačnog broja k tokena, tj. riječi i interpunkcija:*

$$C = (t_1, t_2, \ldots, t_k) \in (W \cup P)^k. \tag{2.1}$$

Neka je $W^+ = \bigcup_{n-1}^{\infty} W^n$ skup svih nizova riječi. N-gram se definira kao uređena n-torka $(w_1, w_2, \ldots, w_n) \in W^+$. U prvom koraku korpus se zapisuje kao skup S svih tokena i njihovih pojavljivanja u korpusu:

$$S = \left\{(w, i) \in W^+ \times N : (i \leq k) \wedge (w = t_i)\right\}. \tag{2.2}$$

Slijedeći korak je lematizacija. Riječi u korpusu pojavljuju se u različitim flektivnim oblicima, što dovodi do velikog broja različitih oblika pojedinog n-grama. Kako bi svi oblici jednog n-grama bili svedeni na jedan oblik, koristimo se postupkom *lematizacije*. Bitno je naglasiti da jedan oblik riječi u hrvatskom

jeziku može biti sveden na više različitih lema. Primjer je riječ *"pile"* koje se lematizacijom svodi na:

- *"pile" (imenica, nominativ jednine)* – vrsta životinje
- *"pile" (imenica, nominativ množine)* – alat za rezanje drva
- *"piliti" (glagol, infinitiv)* – čin rezanja nečega pilom

Definicija 2.2. *Lematizacijska funkcija* $lm : W \to \wp(W)$ *je funkcija koja preslikava svaku riječ na skup njenih lema. Ako se riječ* $w \in W$ *ne može lematizirati, tada vrijedi* $lm(w) = \{w\}$.

Od lematizacije dobivamo i dodatnu informaciju o vrsti riječi (engl. *part-of-speech, POS*), koja može biti *glagol, pridjev, imenica, veznik, itd.* Za imenice i pridjeve dodatno se pohranjuje i informacija o padežu. Neka je POS skup svih mogućih POS oznaka. Funkciju pridjeljivanja POS–oznake $pos : W \to \wp(POS)$ definiramo kao preslikavanje riječi $w \in W$ na skup svih POS–oznaka koje joj se pridjeljuju s obzirom na pripadni skup lema $lm(w)$. Ako se riječ $w \in W$ ne može lematizirati, tada joj je POS–oznaka nepoznata i označava se sa "F". Skup POS uzoraka $POS^+ \bigcup_{n=1}^{\infty} POS^n$ je skup svih nizova POS oznaka. Posebnu pozornost pridaje se skupu tzv. zaustavnih riječi (engl. *stop words*). Ovdje spadaju brojevi, prijedlozi, čestice, zamjenice i veznici.

Prilikom obrade velikih korpusa stvara se velik broj n-grama kao potencijalnih kandidata za kolokacije. Kako bi se taj broj smanjio i olakšala obrada velike količine podataka, koristi se postupak filtriranja po vrsti riječi (POS–filtriranje). U ovom postupku jednostavno se definiraju dopušteni POS–uzorci te se za svaki n-gram provjerava je li neki od njemu pridijeljenih POS–uzoraka odgovara nekom od dopuštenih.

Definicija 2.3. *Neka je* $POS_f \subseteq POS^+$ *skup svih dozvoljenih POS uzoraka koji definira POS filter. Za* n*-gram* $(w_1, w_2 \dots w_n)$ *kažemo da zadovoljava uvjete POS filtra ako:*

$$POS_f \cap \prod_{j=0}^{n} pos(w_j) \neq \emptyset \tag{2.3}$$

,gdje Π *predstavlja Kartezijev produkt skupova.*

POS filtar korišten u ovom radu propušta sve n-grame koji u sebi ne sadrže glagole. Ako se neka od riječi unutar n-grama ne može lematizirati, a ostale prolaze filtar, cijeli n-gram se propušta kroz filtar. Ovaj je uvjet bitan pošto moduli za lematizaciju ne mogu pokriti cijeli skup riječi nekog jezika. Nelematizirana riječ u takvom slučaju može biti i imenica, a takav n-gram ispunjava uvjete pro-

laska kroz filtar te je potencijalna kolokacija. U suprotnom slučaju bio bi narušen odaziv sustava.

Posljednja faza obrade korpusa je određivanje frekvencija. Za svaki lematizirani n-gram utvrđuje se koliko se puta pojavljuje danom korpusu. Dobivena frekvencija nije ovisna o flektivnim oblicima n-grama zbog lematizacije. Strogo gledano, konstruirana je funkcija $f : W^+ \rightarrow \mathbb{N}_0$ kao:

$$f(w_1 \dots w_n) = |\{(w_1^{'} \dots w_n^{'}, i) \in S : (1 \leq j \leq n)(lm(w_j) \cap lm(w_j^{'} \neq \emptyset))\}|. \quad (2.4)$$

Na ovaj način obrađen korpus spreman je za ekstrakciju višerječnih izraza. Na osnovi dobivenih frekvencija rade se popisi n-grama poredani prema vrijednosti mjere asocijacije.

Za programsko ostvarenje opisanih postupaka koristi se programski paket *TMT* (engl. *Text Mining Tools*) (Šilić et al., 2007). TMT je programski paket koji, pored ostalog, sadrži podršku za predobradu teksta, poput tokenizacije i lematizacije. Korištenjem tog paketa razvijen je programski sustav *TermeX* za ekstrakciju kolokacija (Delač et al., 2009). Dijelovi programskog sustava TermeX korišteni su za obradu korpusa kod provedenih pokusa.

3. Leksičke asocijacijske mjere

Leksičke asocijacijske mjere statistički su pokazatelji snage veze između dvije riječi. U svom izvornom smislu danom u Manning i Schütze (1999) ove su mjere primjenjive samo na dvograme. Razvoj tih dvogramskih mjera i njihova primjenjivost na n-grame, gdje je $n > 2$ razrađeno je u radu (Petrović et al., 2008). Odabir prave mjere kritični je dio ekstrakcije kolokacija. U ovom poglavlju dan je kratak uvid u asocijacijske mjere za ekstrakciju kolokacija. U prvom djelu dan je uvod u mjere asocijacije te su opisane važnije dvogramske mjere. Drugi dio opisuje proširenja dvogramskih mjera na trigrame i četverograme.

3.1. Asocijacijske mjere za dvograme

Postoji veliki broj dvogramskih mjera koje se danas koriste. Te se mjere dijele na četiri kategorije ovisno od pristupa iz kojeg proizlaze:

- **Sortiranje po frekvencijama** - najjednostavnija mjera koja unatoč jednostavnosti pokazuje vrlo dobre rezultate.
- **Testiranje hipoteze** - mjere provjeravaju nul-hipotezu koja kaže da između dvije riječi ne postoji jača veza od slučajne. Najpoznatije mjere ove vrste su *t–test* i *chi–kvadrat*.
- **Mjere iz teorije informacija** - najpoznatija mjera je PMI (engl. *Pointwise mutual information*), a daje podatak koliko se informacija o pojavi neke riječi na mjestu $i + 1$ poveća ako imamo informaciju o pojavi neke druge riječi na mjestu i.
- **Heurističke mjere** - mjere bez formalne pozadine i interpretacije. Sve mjere se iskazuju i zasnivaju na ideji da se dvije riječi vjerojatnije pojavljuju zajedno nego same. Najpoznatiji primjer takve mjere je Diceov koeficijent, koji pokazuje dobre rezultate na malim korpusima. Ostali su

Kulczinskyjev koeficijent, Ochiaijev koeficijent, Fagerov i McGowanov koeficijent.

Od zanimljivijih mjera valja istaknuti PMI, Diceov koeficijent i chi kvadrat.

Pointwise mutual information (PMI) je dana formulom:

$$PMI(xy) = \log_2 \frac{P(xy)}{P(y)P(y)}, \tag{3.1}$$

gdje su x i y riječi, a $P(x)$, $P(y)$, $P(xy)$ vjerojatnosti pojavljivanja riječi x i y te dvograma xy. PMI u ovom obliku favorizira rijetke događaje pa se ponekad koristi i malo izmijenjeni oblik:

$$PMI^{'}(xy) = \log_2 \frac{f(xy)P(xy)}{P(x)P(y)}. \tag{3.2}$$

Slijedeća mjera je *Diceov koeficijent* dan formulom:

$$DICE(xy) = \frac{2f(xy)}{f(x) + f(y)}, \tag{3.3}$$

gdje su $f(x)$ i $f(y)$ frekvencije pojavljivanja riječi x i y, a $f(xy)$ frekvencija pojavljivanja dvograma xy. Diceov koeficijent je učinkovitiji od PMI-a kod obrade malenih korpusa ili strojnog prevođenja korištenjem bilingualnog korpusa Manning i Schütze (1999).

Zadnja mjera koja se razmatra je *chi–kvadrat* (engl. *Chi–square*) koja je dana formulom:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{3.4}$$

gdje su O_{ij} i E_{ij} stvarna i očekivana vjerojatnost objašnjene u (Manning i Schütze, 1999).

3.2. Proširivanje asocijacijskih mjera

Postoji više načina da se postupak ekstrakcije kolokacija proširi za uporabu na n-gramima gdje je $n > 2$. Ta proširenja najčešće se zasnivaju na dvogramskih mjerama opisanim u prethodnom poglavlju. Jedan od interesantnih pristupa ekstrakciji n-grama proizvoljne duljine dan je u radu Daudaravicius (2009). Autor predlaže korištenje *Diceovog koeficijenta* za određivanje mjere povezanosti susjednih riječi. Tekst se rastavlja na n-grame proizvoljne duljine tako da se gleda povezanost dvije susjedne riječi. Ako je snaga povezanosti tih riječi veća od unaprijed zadanog praga tada su te riječi dio istog pojma. Dobiveni n-grami skaliraju se po frekvencijama pojavljivanja te se dobiva kvalitetna lista kolokacija.

Prednost pristupa je u tome što Diceov koeficijent nije ovisan o veličini korpusa te se dobivaju zadovoljavajući rezultati na razini pojedinih dokumenata. Ovaj pristup ne zahtjeva proširenje dvogramskih mjera pošto se samo određuju povezanosti između dvije susjedne riječi.

Drugi pristup jest proširenje asocijacijskih mjera za dvograme kako bi se one mogle koristiti za n-grame gdje je $n > 2$. Proširenja mjera asocijacije najbolje su opisana u radovima Petrović et al. (2008) i Petrović et al. (2006). Postoje tri pristupa proširivanju dvogramskih funkcija:

1. **Uzorci za proširivanje mjera:** Definiraju se uzorci za proširenje prema kojima se, uz pomoć dvogramske mjere, izračunava vrijednost mjere asocijacije za n-grame proizvoljne duljine.

2. **Heuristička proširenja:** Definiraju se nove asocijacijske mjere temeljene na lingvističkom znanju o položaju riječi unutar kolokacije.

3. **Temeljna proširenja:** Neke mjere kao PMI i Diceov koeficijent imaju temeljna proširenja za n-grame gdje je $n > 2$.

Temeljno proširenje je najjednostavniji oblik proširenja asocijacijske mjere. Temeljna proširenja asocijacijskih mjera su funkcije $g_n : W^n \to \mathbb{R}$, gdje n predstavlja broj riječi koje čine n-gram. U širem smislu definiramo PMI i Diceov koeficijent kao:

$$PMI_n(w_1 \cdots w_n) = \log_2 \frac{P(w_1 \cdots w_n)}{\prod_{i=1}^{n} P(w_i)}, \tag{3.5}$$

$$DICE_n(w_1 \cdots w_n) = \frac{n f(w_1 \cdots w_n)}{\sum_{i=1}^{n} f(w_i)}. \tag{3.6}$$

Chi–kvadrat je definirana sama po sebi za n elemenata te je samo potrebno pratiti tablicu za očekivane i stvarne vjerojatnosti (Manning i Schütze, 1999).

Proširivanje mjera uzorcima za proširenje (engl. *extension pattern, EP*) je slijedeći pristup koji ćemo razmatrati. Formalnu definiciju preuzimamo ponovo iz rada Petrović et al. (2008) i ona glasi:

Definicija 3.1. *Neka je W^+ skup svih n-grama i $\mathcal{F}$ skup asocijacijskih mjera za dvograme definirana kao $\mathcal{F} = \{g | g : W^2 \to \mathbb{R}\}$, gdje je g funkcija koja uzima dvogram kao argument, a vraća realni broj. Uzorak za proširenje je funkcija G koja kao argumente uzima asocijacijsku mjeru g, duljinu n-grama i n-gram te*

vraća vrijednost proširenja od g za zadani n-gram:

$$G : \mathcal{F} \times \mathbb{N} \times W^+ \to \mathbb{R}. \tag{3.7}$$

Postoji beskonačno mnogo ovakvih proširenja. U sklopu ovog rada navesti ćemo samo ona koja se spominju u radu Petrović et al. (2008), a bitna su za ovo istraživanje:

$$G_1(g, w_1 \cdots w_n) = \frac{g(w_1, w_2 \cdots w_n) + g(w_1 \cdots w_{n-1}, w_n)}{2}, \tag{3.8}$$

ovo proširenje daje srednju vrijednost snage veze između prve riječi i zadnjih $(n-1)$ te prvih $(n-1)$ i zadnje riječi.

$$G_2(g, w_1 \cdots w_n) = \frac{g(w_1 \cdots w_{\lfloor n/2 \rfloor}, w_{\lceil n/2 \rceil} \cdots w_n) + g(w_1 \cdots w_{\lfloor n/2+1 \rfloor}, w_{\lceil n/2+1 \rceil} \cdots w_n)}{2}, \tag{3.9}$$

ovo proširenje, u grubo, daje snagu veze između prve polovice n-grama (prvih $\frac{n}{2}$ riječi) i druge polovice n-grama.

Sljedeće proširenje daje nam srednju vrijednost snage veze između svih susjednih riječi u n-gramu:

$$G_3(g, w_1 \cdots w_n) = \frac{1}{n-1} \sum_{i=1}^{n-1} g(w_i, w_{i+1}). \tag{3.10}$$

Proširenje G_4 računa snagu veze između prvih i zadnjih $(n-1)$ riječi unutar n-grama

$$G_4(g, w_1 \cdots w_n) = g(w_1 \cdots w_{n-1}, w_2 \cdots w_n), \tag{3.11}$$

Posljednje bitno proširenje G_5 kao vrijednost asocijacijske mjere daje srednju vrijednost snage veza između dijelova n-grama ako se u obzir uzmu svi načini njegove podjele na dva dijela:

$$G_5(g, w_1 \cdots w_n) = \frac{1}{n-1} \sum_{i=1}^{n-1} g(w_1 \cdots w_i, w_{i+1} \cdots w_n), \tag{3.12}$$

U istraživanjima koje provodi Petrović et al. (2008) pokazuje se da ovako osmišljeni uzorci za proširivanje ne daju dobre rezultate ako se unutar n-grama pojavljuje zaustavna riječ. U tu svrhu razvijaju se heurističke mjere poput trigramske mjere:

$$H(g, w_1 w_2 w_3) = \begin{cases} \alpha_1 G_0^*(g, w_1 w_2 w_3, \{w_2\}) & \text{if } stop(w_2) \\ \alpha_2 G_5^*(g, w_1 w_2 w_3, \varnothing) & \text{otherwise,} \end{cases} \tag{3.13}$$

gdje je $G_0^*(g, w_1w_2w_3, I)$ uzorak koji imitira temeljno proširenje mjere PMI uz iznimku da se članovi skupa I ignoriraju u nazivniku, a G_5^* mjera ista kao G_5 samo uz mogućnost ignoriranja nekih članova n-grama. Kod heurističke mjere H definira se različito ponašanje za POS uzorke sa zaustavnom riječi u sredini. Ova funkcija pokazala je dobre rezultate kod ekstrakcije trigrama, a u kasnijem istraživanju (Šnajder et al., 2009) funkcija slična mjeri H pokazala se kao optimalno rješenje za trigrame. U tu svrhu kao metoda optimizacije koristi se genetsko programiranje slično opisanom u poglavlju 4.

4. Automatsko izvođenje asocijacijskih mjera genetskim programiranjem

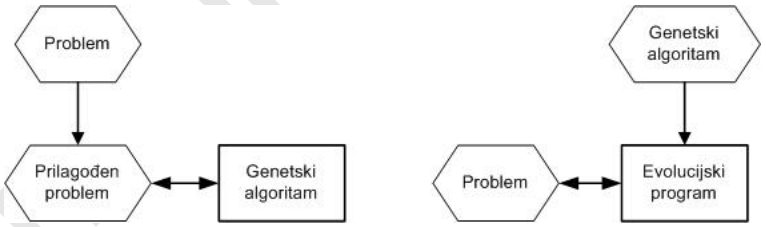
Kako bi pronašli što bolju asocijacijsku mjeru potrebno je pretražiti prostor svih asocijacijskih mjera. Kao što je i ranije navedeno, takvih mjera ima beskonačno mnogo. Ponekad je teško i dugotrajno doći do željenih optimalnih rješenja algoritmima iscrpne pretrage, kao što su *pretraživanje u širinu* ili *A**. Algoritmi koji imitiraju pojave iz prirode mogu dati zadovoljavajuća približna rješenja za teške probleme optimizacije u realnom vremenu. Postoje raznovrsni optimizacijski algoritmi inspirirani prirodnim pojavama. Među najpoznatijima su: genetski algoritam, algoritam kolonije mrava, algoritam roja čestica, algoritam umjetnog imunološkog sustava i dr. Navedeni algoritmi sve više dobivaju na važnosti pojavom jačih računala. Danas se ti algoritmi koriste na poljima matematike, fizike, ekonomije, bioinformatike, računalne znanosti, itd. Neke od primjena su rješavanje numeričkih problema, optimizacija funkcija, kao i svakodnevni problemi poput izrade rasporeda ispita na sveučilištu.

U ovom poglavlju opisana je primjena genetskog programiranja za evoluciju optimalne asocijacijske mjere. U prvom djelu dan je uvod u genetske algoritme. U drugom dijelu opisana je primjena genetskog algoritma na konkretni problem evolucije funkcija.

4.1. Genetski algoritmi

Genetski algoritmi (Golub, 2004) oponašaju pojavu evolucije u prirodi. Apstraktni prikazi kandidata rješenja, zvan *kromosom*, predstavlja pojedino rješenje problema.*Populacija* je skup kromosoma. Cilj algoritma jest genetskim operatorima djelovati na kromosome unutar populacije kako bi se razvilo rješenje blisko op-

timalnom. Kromosomi se najčešće prikazuju kao binarni nizovi, ali mogući su i drugi prikazi kao što će biti opisano u sljedećem poglavlju. Binarni nizovi mogu se kodirati na više načina, a najčešći su prirodni binarni kod i Grayev kod. Evolucija počinje populacijom slučajno stvorenih kromosoma te se odvija u iteracijama zvanim *generacije*. U svakoj generaciji koriste se slučajno odabrani kromosomi te se na osnovi njihove funkcije dobrote i genetskih operatora stvara nova populacija za novu iteraciju algoritma. Algoritam se izvodi dok nije zadovoljen uvjet zaustavljanja. Uvjet zaustavljanja može biti zadovoljenje odgovarajućeg iznosa funkcije dobrote, prekoračenje ranije zadanog broja iteracija ili neki drugi uvjet koji navodi da je trenutno rješenje optimum koji se traži. Dva su pristupa rješavanju problema genetskim algoritmom: *prilagodba problema genetskom algoritmu* i *prilagodba genetskog algoritma problemu*. Dijagrami rješavanja problema dani su slikom 4.1. Za rješavanje problema prvim pristupom postoje specijalizirani genetski algoritmi zvani *evolucijski programi* kojima se prilagode strukture kromosoma i genetski operatori kako bi se efikasno riješio problem. Kod drugog pristupa definira se genetski algoritam usko specijaliziran za rješavanje određenih problema.



Slika 4.1: Pristupi rješavanju problema pomoću genetskog algoritma

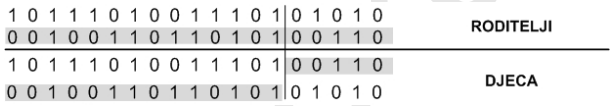
Iz prethodnog objašnjenja jasno je da se moraju definirati tri stvari za svaki genetski algoritam:

- *genetski prikaz domene*,
- *genetski operatori*,
- *funkcija dobrote*.

Funkciju dobrote definiramo kao mjeru kvalitete rješenja. Ta mjera uvijek mora biti prilagođena problemu koji se rješava. Jednostavan primjer je traženje ekstrema neke funkcije gdje sama funkcija predstavlja funkciju dobrote, npr. $f(x) = x^2 + 2x + 1$. Ako se traži minimum, kao bolja vrijednost uzima se što manji iznos funkcije dobrote, a ako se traži maksimum cilj je imati što veći iznos

funkcije dobrote. U danom primjeru kvadratne funkcije ekstrem je minimum u točki $x = -1$ te je iznos funkcije dobrote za taj minimum $f(-1) = 0$. U ovom slučaju bolja je što manja vrijednost.

Genetski operatori su *operator križanja* i *operator mutacije*. Operator križanja binarni je operator koji kao parametre uzima dva kromosoma te daje novi kromosom koji sadrži dio genetskog materijala prvog kromosoma i dio genetskog materijala drugog kromosoma. U klasičnom prikazu kromosoma binarnim nizom postoji više izvedbi križanja. Kod križanja definiranog prekidnim točkama uzima se proizvoljan broj točaka u kromosomu između kojih se uzima genetski materijal pojedinog roditelja. To bi značilo da ako uzmemo jednu prekidnu točku, koju slučajno odaberemo, kao djecu bi dobili dva kromosoma prikazana slikom 4.2. Drugi najčešći način križanja je uniformno križanje gdje se dijete dobiva kao



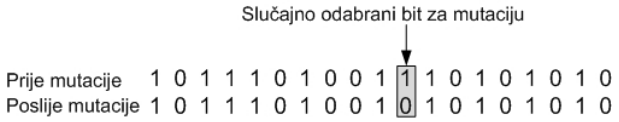
Slika 4.2: Prikaz križanja dvaju binarno kodiranih kromosoma

rezultat sljedeće logičke funkcije:

$$DIJETE = A \wedge B \vee R \wedge (A \otimes B), \tag{4.1}$$

gdje su A i B binarne reprezentacije kromosoma, a R kromosom dobiven slučajnim procesom. Kod ovakvog križanja rezultat je kromosom koji na mjestima gdje se roditeljima poklapaju vrijednosti bitova ima iste vrijednosti kao i roditelji, a na ostalima slučajno izabranu vrijednost bita prvog ili drugog roditelja. Mutacija je unarni operator koji služi za stvaranje novog genetskog materijala. Mutacija je promjena jednog ili više gena unutar kromosoma. Ovaj operator ključan je da algoritam ne bi zaglavio u lokalnom optimumu funkcije dobrote uslijed lošeg slučajnog izbora početne populacije. Kod binarnih prikaza kromosoma mutacija se izvodi tako da se, s prethodno definiranom vjerojatnošću mutacije, promjeni vrijednost slučajno odabranog bita. Primjer je dan slikom 4.3.

Struktura genetskog algoritma može se opisati jednostavnim pseudo–kodom koji je dan algoritmom 1. U kodu algoritma vidljive su četiri osnovne faze: *stvaranje populacije*, *selekcija*, *križanje* i *mutacija*. U fazi stvaranja populacije slučajnim se odabirom određuje ranije definirani broj slučajno generiranih kromosoma. Ti kromosomi čine početnu populaciju. Broj kromosoma u početnoj



Slika 4.3: Prikaz mutacije binarno kodiranog kromosoma

Algoritam 1 Genetski algoritam

1: $t \leftarrow 0$;
2: generiraj početnu populaciju $P(t)$;
3: **while** ¬ ispunjen uvjet zaustavljanja **do**
4: $t \leftarrow t + 1$;
5: selektiraj $P^{'}(t)$ iz $P(t-1)$;
6: križaj jedinke iz $P^{'}(t)$ i djecu spremi u $P(t)$;
7: mutiraj jedinke iz $P(t)$;
8: **end while**

populaciji zove se *"veličina populacije"*. Nakon ove faze slijedi petlja u kojoj se stvaraju nove populacije dok se ne zadovolji uvjet zaustavljanja algoritma. Stvaranje nove populacije sastoji se od tri koraka. Prvi korak je *selekcija*. U grubo, selekcija je postupak izdvajanja kromosoma čijim će se križanjem stvoriti nova populacija. Postoji više metoda selekcije kao što su: *jednostavna selekcija*, *eliminacijska selekcija*, *turnirska selekcija*, itd. Križanje i mutacija su koraci u kojima se koriste prethodno definirani genetski operatori. Kod križanja se iz populacije odabrane selekcijom stvaraju nove jedinke koje će biti dio nove populacije. Kod mutacije jedinke nove populacije mutiraju uz vjerojatnost određenu parametrom algoritma. Nakon faze mutacije stvorena je nova populacija koja ulazi u sljedeću iteraciju algoritma.

4.2. Primjena na optimizaciju asocijacijskih mjera

Genetski algoritam korišten za rješavanje problema optimizacije asocijacijskih mjera razvijen je metodom prilagodbe algoritma problemu. Riječ je selekcijskom algoritmu s 3-turnirskom selekcijom. Kromosome čine matematički izrazi koji predstavljaju asocijacijske mjere. Svaki kromosom modeliran je stablastom strukturom (Koza, 1992; Šnajder et al., 2009; Sikirić, 2007). Listove stabla čine konstante ili statističke informacije o n-gramima. Takve informacije mogu biti frekvencija pojavljivanja n-grama, vjerojatnost pojavljivanja n-grama i ukupan

broj različitih n-grama. Statistička informacija o n-gramu na primjeru trigrama abc može biti frekvencija $f(abc)$ samog trigrama, frekvencije dvograma $f(ab)$ i $f(bc)$ ili frekvencije pojavljivanja riječi $f(a)$, $f(b)$ i $f(c)$. Svi ovi podatci bitni su za dobivanje niza izvedenih asocijacijskih mjera. Dodatno, čuva se i lingvistička informacija o pojedinom n-gramu u formi POS oznaka svake riječi koja ga čini. Unutarnji čvorovi su operatori. Postoje tri vrste operatora koje se mogu pojaviti kao unutarnji čvorovi stabla, i to:

1. *Unarni operator* - Prirodni logaritam ($ln(x)$),
2. *Binarni operatori*:
   (a) Zbrajanje - $sum(x, y) = x + y$,
   (b) Oduzimanje - $sub(x, y) = x - y$,
   (c) Množenje - $mul(x, y) = xy$,
   (d) Dijeljenje - $div(x, y) = \frac{x}{y}$,
3. Ternarni operator - operator $IF - THEN - ELSE$.

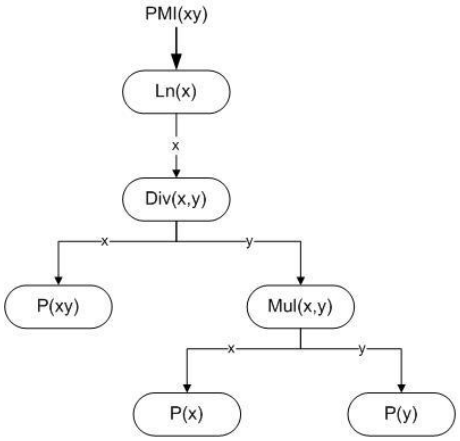
Operator $IF - THEN - ELSE$ nije ništa drugo nego funkcija $ITE : \mathbb{R} \times \mathbb{R} \times \{0, 1\} \to \mathbb{R}$ za koju vrijedi:

$$ITE(a, b, c) = \begin{cases} a & \text{ako} c = 1 \\ b & \text{ako} c = 0 \end{cases}. \tag{4.2}$$

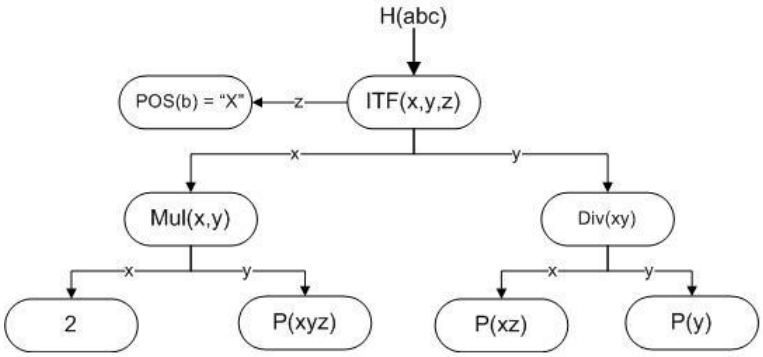
Vrijednost parametra c određuje se prema POS oznaci određene riječi. Uvjet provjere zadan je kao dio ITE operatora, npr. $POS(w_3)$ =“Nn” predstavlja uvjet da POS oznaka treće riječi mora biti imenica u nominativu. Dopuštene POS oznake su: “Nn”, “Ng”, “Nd”, “Na”, “Nv”, “Nl”, “Ni”, “An”, “Ag”, “Ad”, “Aa”, “Av”, “Al”, “Ai”, “X”, “F”; koje redom predstavljaju imenice u svim padežima (“Nx”), pridjeve u svim padežima (“Ax”), zaustavne riječi i riječi koje se ne mogu lematizirati. Ovako opisana asocijacijska mjera može se dobiti i izračunati pravilnim obilaskom po stablu. Kao primjer slikom 4.4 4.3. dano je stablo za mjeru PMI, a slikom 4.5 dano je stablo za funkciju:

$$H_{primjer}(w_1w_2w_3) = \begin{cases} 2P(w_1w_2w_3) & \text{ako } “X” \in POS(w_2) \\ \frac{P(w_1w_3)}{P(w_2)} & \text{inače,} \end{cases} \tag{4.3}$$

4.3.

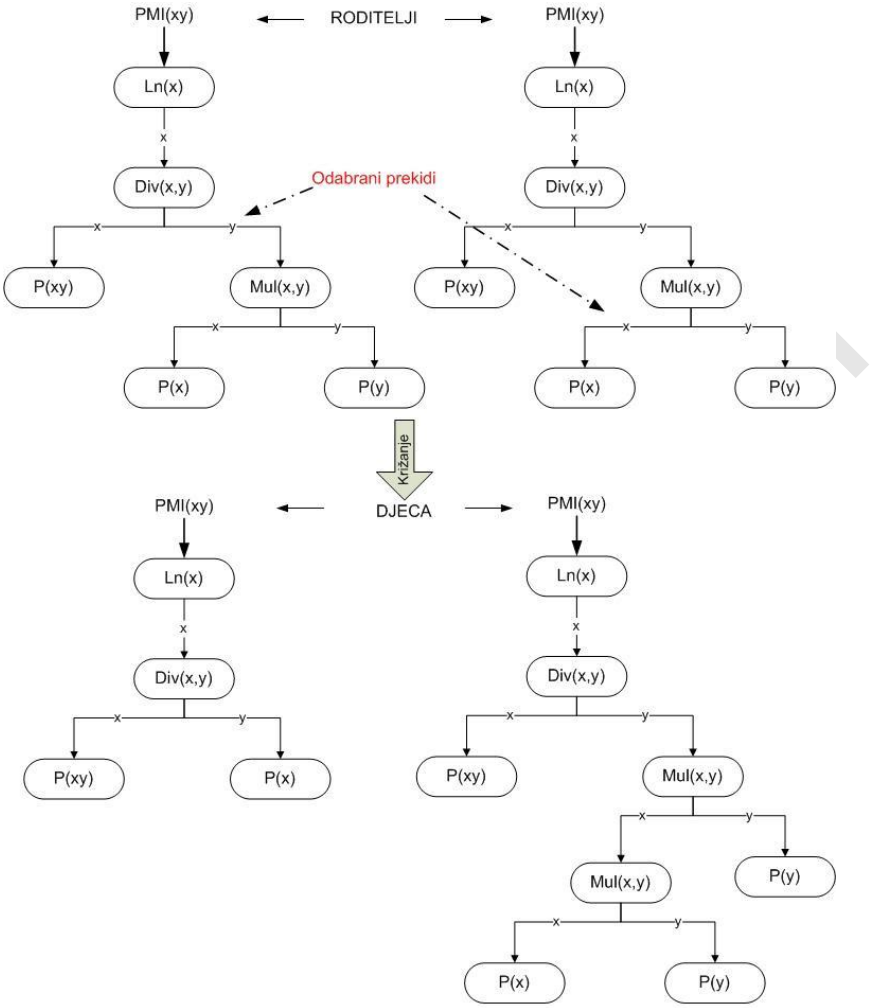


Slika 4.4: Kodiranje funkcije PMI stablom



Slika 4.5: Kodiranje funkcije $H_{primjer}$ stablom

Operator križanja kombinira dva kromosoma kako bi se dobilo novo rješenje. Križanje je ostvareno principom sličnim križanju sa jednom točkom prekida. Slučajnim odabirom uzimaju se po jedan čvor iz svakog kromosoma pod uvijenom da izabrani čvorovi nisu listovi niti čvorovi uvjeta operatora $IF - THEN - ELSE$. Odabranim čvorovima mijenjaju se podstabla i to tako da se kod binarnih i ternarnih operatora slučajnim principom izabere podstablo koje se mijenja slučajno izabranim podstablom drugog roditelja. Princip je preuzet iz rada Gordon et al. (2006), a korišten u radu Šnajder et al. (2009). Slika 4.6 prikazuje princip križanja dvaju kromosoma. Operator mutacije može rezultirati dvama učincima. Prvo, s vjerojatnošću od 25% može obrisati slučajno izabran čvor iz stabla. Kad se čvor obriše, jedno od njegove djece dolazi na njegovo mjesto, pri čemu operator $IF - THEN - ELSE$ ne može biti zamijenjen čvorom uvjeta. Drugi oblik djelovanja mutacije, uz vjerojatnost od 75% jest dodavanje čvora na slučajno mjesto u stablu. Ako dodani čvor nije unarni operator, tada se generira potreban broj



Slika 4.6: Prikaz križanja dva stablasta kromosoma

podstabla da popune mjesta djece.

Funkcija dobrote temelji se na mjeri F_1. Mjera F_1, kao i mjere preciznosti i odaziva, opisana je u poglavlju 5.1. Kao maksimalni iznos mjere F_1 na skupu za evaluaciju uzima se dobrota kromosoma. Dobrota kromosoma proširena je i na način da favorizira kraća rješenja po broju čvorova. To je ostvareno uvođenjem "kazne" na broj čvorova kromosoma. Konačna funkcija dobrote preuzeta je iz rada (Šnajder et al., 2009) i glasi:

$$\mathit{fitness}(j) = F_1(j) + \eta \frac{L_{max} - L(j)}{L_{max}}, \tag{4.4}$$

gdje je $F_1(j)$ mjera F_1 za kromosom j, η koeficijent kazne, L_{max} maksimalan broj čvorova, a $L(j)$ broj čvorova u j-tom kromosomu.

5. Ispitivanje učinkovitosti asocijacijskih mjera

Kako bi se odredila što bolja asocijacijska mjera za ekstrakciju kolokacija potrebno je na pravilan način vrednovati njihovu učinkovitost. Kako bi postupci vrednovanja bili što učinkovitiji posebnu pažnji treba obratiti na podatke koji se koriste u postupku, kao i na korpuse nad kojima se radi ekstrakcija. U ovom poglavlju opisane su korištene metode prikupljanja podataka za vrednovanje i same metode vrednovanja asocijacijskih mjera.

5.1. Korpusi i podatci za evaluaciju

Za sve pokuse koristi se korpus *"Glas Slavonije"*. "Glas Slavonije" korpus je novinskih članaka izdanih u istoimenom listu koji se sastoji od 31.056 novinskih članaka, to jest, 27.594.874 riječi od čega je 423.160 različitih riječi. Nakon provedene obrade korpusa dobiva se 4.314.773 dvograma, 11.405.863 trigrama i 14,232,860 četverograma.

Kako bi se mogli vrednovati postupci ekstrakcije potrebno je ručno označiti skup za vrednovanje. Skup za vrednovanje sastoji se od 3000 n–grama i to po 1000 dvograma, trigrama i četverograma. Izrazi u skupu za vrednovanje dobiveni su slučajnim odabirom iz skupa n–grama koji prolaze POS filtar definiran u poglavlju 2.1. Skup je ručno označavala grupa studenata. Svaki je student morao posebno proći kroz sve liste i označiti kolokacije poštujući pritom definiciju iz poglavlja 1.1. Potrebno je bilo i označiti kojoj vrsti višerječne imenske skupine označene kolokacija pripada. Kako bi se ustvrdila kvaliteta ovako prikupljenih podataka računa se *kappa koeficijent* (Krenn et al., 2004). Kappa koeficijent mjera je slaganja među označivačima. Prednost ove mjere jest u tome što odstranjuje pogrešku u mjeri slaganja koja nastaje kod označavanja slučajnim odabirom. Ove se pogreška javlja ako oba označivača znaju koji je omjer prihvaćenih i odbačenih

kandidata te na osnovi toga slučajno označe uzorak. Kappa koeficijent dan je izrazom:

$$\hat{\kappa} := \frac{p_o - p_c}{1 - p_c}, \tag{5.1}$$

gdje je p_o omjer slaganja između dva označivača, a p_c slučajno slaganje. Slučajno slaganje računa se prema izrazu:

$$p_c = p_1 p_{1\cdot} + p_2 p_{2\cdot}, \tag{5.2}$$

gdje su vrijednosti p_1, $p_{1\cdot}$, p_2, $p_{2\cdot}$ dane tablicom 5.1. Oznake $A+$, $A-$, $B+$, $B-$ predstavljaju slučajeve pozitivnog ili negativnog označavanja dvaju označivača koji se uspoređuju.

Tablica 5.1: Parametri kappa koeficijenta.

	$A+$	$A-$	$+$
$B+$	p_{11}	p_{12}	p_1
$B-$	p_{21}	p_{22}	p_2
$+$	$p_{1\cdot}$	$p_{2\cdot}$	n

Kod kappa koeficijenta definirana su tri intervala koja daju informaciju o različitim razinama slaganja (Green, 1997):

1. $\kappa < 0.4$ - malena razina slaganja
2. $0.4 \leq \kappa \leq 0.75$ - osrednja razina slaganja
3. $\kappa > 0.75$ - visoka razina slaganja.

Za izvođenje pokusa dovoljno dobra mjera sličnosti označenih uzoraka biti će kappa vrijednost kappa–koeficijenta veća ili jednaka 0.6 ($\kappa \geq 0.6$). Uzorci koji pokazuju manju mjeru sličnosti biti će odbačeni. Ako ne postoje dva uzorka koja ispunjavaju taj uvjet smatra se da zadatak označavanja nije dobro definiran te je nepotrebno provoditi pokuse za te slučajeve.

5.2. Postupci evaluacije mjera asocijacije

U sklopu vrednovanja postupka ekstrakcije kolokacija provode se pokusi nad dva ključna segmenta: *mjerama asocijacije* i *postupku obrade korpusa*. Prvi dio vrednovanja postupka obrade korpusa je opravdavanje postupka lematizacije. Lematizacija je postupak svođenja flektivnih oblika riječi njihovu zajedničku lemu.

Postupak donosi bolju informaciju o broju n-grama, ali unosi dodatnu složenost u sam postupak obrade korpusa. Pokusom je potrebno utvrditi koliko bolje rezultate daje postupak obrade korpusa sa lematizacijom u usporedbi s postupkom bez lematizacije. Pošto uspoređujemo samo koliko se ta dva pristupa razlikuju, uzimajući pritom u obzir da je pristup s lematizacijom precizniji, mjera za ovaj pokus biti će *Kendallov tau– koeficijent za rangirane liste* (Kendall, 1955). Kendallov tau–koeficijent predstavlja mjeru suodnosa ili sličnosti dvaju rangiranih listi.

Definicija 5.1 *Neka su X_1 i X_2 dvije rangirane liste koje sadrže sve elemente skupa X te neka je $ind : X \times \{X_1, X_2\} \to \mathbb{N}$ funkcija koja daje položaj elementa skupa X u listama X_1 i X_2. Uređeni par $(p_1, p_2) \in X^2$ je konkordantan s obzirom na liste X_1 i X_2 ako vrijedi:*

$$sign(ind(p_1, X_1) - ind(p_2, X_1)) = sign(ind(p_1, X_2) - ind(p_2, X_2)), \tag{5.3}$$

dodatno, za isti uređeni par kažemo da je diskonkordantan ako vrijedi:

$$sign(ind(p_1, X_1) - ind(p_2, X_1)) = -sign(ind(p_1, X_2) - ind(p_2, X_2)). \tag{5.4}$$

Kendallov tau koeficijent za dvije rangirane liste koje sadrže iste elemente računa se kao:

$$\tau = \frac{n_c - n_d}{\frac{n}{2}n(n-1)}, \tag{5.5}$$

gdje je n_c broj konkordantnih parova, a n_d broj diskonkordantnih parova. Kendallov tau koeficijent poprima vrijednosti iz intervala $[-1, 1]$. Interpretacija vrijednosti 1 je da su dva niza identična, -1 znači da su dva niza suprotno poredana, a vrijednost 0 znači da nema sličnosti između dva niza. Kendallov tau–test provesti će se uz korištenje svih asocijacijskih mjera iz poglavlja 3. za dvograme, trigrame i četverograme kako bi se dobili precizniji rezultati.

Jedan od čestih pokazatelja učinkovitosti postupaka ekstrakcije kolokacije je mjera F_1. Kako bismo definirali mjeru F_1 potrebno je prethodno definirati pojmove *preciznosti* i *odaziva*. Neka je X skup n-grama dobiven postupkom ekstrakcije, a X_{pos} skup pozitivno označenih n-grama. Preciznost definiramo kao:

$$P = \frac{|X_{pos} \cap X|}{|X|}, \tag{5.6}$$

a odaziv kao:

$$R = \frac{|X_{pos} \cap X|}{|X_{pos}|}. \tag{5.7}$$

Kod metoda ekstrakcije kolokacija često se događa da su ova dva pokazatelja u negativnoj vezi, što znači da povećanjem preciznosti pada odaziv i obrnuto.

Kako bismo imali optimalni pokazatelj učinkovitosti tih metoda koristimo mjeru F_1. Mjera F_1 posebni je slučaj F_β mjere gdje je $\beta = 1$. Te dvije mjere dane su formulama:

$$f_\beta = (1 + \beta^2)\frac{PR}{\beta^2 P + R}, \tag{5.8}$$

$$f_1 = 2\frac{PR}{P + R} \tag{5.9}$$

Za usporedbu mjera asocijacija koristiti će se Kendallov tau i srednja preciznost (engl. *average precision, AP*) Hull (1993). Srednja preciznost ima naglasak na dobivanju što većeg broja pozitivno označenih elemenata u rangiranoj listi. Za niz X_1 računa se prema sljedećoj formuli:

$$AP = \frac{\sum_{r=1}^{N} P(r)rel(r)}{n_{pos}}, \tag{5.10}$$

gdje je N broj elemenata niza X_1, $P(r)$ je preciznost na podnizu od prvih r članova niza X_1, vrijednost n_{pos} je broj pozitivno označenih članova niza i $rel : [1, N] \rightarrow 0, 1$ je binarna funkcija za koju vrijedi $rel(r) = 1$ ako je r-ti element niza X_1 pozitivno označen. Za ovaj dio pokusa koristiti će se označena lista za vrednovanje. Ovaj postupak koristiti će se za usporedbu asocijacijskih mjera za pojedine tipove kolokacija definiranih u poglavlju 1.1. Ako dvije asocijacijske mjere imaju visoki Kendallov tau–koeficijent zaključak je da je mjera s većom vrijednosti AP daje bolje rezultate, a da druga mjera daje praktički istu informaciju (ekstrahira i slično rangira sličan skup višerječnih izraza) te je nevažna za postupak ekstrakcije.

Učinkovitost rada genetskog algoritma procjenjuje se dobrotom dobivenog rješenja. Traži se asocijacijska mjera s najboljim vrijednosti F_1. Vrijednost F_1 dobiva se ocjenom nad skupom za vrednovanje. Skup za vrednovanje prvo se rangira prema padajućoj vrijednosti mjere asocijacije. Nad dobivenom rangiranom listom računa se maksimalna vrijednost F_1. Dok se prolazi kroz listu na svakoj poziciji računa se F_1 vrijednost te se pamti njen najveći iznos. Dobivena vrijednost F_1 uspoređuje se sa F_1 vrijednostima dobivenim za asocijacijske mjere iz poglavlja 3.

6. Rezultati

U ovom poglavlju dani su rezultati pokusa opisanih u poglavlju 5. Prvo se razmatra postupak ručnog označavanja kolokacija za izradu skupa za vrednovanje. Potom su dani rezultati usporedbe postupka ekstrakcije s lematizacijom i bez nje. U trećem podpoglavlju uspoređene su mjere asocijacije iz poglavlja 3., dok su u zadnjem poglavlju dani rezultati rada genetskog programiranja.

6.1. Rezultati izrade skupa za vrednovanje

Kako bi se napravio dobar uzorak za vrednovanje, u označavanju je sudjelovalo šest studenata: A, B, C, D, E, F. Studenti su neovisno označavali kolokacije u skupovima slučajno izabranih n–grama iz korpusa "Glas Slavonije". Označavaju se tri uzorka: uzorak dvograma, uzorak trigrama i uzorak četverograma. Svaki uzorak sastoji se od 1000 n–grama. Kolokacije se razvrstavaju u grupe definirane u poglavlju 1.1. Cilj je napraviti uzorak za vrednovanje kombinacijom označenih uzoraka koji imaju zajednički kappa koeficijent veći ili jednak 0.6. Ako je podudaranje svih označenih skupova za određenu vrstu kolokacija manje od 0.6 zadatak označavanja tog tipa kolokacija smatrati će se preteškim te se neće raditi pokusi nad tom kolokacija.

Tablicom 6.1 prikazani su osnovni rezultati označavanja dvograma po vrstama kolokacija. Iz tablice je vidljivo da većina označenih kolokacija pripada grupi vlastitih imena. Najveće razlike u označavanju javljaju se kod grupe terminoloških izraza gdje je najmanje označio student D (samo 3), dok student E ima čak 81 označenu ustaljenu frazu.

Tablica 6.2 prikazuje kappa–koeficijente za liste označenih dvograma svih vrsta. Četiri studenta: A, B, D i F imaju međusobno poklapanje veće od 0.6 te će se te četiri liste kombinirati u skup za vrednovanje dvograma. Rezultat je lista od 694 n–grama od kojih su 84 n–grama $(12,25\%)$ označeni kao kolokacije.

Tablica 6.3 prikazuje rezultate kappa koeficijenta podudaranja za liste ustal-

Tablica 6.1: Rezultati označavanja skupa uzoraka za dvograme.

	Frazemi	Vl. imena	Ust. fraze	Term. izr.	Kolokacije	Učest. fraze
A	2	97	36	41	175	89
B	8	101	16	6	128	31
C	4	99	90	24	216	139
D	1	96	31	3	131	76
E	7	122	58	84	253	204
F	2	94	51	24	170	93

Tablica 6.2: Kappa koeficijent za dvogramske kolokacija.

$\kappa(x, y)$	*A*	*B*	*C*	*D*	*E*	*F*
A	–	**0.62**	0.53	**0.61**	0.52	**0.63**
B	0.62	–	0.56	**0.73**	0.50	**0.64**
C	0.53	0.56	–	0.55	0.54	0.58
D	0.61	**0.73**	0.55	–	0.50	**0.65**
E	0.52	0.50	0.54	0.50	–	0.59
F	0.63	**0.64**	0.58	**0.65**	0.59	–

Tablica 6.3: Kappa koeficijent za dvogramske ustaljene fraze.

$\kappa(x, y)$	*A*	*B*	*C*	*D*	*E*	*F*
A	–	0.10	0.22	0.19	0.20	0.23
B	0.10	–	0.10	0.09	0.04	0.22
C	0.22	0.10	–	0.19	0.29	0.23
D	0.19	0.09	0.19	–	0.12	0.23
E	0.20	0.04	0.29	0.12	–	0.29
F	0.23	0.22	0.23	0.23	0.29	–

jenih fraza. Niti jedan par uzoraka nema zadovoljavajuću vrijednost suglasnosti te se zadatak označavanja dvogramskih fraza smatra preteškim za neprofesionalnog označivača.

Tablicom 6.4 dani su rezultati za dvogramska vlastita imena. Ovaj zadatak pokazuje se kao najlakši te sve liste imaju iznimno viski koeficijent poklapanja.

Tablica 6.4: Kappa koeficijent za dvogramska vlastita imena.

$\kappa(x,y)$	A	B	C	D	E	F
A	–	0.89	0.85	0.86	0.80	0.89
B	0.89	–	0.87	0.90	0.81	0.85
C	0.85	0.87	–	0.89	0.80	0.87
D	0.86	0.90	0.89	–	0.81	0.93
E	0.80	0.81	0.80	0.81	–	0.82
F	0.89	0.85	0.87	0.93	0.82	–

Rezultat kombiniranja svih uzoraka je lista od 689 pojmova, od čega je 79 označenih vlastitih imena.

Tablica 6.5: Kappa koeficijent za dvogramske frazeme.

$\kappa(x,y)$	A	B	C	D	E	F
A	–	0.20	0.37	0.59	0.45	0.50
B	0.20	–	0.35	0.27	0.29	0.25
C	0.37	0.35	–	0.40	0.18	0.33
D	0.59	0.27	0.40	–	0.24	**0.67**
E	0.45	0.29	0.18	0.24	–	0.22
F	0.50	0.25	0.33	**0.67**	0.22	–

Kod frazema (tablica 6.5) javljaju se dva uzorka sa podudaranjem 0.67: uzorak D i uzorak F. Unatoč činjenici da zadovoljavaju taj uvjet, ako se bolje promotri tablica 6.1 vidljivo je da zbog premalenog broja frazema rezultat podudaranja nije pouzdan. Kao kombinacija ta dva uzorka dobila bi se lista sa samo jednim frazemom, što nije dovoljno za provođenje pokusa.

Tablica 6.6 pokazuje da je i označavanje terminoloških izraza preteška zadaća za studente. Niti jedan par ne zadovoljava uvjete za daljnje testiranje.

Rezultati označavanja trigrama dani su tablicom 6.7. Ponovno se pokazuje da su frazemi najrjeđi pojmovi, a da najviše problema ima s označavanjem terminoloških izraza. Kappa–koeficijenti za terminološke izraze i ustaljene fraze ne prelaze vrijednost 0,25 te se neće dalje razmatrati.

Kod trigrama koristila se verzija liste kolokacija i učestalih fraza te su rezultati dani tablicom 6.8. Razlog dodavanju učestalih fraza su loši rezultati suglasnosti

Tablica 6.6: Kappa koeficijent za bigramske terminološke izraze.

$\kappa(x,y)$	A	B	C	D	E	F
A	–	0.03	0.20	0.10	0.28	0.23
B	0.03	–	0.15	0.45	0.09	0.27
C	0.20	0.15	–	0.14	0.17	0.40
D	0.10	0.45	0.14	–	0.06	0.22
E	0.28	0.09	0.17	0.06	–	0.17
F	0.23	0.27	0.40	0.22	0.17	–

Tablica 6.7: Rezultati označavanja skupa uzoraka za trigrame.

	Frazemi	Vl. imena	Ust. fraze	Term. izr.	Kolokacije	Učest. fraze
A	0	17	19	15	51	96
B	6	26	8	0	39	104
C	4	36	101	16	155	224
D	0	19	14	0	33	69
E	6	36	8	69	118	208
F	4	19	54	4	81	77

Tablica 6.8: Kappa–koeficijent za trigramske kolokacije uz dodatak učestalih fraza.

$\kappa(x,y)$	A	B	C	D	E	F
A	–	0.38	0.28	0.31	0.32	0.35
B	0.38	–	0.31	0.41	0.40	0.35
C	0.28	0.31	–	0.26	**0.55**	0.32
D	0.31	0.41	0.26	–	0.31	0.47
E	0.32	0.40	**0.55**	0.31	–	0.35
F	0.35	0.35	0.32	0.47	0.35	–

za označene trigrame. Iako niti jedan par uzoraka ne zadovoljava uvjet za kappa vrijednost, za svrhe pokusa uzeti su uzorci C i E. Kombinacijom ta dva uzorka dobiva se lista od 792 pojma od čega je 239 označenih kolokacija.

Vlastita imena i kod trigrama se pokazuju kao najjednostavniji vrsta kolokacija za označavanje. U tablici 6.9 vidljivo je da je uvjet suglasnosti izražen kappa–

Tablica 6.9: Kappa–koeficijent za trigramska vlastita imena.

$\kappa(x,y)$	A	B	C	D	E	F
A	–	0.55	0.48	0.49	0.48	0.38
B	0.55	–	**0.77**	**0.66**	**0.73**	0.52
C	0.48	**0.77**	–	0.57	**0.68**	0.42
D	0.49	**0.66**	0.57	–	0.53	0.57
E	0.48	**0.73**	**0.68**	0.53	–	0.57
F	0.38	0.52	0.42	0.57	0.57	–

koeficijentom zadovoljen kod uzoraka B, C, D i E, te se ta četiri uzorka kombiniraju u jednu listu. Rezultat je list od 565 pojmova od čega je samo 13 vlastitih imena.

Tablica 6.10: Rezultati označavanja skupa uzoraka za četverograme.

	Frazemi	Vl. imena	Ust. fraze	Term. izr.	Kolokacije	Učest. fraze
A	1	18	9	13	43	167
B	3	26	8	1	39	153
C	2	47	67	26	130	332
D	0	38	4	0	42	94
E	2	54	12	130	195	247
F	2	22	51	6	81	74

Četverogrami prema tablici 6.10 imaju još veće razlike u broju označenih tipove nego trigrami. Ustaljene fraze i terminološki izrazi ponovno nemaju suglasnost veću od 0,2, dok frazema ima premalo za razmatranje. Najbolje kombinirana lista frazema sastojala bi se od dva pozitivno označena primjerka.

Kod vlastitih imena rezultati su dani tablicom 6.11. Vidljivo je da uvjet zadovoljavaju liste B, C i D. Njihovom kombinacijom dobivamo 445 pojmova, od čega je 21 označenih vlastitih imena.

U ovom poglavlju pokazano je da većina zadataka ručnog označavanja kolokacija zahtjeva stručno znanje. Koeficijent suglasnosti kod stručnih leksikografa prelazi vrijednosti od 0.7 za trigrame i četverograme (Petrović et al., 2008), dok je u slučaju ovih pokusa iznos koeficijenta oko 0.5 za jednostavnije zadatke koji ne uključuju svrstavanje kolokacije po grupama.

Tablica 6.11: Kappa koeficijent za četverogramska vlastita imena.

$\kappa(x,y)$	A	B	C	D	E	F
A	–	0.31	0.29	0.27	0.29	0.34
B	0.31	–	**0.64**	**0.64**	0.48	0.53
C	0.29	**0.64**	–	**0.73**	0.52	0.41
D	0.27	**0.64**	**0.73**	–	0.50	0.49
E	0.29	0.48	0.52	0.50	–	0.35
F	0.34	0.53	0.41	0.49	0.35	–

6.2. Rezultati ispitivanja postupka predobrade korpusa

Kod obrade korpusa cilj je bio opravdati postupak lematizacije i POS filtriranja kod morfološki složenih jezika poput hrvatskog. Ideja pokusa je da postupak lematizacije u slučaju takvih jezika nosi bolju informaciju od postupaka koji zanemaruju morfološke oblike jedne riječi. Postavlja se pitanje je li ta razlika dovoljno velika da bi lematizacija bila opravdana u programskom ostvarenju sustava za ekstrakciju kolokacija. U ovu svrhu korišten je Kendallov τ koeficijent. U sklopu pokusa koristile su se asocijacijske mjere PMI, Diceov koeficijent i chi-kvadrat, te frekvencija kao osnova za usporedbu. Učestalosti pojavljivanja su se određivale uz lematizaciju i POS filtriranje te bez njih. Rezultati pokusa prikazani su tablicom 6.12.

Tablica 6.12: Usporedba liste dobivene ekstrakcijom uz lematizaciju i POS–filtriranje s listom dobivenom bez lematizacije i bez POS–filtera.

	Frekvencija	PMI	Dice	χ^2
τ	-0.04	-0.03	-0.001	-0.001

Pokus je rađen samo na dvogramima, ali isti zaključci vrijede i za sve ostale n-grame. Zaključak je da postoji velika razlika u dobivenim listama. Prema Kendallovom τ–koeficijentu ne postoji nikakva veza između dvije liste. Ovaj rezultat upućuje na važnost lematizacije i POS–filtriranja kod morfološki bogatih jezika kao što je hrvatski.

6.3. Rezultati usporedbe asocijacijskih mjera

Usporedba asocijacijskih mjera rađena je sa uzorcima dobivenim u poglavlju 6.1. Kako bi se usporedila kvaliteta ekstrakcija kolokacija određenom asocijacijskom mjerom koristi se metoda srednje preciznosti (engl. *average precision*) opisana u poglavlju 5.1. Kao izvor podataka koristi se korpus "Glas Slavonije", a kao osnova za usporedbu koristi se sortiranje po učestalosti pojavljivanja u korpusu (frekvenciji).

6.3.1. Dvogrami

Kod dvograma se uspoređuju tri mjere: PMI, Diceov koeficijent i χ^2. Kod dvograma dobivena su dva uzorka: uzorak svih kolokacija i uzorak vlastitih imena. Rezultati su prikazani tablicom 6.13.

Tablica 6.13: Usporedba asocijacijskih mjera za dvograme.

	Frekvencija [%]	PMI [%]	Dice [%]	χ^2 [%]
AP(sve kolokacije)	18.3	**78.4**	46.0	18.7
AP(vlastita imena)	17.4	**78.2**	45.2	17.4

Kao daleko najbolja mjera pokazao se PMI. Mjera Dice daje slabije, ali još uvijek zadovoljavajuće rezultate. Kao najlošija mjera pokazao se χ^2 koji ne daje ništa bolje rezultate za dvograme od korištenja samih frekvencija.

6.3.2. Trigrami

Kod trigrama su, kao i kod dvograma, dobivena dva uzorka. Koriste se i učestale fraze kako bi se poboljšala kvaliteta uzorka za vrednovanje. Uspoređuje se sedam asocijacijskih mjera iz poglavlja 3. Kao osnova za proširenje funkcija uzorcima G koristi se dvogramska mjera PMI, budući da je to mjera koja je na dvogramima ostvarila najbolji rezultat. Rezultati ovog pokusa dani su tablicom 6.14.

Ponovno se mjera PMI pokazala kao najbolja mjera za ekstrakciju. Sve mjere pokazuju bolje rezultate od frekvencija. Mjere opisane uzorcima G_x daju sumjerljive rezultate s onima dobivenim Diceovim koeficijentom. Mjera G_3 ima drugi najbolji rezultat koji skoro dostiže srednju preciznost mjere PMI. Za vlastita imena najbolja od izvedenih mjera pokazala se mjera G_2.

Tablica 6.14: Usporedba asocijacijskih mjera za trigrame.

	Frekvencija [%]	PMI [%]	Dice [%]	G_1 [%]
AP(sve kolokacije)	34.6	**51.1**	43.4	45.4
AP(vlastita imena)	6.8	**39.3**	12.2	32.5
	G_2 [%]	G_3 [%]	G_4 [%]	G_5 [%]
AP(sve kolokacije)	45.4	49.7	35.9	45.4
AP(vlastita imena)	32.5	26.7	21.4	32.5

6.3.3. Četverogrami

Za četverograme uspoređuju se ponovno uzorci svih kolokacija i vlastitih imena. Mjere čije se performanse uspoređuju iste su kao i kod trigrama. Rezultati su prikazani tablicom 6.15.

Tablica 6.15: Usporedba asocijacijskih mjera za četverograme.

	Frekvencija [%]	PMI [%]	Dice [%]	G_1 [%]
AP(sve kolokacije)	17.1	9.1	14.1	2.9
AP(vlastita imena)	**29.3**	11.1	**19.8**	4.2
	G_2 [%]	G_3 [%]	G_4 [%]	G_5 [%]
AP(sve kolokacije)	4.0	**25.0**	3.6	7.1
AP(vlastita imena)	5.3	3.3	4.7	9.1

Kao daleko najbolja mjera za ekstrakciju četverograma pokazala se mjera G_3 predložena u (Petrović et al., 2008). Iznenađujuće, kod ekstrakcije vlastitih imena najbolja mjera pokazala se sama frekvencija n-grama. Niti jedna od izvedenih mjera nije kod ekstrakcije vlastitih imena pokazala srednju preciznost veću od 10%, a poslije frekvencije, kao najbolja mjera pokazao se Diceov koeficijent.

6.4. Rezultati evolucije asocijacijskih mjera

Cilj evolucije asocijacijske mjere genetskim programiranjem jest pronaći mjeru koja na što bolji način izdvaja kolokacije iz korpusa tekstova. U svrhu ocjenjivanja ovog postupka koriste se uzorci za sve dvogramske, trigramske i četverogramske kolokacije. Sam algoritam opisan je u poglavlju 4. Kao parametri algoritma ko-

rišteni su: veličina populacije 300, vjerojatnost križanja 0,7, vjerojatnost mutacije 0,001, uvjet zaustavljanja 5.000 iteracija bez promjene. Kao rezultati izvođenja dobivene su slijedeće mjere:

$$EV_2(ab) = \frac{P(ab)}{f(b)P(a)P(b)^2}, \tag{6.1}$$

$$EV_3(abc) = \begin{cases} P(ab)(f(b)f(ab) - P(abc) - f(c)^2), akoPOS(b) =' X' \\ P(ab)(f(b)f(ab) - P(abc) - f(c)f(abc)), inače \end{cases}, \tag{6.2}$$

$$EV_4(abcd) = \begin{cases} (\frac{P(bc)P(d)}{P(cd)} + f(d))\frac{P(ab)P(cd)^2}{P(d)^4} + f(c) + \frac{P(abcd)}{P(d)}, akoPOS(b) =' Ng' \\ (P(d) + f(d))\frac{P(ab)P(cd)^2}{P(d)^4} + f(c) + \frac{P(abcd)}{P(d)}, inače \end{cases}. \tag{6.3}$$

Rezultati i usporedbe sa postojećim mjerama dani su u tablici 6.16. Stupac poboljšanje F_1 prikazuje u postotcima koliko je maksimalni F_1 veći kod mjera dobivenih genetskim programiranjem te je dan formulom:

$$poboljsanje = \frac{F_{1novo} - F_{1staro}}{F_{1staro}}. \tag{6.4}$$

Za pokus se je koristio korpus "Glas Slavonije". Uzorak za vrednovanje isti je kao i kod prethodnih pokusa. Za najbolje mjere dobivene genetskim programiranjem koristi se oznaka EV_n.

Tablica 6.16: Usporedba rezultata evolucije novih asocijacijskih mjera genetskim programiranjem. Oznaka B_n koristi se za najbolju mjeru iz poglavlja 6.3.,

n-gram	$F_1(EV_n)$ [%]	B_n	$F_1(B_n)$ [%]	poboljšanje F_1 [%]
Dvogrami	95.8	PMI	79.4	20.7
Trigrami	82.1	PMI	57.4	40.3
Četverogrami	51.6	Dice	17.3	198.3

Mjere dobivene evolucijom pokazuju osjetno bolje rezultate od klasičnih mjera razmatranih u poglavljima 3. i 6.3.. Kod dvograma se dobiva mjera vrlo slična mjeri PMI. Mjera naglašava što manji broj pojavljivanja druge riječi u dvogramu. Mjere za trigrame i četverograme koriste se POS oznakama. Posebno je zanimljiva mjera za trigrame koja provjerava je li srednja riječ u trigramu zaustavna riječ. Upravo tu provjeru koristio su Petrović et al. (2008) za heurističke mjere, a sličan rezultat dobivali su i Šnajder et al. (2009). Problem ovog pristupa je izražena

ovisnost o uzorku za vrednovanje. Ako uzorak nije dobro označen dobiti će se mjera koja dobro radi na uzorku, ali loše radi sa kolokacijama općenito. Pošto se koristi isti uzorak za usporedbu mjera asocijacija i računanje funkcije dobrote u sklopu genetskog algoritma, dolazi do velikog nesrazmjera u rezultatima usporedbe danim tablicom 6.16. U svrhu pravilnog vrednovanja mjera asocijacije dobivenih genetskim programiranjem potrebno je napraviti još jedan uzorak za vrednovanje. Taj bi se uzorak koristio za usporedbu mjera asocijacije razmatranih u poglavlju 3. i onih dobivenih genetskim programiranjem. Pošto je postojeći uzorak premalen da bi se dijelio na dva dijela, ovakvo vrednovanje nije bilo moguće.

7. Zaključak

Cilj ovog rada bilo je usporediti postupke ekstrakcije kolokacija iz zbirki tekstova. Prikazan je postupak obrade korpusa i ideja proširenja leksičkih asocijacijskih mjera za kolokacije koje se sastoje od više riječi. Dana je definicija kolokacija kao višerječne imenske skupine. Uvedena je i podjela kolokacija na četiri vrste: vlastita imena, frazemi, ustaljene fraze i terminološki izrazi. Opisano je i stvaranje uzorka za vrednovanje postupaka ekstrakcije kolokacija.

U sklopu rada uspoređeno je osam asocijacijskih mjera: PMI, Diceov koeficijent, χ^2, G_1, G_2, G_3, G_4 i G_5. Ispitivala se ekstrakcija različitih tipova kolokacija. Rezultati su pokazali da je najbolja mjera za ekstrakciju dvogramskih i trigramskih kolokacija PMI, dok se kod četverogramskih kolokacija kao najbolja općenito pokazala mjera G_3, a najbolja za vlastita imena (nakon frekvencije) Diceov koeficijent.

Genetskim programiranjem pokušalo se pronaći optimalnu asocijacijsku mjeru. Mjere dobivene genetskim programiranjem imale su veću vrijednost F_1 od klasičnih mjera. Kod dvograma se kao najbolja pokazala mjera slična PMI, a kod trigrama se dobila mjera osjetljiva na isti POS–uzorak kao i onaj korišten kod istraživanja Petrović et al. (2008) i Šnajder et al. (2009). Drugi cilj pristupa temeljenog na genetskom programiranju bio je da se izvedu POS uzorci koji bi predstavljali podjelu kolokacija u hrvatskom jeziku na morfo-sintaktičkoj razini. Ova pretpostavka pokazala se pogrešnom pošto se dobivaju rješenja slična heurističkim uzorcima koji se zasnivaju samo na zaustavnim riječima.

U radu je pokazano i da je postupak lematizacije nužan kod obrade korpusa pisanih morfološki složenim jezicima poput hrvatskog jezika. Istraživanje je pokazalo da su liste dobivene korištenjem lematizacije i POS–filtriranja i bez njih bitno različite te je, uz pretpostavku da lematizacijom dobivamo bolji uvid u učestalosti pojedinih riječ, taj postupak opravdan unatoč povećanju složenosti.

Na posljetku, kao najveći problem pokazalo se stvaranje uzoraka za vrednovanje. Označavanje kolokacija pokazalo se kao zadaća koja zahtjeva veću količinu

stručnog znanja. Također, zbog prevelike razlike u označenim uzorcima, bilo je ne moguće usporediti asocijacijske mjere za ekstrakciju terminoloških izraza, ustaljenih fraza i frazema. Problem s uzorkom bitan je i za evoluciju novih mjera genetskim programiranjem. Ovaj postupak vrlo ovisi o uzorku koji se koristi te je moguće da se uz loši uzorak dobiju mjere koje pokazuju visoki F_1 samo na razini uzorka, a ne općenito.

Prije bilokakvog daljnjeg istraživanja postupaka ekstrakcije kolokacija trebalo bi ispitivanja ponoviti s boljim uzorkom. Isto tako, potrebno je usporediti asocijacijske mjere na korpusima iz drugih domena kao što su znanstveni članci ili pravni tekstovi.

Literatura

Wirote Aroonmanakun. Collocation extract. `http://pioneer.chula.ac.th/~awirote/colloc/`, 2000.

Michael Barlow. *Collocate 1.0: Locating collocations and terminology*. TX: Athelstan, 2004.

Morton Benson. Collocations and general-purpose dictionaries. *International Journal of Lexicography*, 3(1):23–35, 1990.

Chamblon Systems, Inc. Terminology extractor. `http://www.chamblon.com/terminologyextractor.htm`, 2004.

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

Vidas Daudaravicius. Automatic identification of lexical units. *Informatica*, 2009.

Davor Delač, Zoran Krleža, Jan Šnajder, Bojana Dalbelo Bašić, and Frane Šarić. *TermeX*: A tool for collocation extraction. In *Proceedings of the 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, pages 149–157, 2009.

Marin Golub. *Genetski algoritam*. FER, Sveučilište u Zagrebu, 2004. Nastavni materijali.

Michael Gordon, Weiguo Fan, and Praveen Pathak. Adaptive web search: Evolving a program that finds information. *IEEE Intelligent Systems*, 21(5):72 – 77, 2006.

Annette M. Green. Kappa statistics for multiple raters using categorical classifications. In *The Twenty-Second Annual SAS Users Group International Conference (online)*, 1997.

David Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 329–338, 1993.

Maurice Kendall. *Rank Correlation Methods*. Hafner Publishing Co., 1955.

John R. Koza. *Genetic programming: On the programming of computers by means of natural selection*. MIT Press, 1992.

Brigitte Krenn, Stefan Evert, and Heike Zinsmeister. Determining intercoder agreement for a collocation identification task. In *KONVENS*, 2004.

C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

Rita McCardell Doerr. A lexical semantic and statistical approach to lexical collocation extraction for natural language generation. *AI Magazine*, 16:105, 1995.

Rada Mihalcea and Ehsanul Faruque. Senselearner: Minimally supervised word sense disambiguation for all words in open text. In *Proceedings of ACL/SIGLEX Senseval*, 2004.

Brigitte Orilac i Mike Dillinger. Collocation extraction for machine translation. In *Machine Translation Summit IX*, pages 292–298, 2003.

Saša Petrović, Jan Šnajder, Bojana Dalbelo Bašić, and Mladen Kolar. Comparison of collocation extraction measures for document indexing. In *Information Technology Interfaces (ITI 2006)*, 2006.

Saša Petrović, Jan Šnajder, and Bojana Dalbelo Bašić. Extending lexical association measures for collocation extraction. Computer, Speech and Language, 2009.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, 2002.

Ivan Sikirić. *Primjena evolucijskog programiranja na nalaženje optimalnih mjera za ekstrakciju kolokacija iz teksta*. diplomski rad, Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu, 2007.

Frank Smadja. From n-grams to collocations: An evaluation of xtract. In *Proceedings of 29th Annual Meeting of the Association for Computational Linguistics*, pages 279–284, 1991.

Frank Smadja. Retrieving collocations from text: Xtract. In *Proceedings of 31th Annual Meeting of the Association for Computational Linguistics*, volume 19, pages 143–177, 1993.

Artur Šilić, Frane Šarić, Bojana Dalbelo Bašić, and Jan Šnajder. Tmt: Object-oriented text classification library. In *29th International Conference on INFORMATION TECHNOLOGY INTERFACES*, 2007.

Jan Šnajder, Bojana Dalbelo Bašić, and Marko Tadić. Automatic acquisition of inflectional lexica for morphological normalisation. *Information Processing and Management*, 44(5):1720–1731, 2009.

D. Yarowsky. One sense per collocation. In *Proceedings of ARPA Human Language Technology Workshop*, 1993.

Dodatak A
Asocijacijske mjere

Mjere asocijacije PMI, Diceov koeficijent, χ^2 imaju temeljna proširenja te su primjenjiva na bilokoje n-grame, gdje je $n \in (N)$.

$$PMI_n(w_1 \cdots w_n) = \log_2 \frac{P(w_1 \cdots w_n)}{\prod_{i=1}^{n} P(w_i)}, \tag{A1}$$

$$DICE_n(w_1 \cdots w_n) = \frac{nf(w_1 \cdots w_n)}{\sum_{i=1}^{n} f(w_i)}. \tag{A2}$$

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{A3}$$

Slijedeće mjere asocijacije dobivene su korištenjem uzoraka za proširivanje te imaju smisla samo ako se koriste na n-gramima gdje je $n > 2$.

$$G : \mathcal{F} \times \mathbb{N} \times W^+ \to \mathbb{R}. \tag{A4}$$

$$G_1(g, w_1 \cdots w_n) = \frac{g(w_1, w_2 \cdots w_n) + g(w_1 \cdots w_{n-1}, w_n)}{2}, \tag{A5}$$

$$G_2(g, w_1 \cdots w_n) = \frac{g(w_1 \cdots w_{\lfloor n/2 \rfloor}, w_{\lceil n/2 \rceil} \cdots w_n) + g(w_1 \cdots w_{\lfloor n/2+1 \rfloor}, w_{\lceil n/2+1 \rceil} \cdots w_n)}{2}, \tag{A6}$$

$$G_3(g, w_1 \cdots w_n) = \frac{1}{n-1} \sum_{i=1}^{n-1} g(w_i, w_{i+1}). \tag{A7}$$

$$G_4(g, w_1 \cdots w_n) = g(w_1 \cdots w_{n-1}, w_2 \cdots w_n), \tag{A8}$$

$$G_5(g, w_1 \cdots w_n) = \frac{1}{n-1} \sum_{i=1}^{n-1} g(w_1 \cdots w_i, w_{i+1} \cdots w_n), \tag{A9}$$

Dodatak B
Primjeri n-grama dobivenih ekstrakcijom kolokacija

U ovom poglavlju dani su primjeri lista n-grama dobivenih postupkom ekstrakcije kolokacija opisanom u ovom radu. Dana su po dva primjerka, rezultati sa frekvencijom kao mjerom asocijacije te rezultati sa najboljom mjerom asocijacije iz poglavlja 6.3., za dvograme, trigrame i četverograme. Svaka tablica ima po tri stupca. Prvi stupac je oznaka je li dani n-gram ručno označen kao kolokacija ("+"), drugi stupac je sam n-gram, a zadnji stupac je vrijednost mjere asocijacije za taj n-gram. Tablicama je prikazano po 20 najbolje rangiranih n-grama.

Tablica B1: Rezultati ekstrakcije dvograma sa frekvencijom kao mjerom asocijacije

kolokacija	n-gram	frekvencija
	premijerom Račanom	280
	posljednje utakmice	268
+	Antuna Novalića	173
	pripremnog razdoblja	171
	papa Ivan	151
	novi ministar	130
	najmanje pet	124
	cijene kruha	121
+	Andru Vlahušića	116
	kvalitete vode	114
+	Ladislav TOMIČIĆ	111
	Dan općine	95
	šest kuna	82
+	Damir Macanić	67
+	Božo Kovačević	66
+	Andy Roddick	62
	velika prigoda	60
+	Damir Vuica	58
	susjedne zgrade	55
	druga nagrada	55

Tablica B2: Rezultati ekstrakcije dvograma sa PMI kao mjerom asocijacije

kolokacija	n-gram	PMI
+	Egidio Ćepulić	14,60
+	Matthias Platzek	13,79
+	Aslana Mashadova	13,66
+	Rohan Gunaratna	13,42
+	Olena Popik	13,36
+	Jeremy Greenstock	13,25
	uvale Žrnovnica	12,77
+	Luciano Delbianco	12,55
+	Lada Niva	12,53
+	Bartola Kašića	12,49
+	Areta Ćurković	11,94
+	Faks Helizim	11,68
+	James Tomkins	11,48
+	Ljiljanka Grabić	11,35
+	Milojko Tankosić	11,34
+	Augusta Harambašića	11,28
+	Andy Roddick	11,18
+	Andru Vlahušića	11,07
	Njemica Hilde	10,99
	Austrijanac Rainer	10,69

Tablica B3: Rezultati ekstrakcije trigrama sa frekvencijom kao mjerom asocijacije

kolokacija	n-gram	frekvencija
+	Međunarodni kazneni sud	259
+	osječke Kliničke bolnice	222
	predsjedniku Republike Stjepanu	164
	predsjednik Republike Stjepan	164
	rada i poduzetništva	114
	početka sljedeće godine	113
+	visoke stručne spreme	110
	tužiteljice Carle del	108
	ožujka prošle godine	102
	centra za prevenciju	91
	poljoprivrede Petru Čobankoviću	86
+	Hrvatske biskupske konferencije	83
+	izbornik Otto Barić	83
+	naslov svjetskog prvaka	82
	Vlade Jadranka Kosor	78
+	župan Ivan Begović	71
	struje i vode	69
	sabora Zlatkom Tomčićem	69
	kuna po osobi	66
	put ove sezone	64

Tablica B4: Rezultati ekstrakcije trigrama sa PMI kao mjerom asocijacije

kolokacija	n-gram	PMI
+	Susilo Bambang Yudhoyono	29,47
	You need ground	25,89
+	Stolnotenisačica Tamara Boroš	20,64
+	kardinal Joseph Ratzinger	20,37
+	Damir Zlatar Frey	19,69
+	GP Partners Baranja	19,67
	Ivičin trener Vincencij	19,51
	obrane Donald Rumsfeld	19,19
	Petračev sin Novica	18,74
	tužiteljice Carle del	18,22
+	David Junior Lopes	18,05
	poslova Dimitrij Rupel	17,50
+	premijer Silvio Berlusconi	17,17
+	akademika Ivana Supeka	16,80
+	izbornik Otto Barić	16,67
	Panturista Stjepan Frigan	16,14
	poljoprivrede Petru Čobankoviću	15,81
+	gradonačelnik Radoslav Jurić	15,21
+	dogradonačelnik Petar Mlinarić	15,18
+	pobjednici Davis Cupa	14,92

Tablica B5: Rezultati ekstrakcije četverograma sa frekvencijom kao mjerom asocijacije

kolokacija	n-gram	frekvencija
+	Zavod za javno zdravstvo	684
	branitelja i međugeneracijske solidarnosti	211
	socijalnu skrb i zdravstvo	98
	Hrvatskog sabora Vladimiru Šeksu	69
	Udruga za šport djece	56
+	Vijeću sigurnosti Ujedinjenih naroda	48
	vanjskih poslova Miomirom Žužulom	47
	agencije za atomsku energiju	46
	Upravni odbor Hrvatskog fonda	46
	satno zadržavanje u pritvoru	39
	izjednačavanje krivnje i pretvaranje	38
	temelju ugovora o radu	38
	pojavljivanje dokumenata u listu	34
	događanja s objavljivanjem transkripata	34
	odnosu na prošlu sezonu	33
	sklopu Osječkoga ljeta kulture	31
	vanjskih poslova Jack Straw	29
+	Centar za profesionalnu rehabilitaciju	29
	područja grada Belog Manastira	28
	Rukometni klub Osijek Elektromodul	27

Tablica B6: Rezultati ekstrakcije četverograma sa G_3 kao mjerom asocijacije

kolokacija	n-gram	G_3
	Općine Lovas Željko Cirba	6,53
+	Zavod za javno zdravstvo	6,32
	uprave Holdinga Dragan Marčinko	6,23
+	Interov trener Srećko Bogdan	6,12
	ministrice pravosuđa Boris Koketi	6,02
+	Klub dizača utega Osijek	6,01
	Končar elektroindustrija rastom cijene	5,65
	izboru članova predstavničkih tijela	5,64
	Pivovare Osijek Dario Fančović	5,08
	sustava unaprjeđenja kvalitete obrazovanja	5,02
	ministarstva kulture Jadran Antolović	4,97
+	Filozofski fakultet u Osijeku	4.85
	Načelnik Općine Ernestinovo Matija	4,78
	šećerana Kandit Premijer d	4,61
	sukob interesa Antun Kapraljević	4,19
	generala Gotovine Luka Mišetić	3,86
	liga RUKOMET Dvorana Jug	3,65
	dio Parka prirode Papuk	3,63
	područja grada Belog Manastira	3,59
	radnicima PPK Valpovo d	3,56

Dodatak C
Sažetak i ključne riječi

Sažetak:

Kolokacije, kombinacije riječi koje se skupa pojavljuju češće nego slučajno, imaju velik broj primjena u obradi prirodnog jezika. U literaturi se pojavljuje mnogo pristupa automatskoj ekstrakciji kolokacija zasnovanih na mjerama asocijacija. U ovom radu vrednuje se postupak ekstrakcije kolokacija korištenjem različitih mjera asocijacije. Uspoređeno je mnogo mjera te su genetskim programiranjem izvedene nove mjere za ekstrakciju kolokacija. U ovom radu su opisani postupci lematizacije i POS filtriranja te dana usporedba procesa ekstrakcije kolokacija sa i bez ta dva koraka.

Ključne riječi:

ekstrakcija kolokacija, mjere asocijacije, lematizacija, POS filter

Abstract:

Collocations – word combinations occurring together more often than by chance – have a wide range of NLP applications. Many approaches for automating collocation extraction based on lexical association measures have been proposed in the literature. In this thesis we evaluate the use of lexical measures of association in collocation extraction. A wide range of different measures are compared. New association measures are evolved using genetic programming. This thesis also addresses the use of lemmatization and POS–filtering in collocation extraction with application.

Keywords:

collocation extraction, lexical association measures, lemmatization, POS filtering