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. 144

POSTUPAK EKSTRAKCIJE SLOŽENIH KRATICA HRVATSKOGA JEZIKA

Fran Dragomanović

Zagreb, siječanj 2011.

Sadržaj

Popis slika:

Popis tablica:

1. Uvod

Svakim danom u jezik se uvodi sve veći broja složenih kratica[1] (ili akronima) u različitim područjima. Mnogo organizacija na Internetu posjeduje vrlo velik broj dokumenata koji sadrže mnogo kratica. U mnogo se slučajeva kratice javljaju toliko često da ljudi drugih struka imaju problema s razumjevanjem teksta. Zbog toga je potrebno bilježiti kratice i njihova proširenja te na taj način stvarati rječnik kratica pomoću kojeg će biti jednostavnije razumijevanje i obrada dokumenata. Mnogo ručno sakupljenih kratica je dostupno na Internetu. Međutim, mnoge tako skupljene kratice ograničene su na određena tematska područja. Uz vrlo veliku brzinu rasta broja kratica, ručno održavanje takvog rječnika prilično je težak posao. Stoga bi svakako bilo od velike koristi razviti sustav za automatsko prepoznavanje kratica i njihovih proširenica u tekstu.

Automatsko traženje svih kratica i njihovih proširenica problem je obrade teksta koji se do sada rješavao primjenom *ad hoc* heuristika. Svaki prijašnji pristup rješavanju tog problema pripada jednoj od sljedeće tri kategorije: (1) pristupi temeljeni na prirodnom jeziku (engl. *natural language-based approaches*), (2) pristupi temeljeni na pravilima (engl. *rule-based approaches*) i (3) pristupi temeljeni na višestrukom poravnavanju (engl. *multiple alignment-based approaches*).

U ovom radu predlaže se pristup strojnog učenja za ekstrahiranje kratica iz tekstova na hrvatskome jeziku.

Sustav se može podijeliti na tri osnovna koraka:

1. Heuristikom se izdvajaju svi potencijalni kandidati kratica,
2. Iz okoline svakog kandidata generiraju se kandidati proširenica,
3. Koristi se stroj s potpornim vektorima (engl. *support vector machine*, SVM) za odabir ispravne proširenice zadane kratice.

U usporedbi s konvencionalnim pristupom temeljenim na uzorcima, predložen pristup strojnog učenja ima nekoliko prednosti:

[1] U nastavku ćemo radi jednostavnosti "složene kratice" oslovljavati jednostavno "kraticama"

1. Ljudski faktor pisanja i reguliranja uzoraka, odnosno pravila nije potreban;
2. Za odabir ispravne proširenice moguće je iskoristiti više dokaza nego što je moguće kada se koriste uzorci. Kod uzoraka je moguće koristiti jedino ”čvrste dokaze”. Nasuprot tome, strojno učenje zajedno koristi i čvrste i slabe dokaze;
3. Lakše je upravljati adaptacijom domene. Uporabom odgovarajućih značajki, model naučen nad jednom domenom može ostvariti vrlo dobre rezultate i nad drugim domenama.

Ostala su poglavlja organizirana na sljedeći način. U drugom se poglavlju razmatraju srodni radovi. U trećem poglavlju opisan je postupak ekstrakcije kratica, a u četvrtom implementacija programa. Naposljetku se tumače rezultati analize i programa. Rad završava zaključkom i pregledom korištene literature.

2. Srodni radovi

U ovom se odjeljku ukratko opisuju dosadašnji radovi na području automatske ekstrakcije kratica iz teksta. Kao što je u uvodu navedeno, pristupi automatskoj ekstrakciji kratica mogu se svrstati u jednu od tri kategorije: (1) pristupi temeljeni na primjeni prirodnog jezika, (2) pristupi temeljeni na pravilima i (3) pristupi temeljeni na višestrukom prilagođavanju.

Pristupima temeljenim na primjeni prirodnog jezika pokušavaju se iskoristiti rezultati obrade prirodnog jezika kao što je označavanje vrste riječi (engl. *part-of-speech tagging*) pri pronalaženju proširenica koje se u tekstu pojavljuju u blizini kratica. Primjer takvog sustava za traženje parova kratica i proširenica iz medicinskih dokumenata je AcroMed [1]. Međutim, navedeni pristup vrlo je teško ostvariti u praksi zbog složenosti algoritma podudaranja uzoraka.

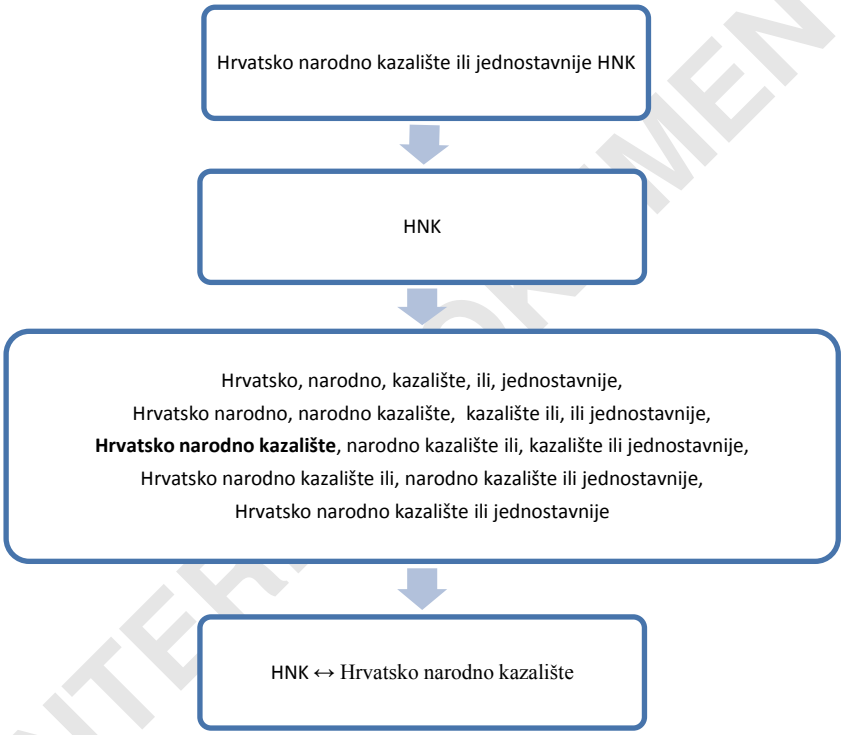
Primjer pristupa temeljenog na pravilima jest Acrophile [2]. Njegovi su autori detaljno proučavali web-dokumente i stvorili velik broj uzoraka i pravila za identifikaciju kratica i njihovih proširenica iz teksta. Nažalost, zbog složenosti veza između kratica i njihovih proširenica u web-dokumentima, preciznost Acrophilea je niska. Osim njega, postoji još radova s pristupima temeljenih na pravilima [3, 4].

Kao primjer pristupa temeljenih na višestrukom prilagođavanju, program pronalaženja kratica (engl. *Acronym Finding Program*, AFP) upravlja ekstrakcijom u okruženju OCR-a [5]. AFP se temenji na nepreciznom algoritmu podudaranja uzoraka koji se primjenjuje na tekst koji okružuje mogući kraticu. Među ostalim radovima koji pripadaju pristupima temeljenim na višestrukom prilagođavanju su [6, 7].

2.1.Korištenje SVM-a u ekstrakciji kratica

Pristup opisan u ovom radu najsličniji je [8]. Taj se pristup sastoji od tri osnovna koraka:

1. Identifikacija mogućih kratica,
2. Generiranje kandidata ekspanzija,
3. Selekcija pravih ekspanzija.



Slika 1. Dijagram ekstrakcije kratica i ekspanzija

Slika 1. pokazuje obradu primjera "*Hrvatsko narodno kazalište ili jednostavnije HNK*", koji je ujedno u ulazni tekst u postupak ekstrakcije složenih kratica. Prolaskom kroz tekst ekstrahiraju se sve moguće kratice. Kao moguća kratica odabire se "*HNK*" te se potom obrađuje okruženje oko kratice i izvlače kandidati ekspanzija za navedenu kraticu. U konkretnom slučaju to su nizovi riječi: "*Hrvatsko, narodno, kazalište"*,…, *"Hrvatsko narodno kazalište"*,…, *"Hrvatsko narodno kazalište ili jednostavnije*". Na kraju, od navedenih se nizova riječi odabire jedan koji se uzima kao ispravnu ekspanziju za kraticu. Detaljniji opis navedenih koraka slijedi u nastavku.

2.1.1. Identifikacija mogućih kratica

Cilj ovog koraka jest identificirati sve moguće kratice iz originalnog teksta. To se odnosi na prvi korak (*moguće kratice*) sa slike 1, gdje se identificira kratica "*HNK*" iz teksta "*Hrvatsko narodno kazalište ili jednostavnije HNK*". Ako pojavnica zadovoljava sljedeće zahtjeve, možemo je smatrati mogućom kraticom:

1. Pojavnica je duljine između dva i deset znakova i sadrži najviše jedan razmak;
2. Prvi je znak slovo ili broj, pri čemu je barem jedno slovo veliko;
3. Pojavnica nije poznata riječ u rječniku. Nije ime osobe, mjesta ili riječ iz predefinirane liste riječi zaustavljanja (funkcijske riječi).

Prvi se uvjet odnosi na veličinu kratice. Drug i treći sprječavaju ulazak imena, mjesta i čestih riječi u listu mogućih kratica. Na temelju navedenih ograničenja formira se lista mogućih kratica. Kako su ograničenja prilično generalizirana za mnogo mogućih kratica u listi nije moguće pronaći ekspanziju. Pogrešno identificirane kratice neće utjecati na završne rezultate jer se oni automatski likvidiraju u sljedećim koracima.

2.1.2. Generiranje kandidata ekspanzija

U ovom se koraku generiraju svi kandidati ekspanzija za prethodno identificirane kratice. Kao u primjeru na slici 1, u trećem bloku se nalazi skup svih kandidata ekspanzija. Možemo primijetiti kako se ekspanzije uvijek pojavljuju u tekstu oko kratica. Prema tome, iz okoline koja je u istoj rečenici generiraju se kandidati ekspanzija za kraticu. Jedan od navedenih kandidata je moguća ispravna ekspanzija za zadanu kraticu. Prije generiranja kandidata ekspanzija, rečenica se segmentira razmacima, pri čemu se znakovi poput "*,*", "*.*", "*)*", "*(*", "*!*" i njima slični smatraju zasebnim pojavnicama.

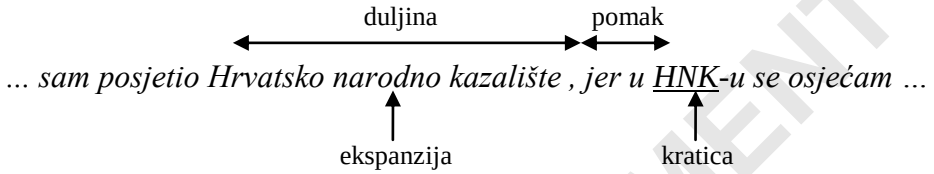
Kratica dijeli rečenicu na dva dijela:

1. Niz koji prethodi kratici (lijevi kontekst),
2. Niz koji slijedi nakon kratice (desni kontekst).

Svi nizovi u lijevom i desnom kontekstu smatraju se kandidatima. Koriste se dva parametra pri identifikaciji kandidata:

1. *duljina* – broj pojavnica kandidata,
2. *pomak* – broj pojavnica između kandidata i kratica.

Na slici 2. prikazan je primjer kratice (HNK) i njegove ekspanzije (*Hrvatsko narodno kazalište*) pri čemu je *duljina*=3 i *pomak*=3. Bitno je primijetiti da se znak "," gleda kao zasebna riječ.



Slika 2 - Generiranje kandidata ekspanzija

Teoretski, svi nizovi unutar rečenice mogli bi se označiti kao kandidati ekspanzije. Međutim, to nije potrebno. Možemo primijetiti da se prave ekspanzije javljaju uvijek blizu kratica. Duljina ispravne ekspanzije (izražena u broju pojavnica) je uvijek približno jednaka duljini odgovarajuće kratice (izražene u znakovima):

1. *maxpomak* – maksimalna udaljenost između kandidata i kratice. U ovom je radu ta udaljenost postavljena na deset znakova. Uvidom u podatke zaključeno je da je to dovoljno velik pomak za generiranje svih ispravnih ekspanzija.
2. *maxduljina* – maksimalna duljina ekspanzija. Za duge kratice, duljina ekspanzija (u pojavnicama) kraća je od duljine kratice (u znakovima) uvećane za 5. Za kratke kratice duljina ekspanzije je kraća od dvostruke duljine kratice. Računanje maksimalne duljine možemo dakle definirati kao:

$$maxduljina = min\,(duljina\,(kratica) + 5,\ duljina\,(kratica) \times 2), \quad (1)$$

koja je preuzeta iz [8].

2.1.3. Selekcija ispravnih ekspanzija

U ovom se koraku koristi SVM-model kako bi se odabrale ispravne ekspanzije za kratice iz skupa kandidata. Zadnji korak primjera na slici 2. je odabir kandidata "*Hrvatsko narodno kazalište*" pomoću značajki kao pravu ekspanziju za kraticu.

2.1.3.1. SVM-model

Za rješavanje ovog problema upotrijebljen je statistički klasifikator. Skup kandidata kratica-ekspanzija je unaprijed ručno označen. On se koristi za učenje SVM-modela i za testiranje performansi sustava za ekstrakciju. Formalnije se problem može opisati na sljedeći način: Za skup za učenje $D = \{x_i, y_i\}_{i=1}^{n}$, konstruiramo model koji minimizira pogrešku u predikciji *y* za neki *x*. Pritom je $x_i \in X$ kandidat sastavljen od para kratica-ekspanzija, a $y_i \in \{+1, -1\}$ oznaku koja indicira je li kandidat ispravna ekspanzija za danu kraticu. Kad se primijeni na novog kandidata *x*, model predviđa odgovarajući *y* i vraća rezultat predikcije. Za instancu *x*, SVM dodijeljuje rezultat *f(x)* prema

$$f(x) = w^T x + b, \tag{2}$$

gdje je *w* vektor težina, a *b* posmak. Ako je vrijednost *f(x)* pozitivna, *x* se klasificira u pozitivnu kategoriju, a inače u negativnu. Za konstrukciju SVM-a potreban je unaprijed označen skup za učenje. Detalji o algoritmu nalaze se u [9]. Ukratko, u (2) algoritam učenja stvara hiperravninu takvu da ona odjeljuje pozitivne od negativnih primjera u skupu za učenje, i to tako da granični pojas između tih dvaju skupova bude najveći mogući.

U nekim je slučajevima moguće da se za istu kraticu dodijeli više kandidata ekspanzija. Tada se kao prava ekspanzija odabire onaj kandidat koji ima najveći rezultat predikcije. S druge strane, postoje slučajevi u kojima se niti jedan kandidat ne identificira kao ekspanzija za danu kraticu. U tom slučaju kratica se automatski odbacuje.

2.1.3.2. Značajke

U SVM-modelu koristimo značajke binarnih i realnih vrijednosti opisane u nastavku. One se stvaraju kako bi okarakterizirale moguće kratice, kandidate ekspanzija te kontekst u kojem se javljaju.

- Značajke koje karakteriziraju kratice i ekspanzije:

1. Duljina kratice, mala slova, brojevi, specijalni znakovi i razmaci u kratici su važne značajke pri odabiru dobre kratice;
2. Duljina ekspanzije, riječi s početnim velikim slovom u ekspanziji, prva/zadnja riječ ekspanzije i prijedlozi/veznici dobri su indikatori za utvrđivanje radi li se o dobroj ekspanziji;
3. Veza između kratice i ekspanzije smatra se bitnom značajkom. Npr. ako se slova kratice podudaraju s prvim slovima riječi u ekspanziji, vrlo je vjerojatno da je kandidat prava ekspanzija za kraticu.

- Značajke koje karakteriziraju kontekst:

Također se oslanjamo na značajke konteksta. Npr. većina se kratica nalazi unutar zagrade i pojavljuju se neposredno nakon ekspanzija. Ako su se kratice već pojavile u prethodnom tekstu, mala je vjerojatnost da je kandidat prava ekspanzija. Stoga su značajke:

1. Nalazi li se kratica unutar zagrade,
2. Pojavljuje li se kratica neposredno nakon ekspanzije,
3. Nalazi li se ekspanzija u lijevom ili desnom kontekstu kratice.

2.2.Prednosti pristupa temeljenog na SVM-u

Ekstrakcija pomoću uzoraka, temeljenih na heurističkim pravilima, izravan je i očit pristup ekstrakciji kratice. Međutim, stvaranje i podešavanje uzoraka vrlo je naporno i dugotrajno. Također, ručno pisanje pravila ograničava korištenje informacija. Samo se jaki dokazi mogu smatrati čvrstim pravilima. Pristup temeljen na strojnom učenju može nadvladati ove poteškoće na prirodan način.

Model strojnog učenja gradi se na temelju zbirke označenih primjera. Označavanje podataka mnogo je lakše i jeftinije od pisanja uzoraka odnosno pravila.

Također, modeli strojnog učenja lako mogu iskoristiti različite vrste dokaza. U strojnom učenju, kratice, ekspanzije i kontekst u kojem se javljaju opisuju se značajkama, od kojih neke predstavljaju jake dokaze. Npr. značajka se definira na sljedeći način: "Tvore li prva slova riječi ekspanzije kraticu?". Tako definirana značajka može se koristiti u metodama temeljenim na pravilima i u pristupima temeljenim na strojnom učenju. Neke druge značajke predstavljaju slabe dokaze. Npr. definira se značajka: "Pojavljuje li se kratica u prethodnom tekstu?". Budući da se kratice često definiraju pri njihovom prvom pojavljivanju u tekstu, značajka nagovještava da kandidat možda nije prava ekspanzija. Međutim, protuprimjeri postoje u stvarnom svijetu. Npr., "*SABA RH je osnovan u Republici Hrvatskoj (RH)*". Takva vrsta značajki može se koristiti samo u pristupima temeljenim na strojnom učenju. To je razlog boljeg rezultata pristupa ekstrakciji kratica pomoću SVM-a.

3. Opis postupka i implementacije

Rješavanju problema pristupilo se dvjema različitim metodama. Prva je referentna metoda koja se temelji na uzorcima, dok se druga, koja je oblikovana po uzoru na [8], temelji na SVM-u. U radu su odabrana dva pristupa kako bismo mogli jasno prikazati usporedbu različitih metoda.

3.1.Referentna metoda

Referentna metoda (eng. *baseline method*) gradi se na temelju uzoraka koji su utvrđeni prethodnom detaljnom analizom. Većina kratica tvori se na temelju početnih slova njegove ekspanzije. Međutim, u korpusu Vjesnika (1999-2009.), nad kojim je izvršena analiza i kasnije testiranje, javlja se mnogo *stranih* kratica. To su kratice koji čine početna slova ekspanzija na stranom jeziku, no umjesto originalne ekspanzije, u tekstu je naveden hrvatski prijevod ekspanzije. Tako da se u rijetko kojem slučaju početna slova originalne ekspanzije i hrvatskog prijevoda podudaraju, što znači da nije moguće točno odrediti ekspanziju za moguću kraticu samo na temelju početnih slova. U tom bi slučaju točnost bila premala. Potrebno je uvesti još neka pravila ili uzorke. Još jedno pravilo koje radi bitniju razliku je pregled nalazi li se kratica unutar zagrada. To je najčešća forma u kojoj se javlja par kratica/ekspanzija. No, ni s tom se formom ne može sa sigurnošću tvrditi da se ekspanzija javlja upravo ispred zagrada. Velik broj slučajeva mogućih kratica koji se nalaze unutar zagrada su kratice političkih stranaka kod kojih se imenima političara pojavljuju ispred zagrade ("*... priopćila je Jadranka Kosor (HDZ)...*") ili kratice država s imenima sportaša ispred zagrade ("*...3. Irvine (VB) +1.796,...*"), jer se mnogo teksta posvećuje politici, odnosno sportskim rezultatima. Zbog navedenih razloga uvode se još neka pravila koja će povećati preciznost referentne metode.

Pseudokodom jednostavnije možemo prikazati tok odabira ekspanzije za odabrana kratica:

```
dok (RečenicaUDatoteciNijePrazan)
{    čitaj rečenica
     ako rečenica sadrži mogući akronim
          TražiProširenicu(rečenica)
}

TražiProširenicu(rečenica)
{    TražiPočetakMogućeProširenice(akronim)
     ako (AkronimUnutarZagrada & !ProširenicaIme)
     {    provjeraLijevo = ProvjeraLijevoOdPočetka(lijevo)
          provjeraDesno = ProvjeraDesnoOdPočetka(desno)
          ako (provjeraLijevo)
               DodajProširenicu (lijevo)
          inače ako (provjeraDesno)
               DodajProširenicu (desno)
          inače
          {    pronašao = TražiPodudarajućaSlova(mjesto)
               ako (pronašao)
                    DodajProširenicu (mjesto)
          }
     }
     inače
     {    pronašao = TražiPodudarajućaSlova(mjesto)
          ako (pronašao)
               DodajProširenicu (mjesto)
     }
}
```

3.1.1. Primjer postupka ekstrakcije

Prikazat će se primjer rada sustava za primjer: "*UN će od zemalja donatora tražiti gotovo pola milijarde dolara radi financiranja pomoći prognanima s Kosova do kraja godine, objavilo je u srijedu Visoko povjerenstvo UN-a za izbjeglice (UNHCR).*"

Korak 1. U tekstu (rečenici) traže se moguće kratice. Ako rečenica sadrži moguću kraticu, obrada se nastavlja.

> ***UN*** *će od zemalja donatora tražiti gotovo pola milijarde dolara radi financiranja pomoći prognanima s Kosova do kraja godine, objavilo je u srijedu Visoko povjerenstvo* ***UN****-a za izbjeglice (****UNHCR****).*

Korak 2. Izbacuju se prijedlozi, veznici i morfološki nastavci kratica iz teksta.

> *UN će __ zemalja donatora tražiti gotovo pola milijarde dolara radi financiranja pomoći prognanima _ Kosova __ kraja godine, objavilo je _ srijedu Visoko povjerenstvo UN__ __ izbjeglice (UNHCR).*

Korak 3. Traži se ekspanzija za moguću kraticu.

> ***UN*** *će zemalja donatora tražiti gotovo pola milijarde dolara radi financiranja pomoći prognanima Kosova kraja godine, objavilo je srijedu Visoko povjerenstvo UN izbjeglice (UNHCR).*

Korak 4**.** Kratica "*UN*" nije unutar zagrada i zato se gleda podudaraju li se slova kratica s početnim slovima niza riječi. U tekstu ne postoji takav niz riječi (zaredom) da prva riječ počinje sa slovom "*u*", a druga slovom "*n*". Prema tome, ekspanzija za "*UN*" ne postoji. Možemo primijetiti da je i sljedeća moguća kratica "*UN*" te istim postupkom dolazimo do odluke da ekspanzija za njega ne postoji**.**

Korak 5. Za moguću kraticu "*UNHCR*" postupak teče malo drugačije budući da se nalazi unutar zagrade**.** Prvo se gleda broj slova (za UNHCR *duljina=5*), što znači da se kao niz uzimaju 5 riječi ispred zagrade. Uz to se provjeravaju prve tri riječi ispred zagrade. Ako su njihova početna slova redom: malo, veliko, veliko slovo, onda ekspanzija sadrži ime i tu postupak staje. No, u našem slučaju redoslijed je: malo, veliko, malo slovo, što znači da ekspanzija ne sadrži ime pa se postupak nastavlja.

> *UN će zemalja donatora tražiti gotovo pola milijarde dolara radi financiranja pomoći prognanima Kosova kraja godine, objavilo je* ***srijedu Visoko povjerenstvo UN izbjeglice*** *(UNHCR).*

Korak 6. Peta riječ ispred zagrade je početna. Ako ona ima početno slovo veliko, ekspanzija se sastoji od tih pet riječi. Naš primjer nije toliko jednostavan pa nastavljamo s potragom. Krećemo nadesno od početne riječi. Nadesno se krećemo maksimalno onoliko riječi od početne kolika je duljina kratica (*duljina=5*). Trenutna riječ je ona za koju se provjeravaju svojstva. Prvo se provjerava završava li riječ sa ",", odnosno nalazi li se iza trenutne riječi, jer se u tom slučaju postupak kretanja nalijevo završava i kreće se nadesno. Zasad to nije slučaj pa se gleda početno slovo (kod nas je malo – "*je*"). U slučaju da je veliko postupak se završava i ekspanzija je niz riječi od trenutne riječi do riječi ispred zagrede.

> *UN će zemalja donatora tražiti gotovo pola milijarde dolara radi financiranja pomoći prognanima Kosova kraja godine, objavilo* ***je*** srijedu Visoko povjerenstvo UN izbjeglice *(UNHCR).*

Korak 7. Postupak ulijevo se nastavlja i sljedeća je riječ "*objavilo*", što je naznaka da se postupak nastavlja budući da je ovo tek druga riječ od pet, što je naša granica. Na riječi "*godine,*" postupak nalijevo staje zbog zareza. To bi značilo da postupkom ulijevo nismo pronašli ekspanziju i kreće postupak udesno.

> *UN će zemalja donatora tražiti gotovo pola milijarde dolara radi financiranja pomoći prognanima Kosova kraja* ***godine,*** *objavilo je* srijedu Visoko povjerenstvo UN izbjeglice *(UNHCR).*

Korak 8. Postupak udesno se odvija riječ po riječ sve do zagrade. Ako se na putu nađe riječ s početnim velikim slovom, tvori se ekspanzija od te riječi do zagrade. U našem slučaju odmah prva riječ do početne ima početno slovo veliko i prema tome tu postupak traženja završava.

> *UN će zemalja donatora tražiti gotovo pola milijarde dolara radi financiranja pomoći prognanima Kosova kraja godine, objavilo je* srijedu **Visoko povjerenstvo UN izbjeglice** *(UNHCR).*

Korak 9. U zadnjem koraku se u se u rečenicu vraćaju prijedlozi i ostale oznake iz originalne rečenice te se označavaju ekspanzija i kratica.

> *UN će od zemalja donatora tražiti gotovo pola milijarde dolara radi financiranja pomoći prognanima s Kosova do kraja godine, objavilo je u srijedu <E1>Visoko povjerenstvo UN-a za izbjeglice</E1> (<A1>UNHCR</A1>).*

Nakon što se postupak odradi za čitav tekst, označene se rečenice zapisuju u datoteku nakon čega se vrši evaluacija.

3.2.SVM-metoda

Kao što je rečeno, SVM-metoda je rađena po uzoru na već ranije opisanu metodu uz neke preinake. Skup mogućih kratica gradio se samo na temelju velikih slova, odnosno bez brojeva, razmaka i specijalnih znakova. Razlog tome jest što je analizom utvrđeno da se na svakih 200-tinjak stvarnih kratica javlja tek jedna koji se ne sastoji samo od velikih slova. Prema tome, takve su značajke, radi pojednostavljivanja modela, a uz uvođenje gotovo neprimjetne greške, izbačene. Posljedica toga je izbacivanje dijela značajki koje više nemaju smisla. Tako su npr. broj malih slova, brojeva, specijalnih znakova i razmaka za definirani skup nebitne informacije, jer niti ne postoje, odnosno svaki je od tih brojeva jednak nuli za svaku kraticu. Izbacivanjem značajki oblikuje se konačni skup značajki:

Tablica 1 - Popis značajki za SVM-metodu

	Opis značajke:	Vrsta vrijednosti:
1.	Duljina kratica	Cjelobrojna
2.	Broj riječi ekspanzije	Cjelobrojna
3.	Broj riječi ekspanzije s početnim velikim slovom	Cjelobrojna
4.	Broj prijedloga	Cjelobrojna
5.	Podudaraju li se slova kratica s početnim slovima ekspanzije	Binarna
6.	Nalazi li se kratica unutar zagrada	Binarna
7.	Nalazi li se ekspanzija u desno ili lijevom kontekstu s obzirom na kraticu	Binarna

Za implementaciju SVM koristila se gotova biblioteka *libSVM* [11]. Kao jezgrena funkcija koristila se radijalna bazna funkcija (engl. *radial basis function*) iz razloga što za problem ekstrakcije kratica daje optimalne rezultate. Za nju su bitna dva parametra: C i γ. Potrebno je identificirati dobar par parametara C i γ kako bi klasifikator mogao ispravno klasificirati nepoznate parove kratica i njihovih ekspanzija. Njihovo se pretraživanje vrši na način da se poznati (označeni) skup odnosno skup za učenje podijeli na v podskupova od kojih se odvaja jedan po jedan te se nad njim vrši testiranje klasifikatora koji se uči na ostalih v-1 podskupova. Taj se postupak naziva unakrsna provjera (engl. c*ross-validation*). Nad različitim se parovima (C i γ) vrijednosti vrši ispitivanje unakrsnom provjerom te se odabire par s najvećom točnošću.

Kao C i γ vrijednosti se uzimaju eksponencijalno rastući slijedovi (npr. $C=2^{-5}, 2^{-3},..., 2^{15}$ te $\gamma=2^{-15}, 2^{-13},..., 2^{3}$), jer se oni prema [12] smatraju najpovoljnijima pri identifikaciji dobrih parametara. Nakon što se oba optimalna parametra pronađu, cijeli se skup ponovo koristi za učenje kako bi se generirao završni klasifikator.

4. Rezultati

Pri izgradnji referentne metode vrlo je važna bila detaljna analiza ručno označenih primjera. Zbog toga je takva analiza provedena prije početka implementacije.

4.1.Analiza ručno označenih rezultata

Program je građen nad korpusom od 10 godina Vjesnika (1999-2009.). Analiza nije provedena nad cijelim korpusom, već samo nad manjim dijelom. No, kako postoji mogućnost da se stil pisanja novinara, odnosno kolumnista mjenjao tokom 10 godina, u analizu je uzet dio starijih i dio novijih vijesti. Starije su vijesti one koje se odnose na početne, a novije na zadnje vijesti tokom navedenih 10 godina. Rezultati analize prikazani su tablicama 2 i 3.

Tablica 2 - Prikaz rezultata analize za starije vijesti

Jezik kratice		E(K)	K(E)	E-K	K-E	E, K	K, E	E…K	K…E	E(K)	ostalo
Hr/Strani	148	141	5	0	0	0	0	1	1	141 (95.27%)	7 (4.73%)
Strani/Strani	16	12	2	0	0	0	0	2	0	12 (75%)	4 (25%)
Hr	164	118	1	2	1	0	0	30	12	118 (59.76%)	46 (40.24%)
Ukupno:	**328**	**271**	**8**	**2**	**1**	**0**	**0**	**33**	**13**	**271 (76.52%)**	**57 (23.48)**

Tablica 3 - Prikaz rezultata analize za novije vijesti

		E(K)	ostalo
Hr/Strani	89	82 (92.13%)	7 (7.87%)
Strani/Strani	20	14 (70%)	6 (30%)
Hr	126	96 (76.19%)	30 (23.81%)
Ukupno:	**235**	**192 (81.7%)**	**43 (18.3%)**

Oznake unutar tablica tumače se na sljedeći način:

E(K) – ekspanzija E se nalazi neposredno prije zagrada unutar kojih se nalazi kratica K,

K(E) – kratica K se nalazi neposredno prije zagrada unutar kojih se nalazi ekspanzija E,

E-K – ekspanzija E i kratica K su odjeljeni samo znakom "-",

K-E – kratica K i ekspanzija E su odjeljeni samo znakom "-",

E, K – ekspanzija E i kratica K su odjeljeni samo zarezom,

K, E – kratica K i ekspanzija E su odjeljeni samo zarezom,

E…K – ekspanzija E i kratica K su odjeljeni proizvoljnom količinom i proizvoljnim sadržajem teksta,

K…E – kratica K i ekspanzija E su odjeljeni proizvoljnom količinom i proizvoljnom sadržajem teksta.

Hr/Strani – kratica se odnosi na kraticu ekspanzije na stranom jeziku, a njegova ekspanzija je na hrvatskom jeziku,

Strani/Strani – kratica i njegova ekspanzija su oboje na stranom, ali istom jeziku,

Hr – kratica i njegova ekspanzija su oboje na hrvatskom jeziku.

Budući da rezultati daju vrlo veliku signifikantnost izrazu E(K), skraćenom tablicom se može s velikom preciznošću predočiti oblik u kojem se nalaze kratice i njihove ekspanzije. Zbog toga je tablica 3 predočena skraćeno. Postojanje razlike odnosno njeno nepostojanje u stilu pisanja trebalo bi utvrditi statističkim testom, no kako to nije fokus ovog rada, neće biti obrađeno.

4.2.Tablica zabune

Evaluacija referentne i SVM-metode izvedena je pomoću tablice zabune i evaluacijskih mjera preciznosti, odziva i F1-mjere. Primjer tablice zabune nalazi se u tablici :

Tablica 4 - Tablica zabune

		Stvarno stanje, `oznacen_par_rucno`	
		ISTINA	LAŽ
Predviđeno stanje, `oznacen_par_program`	ISTINA	Točno pozitivan True Positive = TP	Lažno pozitivan False Positive = FP
	LAŽ	Lažno negativan False Negative = FN	Točno negativan True Negative =TN

Kako korištene evaluacijske mjere preciznost, odziv te F1-mjera ne iziskuju računanje vrijednosti TN, ona se u okviru ovog rada niti ne računa. Korištene evaluacijske mjere [10] mogu se definirati:

- *Preciznost* (engl. *precision*) je udio točno klasificiranih primjera:
  - $Preciznost = \frac{TP}{TP+FP}$
- *Odziv* (engl. *recall*) je udio točno klasificiranih primjera u skupu svih pozitivno klasificiranih primjera:
  - $Odziv = \frac{TP}{TP+FN}$
- *F1-mjera* je harmonijska sredina preciznosti i odziva, koja jednaku važnost pridaje preciznosti i odzivu:
  - $F1 - mjera = \frac{2 \cdot preciznost \cdot odziv}{preciznost+odziv}$

4.3.Analiza referentne metode

Iz prethodne analize je jasno da je forma kratica-ekspanzija većim djelom E(K) i zbog toga je to prvo pravilo po kojem se traži ekspanzija u referentnoj metodi.

Rezultate za metodu opisanu u poglavlju 3.1. možemo prikazati tablicama 5 i 6:

Tablica 5 - Rezultati analize za referentnu metodu

	ISTINA	LAŽ
ISTINA	TP = 80	FP = 31
LAŽ	FN = 20	TN = 0

Tablica 6 - Evaluacijske mjere za referentnu metodu

Evaluacijska mjera	Vrijednost (%)
Preciznost	72,1
Odziv	80,0
F1-mjera	75,8

U tekstu se javlja 31 pogrešno označen par kratica-ekspanzija. To se u najvećem djelu javlja zbog kratkih potencijalnih kratica (poput *UN, EU,...).* Primjeri progrešno označenih parova:

Primjer 1.

UN *utorak novi*

EU *europski uspjeh*

JNA *jedan od najvećih apsurda*

Osim njih pogrešno se označuju i primjeri s kraticom unutar zagrada, kada se prije zagrade javlja ekspanzija koja se sastoji od manjeg broja riječi nego što kratica ima slova, jer se program početno postavlja na onu riječ prije zagrade koliko kratica ima slova. Kako od početne riječi ide ulijevo tražeći riječ s velikim slovom, javlja se problem kada ju pronađe, jer u tom slučaju program završava s radom i ekstrahira ekspanziju od te riječi sve do zagrade.

Primjer 2.

> ***Informativni centar Hrvatskog autokluba** (**HAK**) izvijestio je u utorak da je zbog bure prekinut trajektni promet…*

U primjeru 2. program će se za kraticu *HAK* početno postaviti na riječ *centari* (jer je ona treća riječ prije zagrade, koliko kratica *HAK* ima slova), tražit će sastoji li se ta riječ od početnog velikog slova, što nije istina, te se zaputiti ulijevo. Kako se naša kratica sastoji od 3 slova, upravo toliko riječi će ulijevo program tražiti riječ s početnim slovom. Već prva riječ *Informativni* ima veliko slovo, što znači da program završava s radom za tu kraticu i dodijeljuje ekspanziju od pronađene riječi (*Informativni*) sve do zagrade (u kojoj se nalazi kratica) tako tvoreći ukupnu ekspanziju *Informativni centar Hrvatskog autokluba*, što je pogrešno. Prema tome označuje se par:

HAK Informativni centar Hrvatskog autokluba

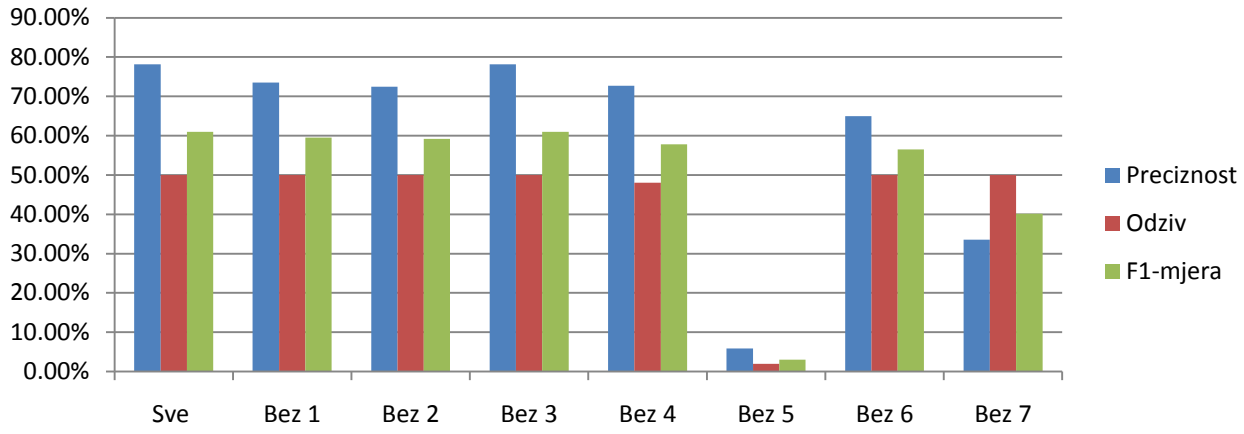
4.4.Analiza SVM metode

Kako bismo lakše pratili analizu, prisjetimo se značajki korištenih u SVM-metodi (tablica 1.). U tablici 7 navedene su rezultati SVM-metode uz nestandardizirane vrijednosti.

Tablica 7 - Rezultati SVM-metode uz nestandardizirane vrijednosti

TP	FP	FN	Preciznost	Odziv	F1-mjera
50	14	50	78.13%	50%	60.98%

Isključivanjem značajki pokušao se postići optimalan rezultat, a uz skup za učenje od 200 parova i skup za ispitivanje od 100 parova rezultati su prikazani slikom 3:



Slika 3 - Usporedba rezultata isključivanjem jedne značajke

Kako su rezultati evaluacije bez treće značajke (*F1-mjera = 60.98%*), odlučujemo izbaciti upravo nju iz skupa značajki. Kako bismo pokušali još više poboljšati rezultate primijenit ćemo standardizaciju vrijednosti značajki iz razloga što njome izbjegavamo dominaciju atributa s većim numeričkim rasponom od onih s manjim numeričkim rasponom. Bitno je napomenuti da i cjelobrojne i binarne vrijednosti u tablici 1 standardizacijom postaju realne. Rezultati evaluacije uz standardizirane vrijednosti značajki nalaze se u tablici 8.

Tablica 8 - Rezultati SVM-metode uz standardizaciju svih vrijednosti

	Preciznost	Odziv	F1-mjera
Sve značajke	78.13%	50%	60.98%
Bez 1.	74.24%	49%	59.04%
Bez 2.	72.46%	50%	59.17%
Bez 3.	79.37%	50%	61.35%
Bez 4.	**83.93%**	47%	60.26%
Bez 5.	5.88%	2%	2.99%
Bez 6.	78.13%	50%	60.98%
Bez 7.	81.97%	**50%**	**62.11%**

Optimalni rezultati za svaku evaluacijsku mjeru korištenjem standardizacije vrjiednosti ostvaruju se ukidanjem sedme značajke. Tako je u tom slučaju *odziv=50%* i *F1-mjera=62.11%*. Jedino je preciznost bolja u slučaju ukidanja četvrte značajke (*preciznost=83.93%*). Standardizacija svih vrijednosti isključivanjem sedme značajke uspjela je poboljšati rezultate, jer je najbolja *F1-mjera = 62.11%*, što je više od dosadašnjih 60.98% .

Sljedeći korak je standardizacija samo cjelobrojnih vrijednosti pri čemu cjelobrojne vrijednosti tablice 1 postaju realne, no binarne ostaju binarne uz malu preinaku. Ako je binarna vrijednost značajke jednaka 1, ona ostaje jednaka 1, no u slučaju da je binarna vrijednost jznačajke jednaka nula, ona postaje jednaka -1. U tom se slučaju dobivaju rezultati prikazani tablicom 9.

Tablica 9 - Rezultati SVM-metode uz standardizaciju cjelobrojnih vrijednosti

	Preciznost	Odziv	F1-mjera
Sve značajke	79.37%	50%	61.35%
Bez 1.	72.46%	50%	59.17%
Bez 2.	72.46%	50%	59.17%
Bez 3.	79.37%	50%	61.35%
Bez 4.	9.69%	47%	16.07%
Bez 5.	5.88%	2%	2.99%
Bez 6.	78.13%	50%	60.98%
Bez 7.	**81.97%**	**50%**	**62.11%**

Rezultati SVM-metode uz standardizaciju cjelobrojnih vrijednosti ne razlikuju se bitno od rezultata uz standardizaciju svih vrijednosti. Ukidanje sedme značajke rezultira optimalnim rezultatima koji su u potpunosti jednaki optimalnim rezultatima uz standardizaciju svih vrijednosti: *preciznost=81.97%, odziv=50%* i *F1-mjera=62.11%.*

U našem slučaju rezultati se bolji kada se standardiziraju sve (ili samo cjelobrojne) vrijednosti iz tablice 1. Prema tome, za izgradnju optimalnog stroja s potpornim vektorima, koristit ćemo standardizirane vrijednosti značajki. Kako je u tom slučaju optimalan rezultat dobiven ukidanjem sedme značajke, uz nju ćemo ukidati još jednu značajku u pokušaju boljih rezultata. Ukidat ćemo treću odnosno šestu značajku budući da su one imale najbolje rezultate nakon rezultata dobivenih ukidanjem sedme značajke.

Tablica 10 - Rezultati SVM-metode uz standardizirane sve vrijednosti ukidanjem 2 značajke

	Preciznost	Odziv	F1-mjera
Bez 7.	**81.97%**	**50%**	**62.11%**
Bez 3. i 7.	**81.97%**	**50%**	**62.11%**
Bez 6. i 7.	**81.97%**	**50%**	**62.11%**

Uspoređujući rezultate SVM-metode uz standardizirane sve vrijednosti i ukidanjem dviju značajki s rezultatima ukidanjem sedme značajke zaključujemo da je dovoljno ukinuti samo sedmu značajku. Uz ukinute dvije značajke iz tablice 10, možemo vidjeti da uz sedmu,

ukidanjem treće ili šeste značajke dobivamo iste rezultate: *preciznost=81.97%, odziv=50%* i *F1-mjera=62.11%*. Kao usporedbu možemo promotriti rezultate koje su dobili autori rada [8]. Naime, njihovom SVM-metodom uspjeli su postići odziv od 84.13% uz preciznost od 90.86%, što je rezultiralo F1-mjerom od 87.37%, a to je znatno bolji rezultat od našeg optimalnog rezultata korištenjem SVM-metode. Međutim, skup nad kojem su oni radili (čije su kratice i njihove ekspanzije na engleskom jeziku) također se bitno razlikuje od našeg skupa čiji su problem opisani u nastavku.

Nakon provedenih rezultata za referentnu i SVM-metodu zaključujemo da referentna metoda (*F1-mjera=75,83%*) ima bolje rezultate od optimalne SVM-metode (*F1-mjera=62,11%*). Razlog tome je što se tekst sastoji od mnogo stranih kratica uz hrvatske prijevode ekspanzija. Klasifikator SVM uz značajke koje ne sadrže prijevod na hrvatski jezik ne može dati bolje rezultate. Takva značajka u radu nije uvedena, no svakako je moguće poboljšanje u tom smjeru.

Kada bi se tekst sastojao samo od hrvatskih kratica s hrvatskim ekspanzijama i stranih kratica sa stranim ekspanzijama (na istom jeziku), rezultati bi bili znatno bolji. U tu svrhu je proveden test nad skupom koji sadrži istojezične kratice i njihove ekspanzije. Ispitivanje je pokrenuto uz standardizaciju svih vrijednosti ukidanjem jedne značajke nad skupom koji se sastoji od 157 primjera za učenje te 100 primjera za ispitivanje. U tablici 11 navedeni su rezultati SVM-metode uz nestandardizirane vrijednosti samo za parametre uz koje je F1-mjera bila veća od 88%. Optimalna vrijednost je postignuta ukidanjem treće značajke budući da je F1-mjera u tom slučaju jednaka 90.57%.

Tablica 11 - Rezultati SVM-metode na skupu istojezičnih kratica i ekspanzija uz standardizirane sve vrijednosti

	Preciznost	Odziv	F1-mjera
Sve značajke	84.96%	96%	90.14%
Bez 1.	81.36%	96%	88.07%
Bez 3.	**85.71%**	**96%**	**90.57%**
Bez 6.	83.33%	95%	88.79%
Bez 7.	81.36%	96%	88.07%

U našem slučaju, gdje se skup sastoji i od kratica te njihovih ekspanzija koje nisu istojezične, kao nadogradnju na SVM-metodu koristila se referentna metoda, koja je rezultirala mješovitom metodom.

4.5.Analiza mješovite metode

Mješovita SVM-metoda je zapravo SVM-metoda uz jednu dodatnu značajku, koja je dio referentne metode. Pretražuje se ekspanzija za moguću kraticu pomoću referentne metode. Ukoliko niz, koji SVM-metoda ekstrahira kako bi se odredile značajke, sadrži tu ekspanziju, pridodaje mu se binarna vrijednost 1, inače nula. Nova značajka je dakle:

8. Podudaranje niza s rezultantnom ekspanzijom za moguću kraticu referentne metode (binarna vrijednost).

Tablica 12 - Rezultati mješovite metode uz nestandardizirane vrijednosti značajki

	Preciznost	Odziv	F1-mjera
Sve	**83.70%**	**77%**	**80.21%**
Bez 1	81.05%	77%	78.97%
Bez 2	77.78%	77%	77.39%
Bez 3	78.35%	76%	77.16%
Bez 4	78.57%	77%	77.78%
Bez 5	16.2%	76%	26.71%
Bez 6	78.57%	77%	77.78%
Bez 7	79.38%	77%	78.17%

U tablici 12 prikazani su rezultati mješovite metode bez standardizacije vrijednosti. Znatno su bolji od rezultata SVM-metode, a također su bolji od rezultata referentne metode. Optimalni rezultati ostvaruju se korištenjem svih značajki: *preciznost=83.7%, odziv=77%* i *F1-mjera=80.21%*.

Tablica 13 - Rezultati mješovite metode uz standardizaciju svih vrijednosti značajki

	Preciznost	Odziv	F1-mjera
Sve	65.25%	77%	70.64%
Bez 1	77%	77%	77%
Bez 2	73.33%	77%	75.12%
Bez 3	76.24%	77%	76.62%
Bez 4	76.24%	77%	76.62%
Bez 5	51.7%	76%	61.54%
Bez 6	24.92%	77%	37.65%
Bez 7	**77.78%**	**77%**	**77.39%**

Rezultati u tablici 13 ne pokazuju poboljšanja obzirom na rezultate postignute uz nestandardizirane vrijednosti. Optimalni rezultati ostvaruju se uz standardizaciju svih vrijednosti te ukidanjem sedme značajke: *preciznost=77.78%, odziv=77%* i *F1-mjera=77.39%.*

Tablica 14 - Rezultati mješovite metode uz standardizaciju cjelobrojnih vrijednosti značajki

	Preciznost	Odziv	F1-mjera
Sve	74.04%	77%	75.49%
Bez 1	**77%**	**77%**	**77%**
Bez 2	73.33%	77%	75.12%
Bez 3	76.24%	77%	76.62%
Bez 4	52.74%	77%	62.6%
Bez 5	52.78%	76%	62.3%
Bez 6	74.04%	77%	75.49%
Bez 7	18.92%	77%	30.37%

Tablica 14 potvrđuje najbolje rezultate nestandardiziranih vrijednosti. Optimalni rezultati uz standardizirane cjelobrojne vrijednosti ostvaruju se ukidanjem prve značajke: *preciznost=77%, odziv=77%* i *F1-mjera=77%.*

Nakon provedenih rezultata za referentnu i SVM-metodu zaključujemo da mješovita metoda (*F1-mjera=80,21%*) ima bolje rezultate od optimalne referentne metode (*F1-mjera=75,83%*), što smo i očekivali.

5. Zaključak

U radu je obrađen problem ekstrakcije akronima i njihovih proširenica iz teksta. Nad pristupom koji je bio fokus rada, metoda stroja s potpornim vektorima, provedeno je mnogo ispitivanja s mnogo različitih parametara među kojima je standardizacijom svih vrijednosti uz ukidanje značajke koja određuje radi li se o lijevom ili desnom kontekstu postignut optimalan rezultat. Taj rezultat nije nadmašio rezultat referentne metode, no razlog tome je prisutstvo stranih akronima uz prijevod njihovih ekspanzija na hrvatski jezik. Kako bi ta konstatacija bila potvrđena, provedeno je ispitivanje nad skupom iz kojeg su izbačeni neistojezični parovi akronim-ekspanzija. Rezultati su tada značajno bolji od referentne metode. Nažalost, to nije realna situacija budući da se takvi primjeri parova redovito javljaju u tekstovima (npr. novinskim).

U pokušaju poboljšanja implementirana je i mješovita metoda koja kao značajku SVM-metode koristi ekstrakciju ekspanzija referentne metode. Metoda je donijela poboljšanja s obzirom na referentnu metodu, no i dalje su rezultati istojezičnog korpusa mnogo bolji.

U budućem radu bilo bi poželjno testirati programsko ostvarenje nad novim korpusom kako bi se dokazala modularnost sustava, a kao moguće poboljšanje svakako bi bilo uvođenje značajki koje sadrže prijevod na hrvatski jezik.

6. Literatura

[1] Pustejovsky, Castano, Cochran, Kotecki, Morrell: "*Automatic extraction of acronym meaning pairs from MEDLINE databases*", Medinfo 10 (Pt 1): 371–375, 2001.

[2] Larkey, Ogilvie, Price, Tamilio. "*Acrophile: An automated acronym extractor and server*", Proceedings of the 5th ACM conference on digital libraries, ACM Press, San Antonio, 205-214, 2000.

[3] Park, Byrd: "*Hybrid text mining for finding abbreviations and their definitions*", Proceedings of the 2001 conference on empirical methods in natural language processing, Pittsburgh, 126-133, 2001.

[4] Yu, Hripcsak, Friedman: "*Mapping abbreviations to full forms in biomedical articles*", J Am Med Inform Assoc 9, 262-272, 2002.

[5] Taghva, Gilbreth: "*Recognizing acronyms and their definitions*", Technical Report, ISRI (Information Science Research Institute), UNLV, 1999.

[6] Bowden, Halstead, Rose: "*Dictionaryless English plural noun singularisation using a corpus-based list of irregular forms*", Proceedings of the 17th international conference on English Language Research on Computerized Corpora, Rodopi, Amsterdam, The Netherlands, 130-137, 2000.

[7] Chang, Schutze, Altman: "*Create an online dictionary of abbreviation from MEDLINE*", J Am Med Inform Assoc, 9(6), 612-620, 2002.

[8] Xu, Huang: "*Using SVM to extract acronyms from text*", Springer-Verlag, Soft Comput, 2006.

[9] Vapnik: "*The nature of statistical learning theory*", Springer, Berlin Heidelberg New York, 1995.

[10] Dalbelo Bašić, Bojana: "*Dodatak: Nadzirano Učenje - Evaluacija*", Fakultet elektrotehnike i računarstva, Zagreb, 2010.

[11] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[12] Hsu, Chang, Lin: "*A Practical Guide to Support Vector Classification*", Department of Computer Science, National Taiwan University, Taipei 106, Taiwan, 2003.

Sažetak

Postupak ekstrakcije složenih kratica hrvatskoga jezika

Cilj ekstrakcije složenih kratica hrvatskoga jezika je razvitak tehnika koje automatski ekstrahiraju kratice i njihove pripadajuće ekspanzije iz teksta. Razvijena su tri različita pristupa: referentna metoda, metoda potpornih vektora te njihova kombinacija. Više je pristupa korišteno kako bi se mogli usporediti te iz njih izvući zaključci. Korištenje metode potpornih vektora zahtjevalo je najviše optimiranja parametara. Tako su njene značajke učene i testirane na tri različita načina: nestandardizirane, standardizirane sve, te standardizirane cjelobrojne značajke. Tekst nad kojim se vršio postupak ekstrakcije je 10 godina Vjesnika (1999-2009.). Uspješnost ostvarenja je evaluirana i uspoređena s ručno označenim parovima kratice i ekspanzije.

Ključne riječi: obrada prirodnog jezika, ekstrakcija složenih kratica, stroj s potpornim vektorima, metoda temeljena na uzorcima

Abstract

Acronym extraction in Croatian language

The objective of acronym extraction in Croatian language is to develop technologies that automatically extract acronyms and their expansions in text. There are three different approaches developed: baseline method, support vector machine method and their combination. More approaches are used so we could compare them and get some conclusions. Use of support vector machine method required most of parameter optimizing. It's features are learned and tested in three different ways: non-standardized, all standardized and only integer standardized features. Text over which extraction is performed is 10 years of "Vjesnik" newspaper (1999-2009.). Performance of implementation is evaluated and compared to manually marked acronym and expansion pairs.

Keywords: natural language processing, acronym extraction, support vector machine, rule-based method