Sveučilište u Zagrebu
Filozofski fakultet
Odsjek za informacijske i komunikacijske znanosti

Lobel Filipić, Tereza Jurić, Marija Stupar

Strojno prepoznavanje naziva u tekstovima pisanima hrvatskim jezikom

Mentor: doc. dr. sc. Nikola Ljubešić

Zagreb, 2012.

Ovaj rad izrađen je na Odsjeku za informacijske znanosti Filozofskog fakulteta Sveučilišta u Zagrebu pod vodstvom doc. dr. sc. Nikole Ljubešića i predan je na natječaj za dodjelu Rektorove nagrade u akademskoj godini 2011/2012.

Sadržaj

1. Uvod

Prepoznavanje naziva (engl. *Named Entity Recognition*) jest problem iz područja obrade prirodnog jezika koji se vrlo aktivno istražuje posljednje desetljeće. Takvim je postao napretkom računalne znanosti, interesom za lingvistička istraživanja unaprijeđena tehnologijom. Stoga prepoznavanje je naziva u punome smislu interdisciplinarno područje.

Strojno prepoznavanje naziva dio je crpenja (ekstrakcije) informacija (engl. *information extraction*) čime se želi pronaći i klasificirati elemente u prethodno određene kategorije kao što su osobe, lokacije, organizacije, vremenske oznake, oznake količina, valute, trgovačke marke, udruge itd.

2. Teorijska podloga

Tema prepoznavanja naziva začeta je na konferenciji Sixth Message Understanding Conference (MUC-6) (Grishman, Sundheim, 1996). Uloga je sustava za strojno prepoznavanje naziva da u neoznačenu tekstu istakne nazive koji mogu biti značajni u procesu ekstrakcije informacija.

Rani sustavi koristili su se ručno izrađenim algoritmima s pravilima, a moderniji se oslanjaju na metode strojnog učenja.

Od 1991. do 1995. malen je broj objavljenih radova na ovome području, a od 1996. osjetno raste zanimanje za razvoj sustava. Upravo te godine organiziran je MUC-6 (Message Understanding Conference), prvi veliki i važan događaj na kojemu su se postavili temelji daljnjemu istraživanju (Grishman, Sundheim, 1996). Uslijedile su konferencije na kojima se predstavljalo dosege u području temeljene na označavanjima, te se domena istraživanja proširila na mnoge druge jezike osim engleskog.

Prve konferencije nakon MUC-a bavile su se kineskim i japanskim (Multilingual Entity Extraction Conference – MET), španjolskim i nizozemskim te njemačkim i engleskim (Conference on Computational Natural Language Learning – CoNLL), portugalskim (Evaluation contest for named entity recognizers in Portuguese – HAREM), francuskim (Speech technology evaluation for the French language – ESTER/ETAPE).

Rani sustavi obrađivali su novinske tekstove da bi kasnije prešli na vojne dopise i izvještaje. Nakon tih prilično formalnih i pravilima utvrđenih tekstova počelo se obrađivati neformalne tekstove s internetskih dnevnika – blogova, transkripata, telefonskih razgovora, a u posljednje vrijeme najveća primjena ostvarila se u strogo stručnim tekstovima molekularne biologije i bioinformatike.

Jedan od najvećih projekata trenutačno je Global Autonomous Language Exploitation (GALE[1]). Program vodi i nadzire DARPA[2], agencija Ministarstva obrane Sjedinjenih Američkih Država. Uloga je te agencije razvoj novih tehnologija za vojsku SAD-a.

[1] http://projects.ldc.upenn.edu/gale/
[2] http://www.darpa.mil/

GALE je računalni program koji prima, analizira i tumači velike količine pisanoga i govorenoga teksta na mnogim jezicima. Važan dio toga projekta upravo je strojno prepoznavanje naziva, iznimno važno kao priprema analitičkome radu obavještajnih službi.

U radu *A survey of named entity recognition and classification* (Nadeau, Sekine, 2007) napominje se kako se mala pozornost u literaturi koja se bavi teorijskom pozadinom područja pridaje prilagodbi sustava žanru i domeni teksta, a upravo je raznovrsnost tekstova jedan od problema prilagodbe sustava koji se ističe u mnogim praktičnim radovima. Kako autori navode, nekoliko je studija koje su specificirale područje tekstualne obrade kao što je rad nad korpusom e-pisama, znanstvenih i religijskih tekstova, rad nad isključivo e--pismima (Minkov et al, 2005).

Kao što se može iščitati iz zaključaka konferencija, izrada sustava koji će jednako kvalitetno obrađivati tekstove različitih žanrova velik je izazov.

Prepoznavanje naziva određeno je dogovorenim pravilima o tome što je točno naziv. Naziv je točno određena pojavnica ili niz pojavnica u tekstu koje se mogu vezati uz samo jedan konkretni pojam u stvarnosti. Ti su pojmovi imena i nazivi, određene prirodne pojave kao biološke vrste. U nazive su uključene i vremenske i brojčane oznake poput datuma i valutnih iznosa te drugih jediničnih oznaka. Potrebno je odmah napomenuti kako se posebna odstupanja lako događaju pri određivanju vremenskih jedinica (npr. „23. lipnja 2012.“ točno je određen dan u vremenu, dok izraz „u lipnju“ ne bi smio biti označen jer ne određuje jednoznačan trenutak, nije određeno koji je to lipanj, u kojoj godini).

Najveći dio radova dijeli nazive na osobe (engl. *person* – PERS), lokacije (engl. *location* – LOC) i organizacije (engl. *organization* – ORG) i te se tri klase nakon MUC-6 grupiraju pod nazivom ENAMEX. Na ranije spomenutome CoNLL-u uvedena je i klasa ostalo (engl. *miscellaneous* – MISC) pod koju se ubrajaju svi nazivi koji se ne mogu svrstati u neku od već postojećih (ENAMEX).

Osim ENAMEX-a definirani su skupovi klasa TIMEX (za datum i vrijeme) i NUMEX (za novac i postotak). Zbog kompleksnosti same zbilje i jezika uvijek je moguće dodati zasebnu klasu izvan neke od zadanih grupa: npr. e-adresa, telefonski broj, naziv projekta, naziv posla itd.

Svaki istraživač može uvesti klase za koje smatra da su važne u crpenju informacija i za koje misli da tekst nad kojim radi obradu njih sadrži. Također dio istraživača sve klase koje se ne mogu svrstati u neke od definiranih skupova ENAMEX, TIMEX, NUMEX stavlja u skup MISC.

Prepoznavanje naziva u novije vrijeme kao važnu problematiku prepoznali su znanstvenici medicinskih i biomedicinskih znanosti. Stoga je bilo potrebno uvesti neke nove klase koje su njima važne: protein, DNA, RNA itd. Stvoren je korpus GENIA nad kojim se te strogo stručne i detaljne klase mogu uvježbavati. Skup je to tekstova biomedicinske literature obrađen i označen u projektu GENIA.[3]

Noviji radovi pokušali su označiti pojmove u širokome opsegu i te se pokušaje naziva radovima otvorene domene. Definirajući najčešće pojmove došlo se do broja od oko 200 klasa (Sekine, Nobata, 2004). Podklase obuhvaćaju detaljno razgranatu mrežu pojmova od muzeja, rijeka, proizvoda, događanja do životinjskih vrsta, religija i boja.

Budućnost prepoznavanja naziva je označavanje multimedijskih uradaka (videomaterijal i audiomaterijal), polunadzirano učenje, upotreba u strojnome prevođenju i dr.

Pregled vezanih radova

Rani radovi temeljili su se na ručno izrađenim pravilima dok se u novije vrijeme pribjeglo nadgledanome strojnome učenju. Stoga, moguće je primijeniti metodu temeljenu na gramatičkim pravilima te onu statističku. Ručno rađeni gramatički modeli obično imaju bolju preciznost ali po cijenu lošijega odaziva rezultata i dugotrajnoga računalnolingvističkog rada na sustavu. Sustav temeljen na statističkoj obradi zahtijeva veliku količinu ručno označena teksta za učenje sustava.

Trenutačno je najčešća metoda nadgledanoga učenja, a ona uključuje pristupe koji se temelje na algoritmima kao što su skriveni Markovljevi modeli, stabla odlučivanja, entropijski modeli i metode uvjetnih nasumičnih polja. Ova se vrsta učenja temelji na velikome označenom korpusu (Nadeau, Sekine, 2004).

Osim te koriste se također i metoda djelomično nadgledanoga učenja te metoda nenadgledanoga učenja koja se temelji na leksičkim resursima kao što je WordNet.

[3] http://www.nactem.ac.uk/genia/genia-corpus

Hrvatski je jezik s malo govornika, relativno maloga političkoga utjecaja i maloga tržišta. Stoga tek domaći stručnjaci prilagođavaju dostupne resurse hrvatskome. Slično vrijedi i za bliske nam slavenske narode i njihove jezike. U nastavku poglavlja razmotrit će se dosezi sustava koji se bave slavenskim jezicima.

Božo Bekavac u svojoj je doktorskoj disertaciji opisao sustav za prepoznavanje i klasifikaciju naziva za hrvatski jezik OZANA (Bekavac, 2005). Sustav sadrži modul za segmentaciju na rečenice, opći leksikon, specijalizirane popise imena i konačne transduktore za automatsko prepoznavanja brojeva i nekih oblika pridjeva. Samom prepoznavanju naziva prethodi obrada teksta i segmentacija na rečenice, obilježavanje pojavnica teksta (leme te morfosintaktičke značajke). Bekavčev sustav, kao i onaj predstavljen u ovom radu, analizu radi na morfosintaktički označenome tekstu, a razlika je u tome što se njegova klasifikacija temelji na pravilima, dok se u ovom radu koristi statistički model. OZANA je postigla F-mjeru od 90%.

Matko Bošnjak opisuje primjenu metode strojnoga učenja u stvaranju sustava (Bošnjak, 2007). Autor rada koristi se slabo nadziranom metodom koja se temelji na popisima naziva i sustavu za ekstrakciju naziva. Ukupni proces sastoji se od modula za označavanje i modula za ekstrakciju novih naziva pomoću pravila generiranih kontekstom. Dio procesa jest i slanje upita tražilici Google i prema rezultatima pretrage proračunava se vjerojatnost da je određena pojavnica ili niz pojavnica naziv.

Vitas i Pavlović-Lažetić sa Sveučilišta u Beogradu proučavali su kompleksnost prilagodbe sustava srpskome (Vitas, Pavlović-Lažetić, 2008). Ističu veliku razliku između engleskoga i srpskoga jezika (pa tako i ostalih slavenskih) koji zbog fleksije zahtijevaju opsežniju pripremu sustava, s obzirom da se radi o sustavu temeljenom na pravilima. Usto, srpski su tekstovi pisani dvama pismima, latinicom i ćirilicom. Ističu i mogućnost izvornoga pisanja naziva i fonetskoga što dodatno otežava pronalaženje naziva.

3. Cilj rada

U ovom radu korišten je postojeći alat Stanford NER[4] kojim se htjelo stvoriti model koji bi uspješno prepoznavao klase naziva u tekstovima na hrvatskom jeziku. Autore je zanimalo kakve su mogućnosti korištenja alata te hoće li izgrađeni sustav jednako reagirati u različitome domenskom okruženju. Kombiniranjem parametara za ekstrakciju karakterističnih obilježja teksta, načina obrade podataka te korištenjem algoritma strojnog učenja mjerila se preciznost, odaziv i F-mjera.

Alat se temelji na konceptu označavanja sekvenci, te za to koristi linearni model uvjetnih nasumičnih polja (engl. *Conditional Random Fields* – CRF). Sustav je dostupan pod GNU GPL licencom, te ga je moguće proširivati i prilagođavati. Uz sam sustav dolaze prethodno serijalizirani modeli za prepoznavanje naziva. Nekoliko je inačica tih modela ovisno o tome koliko skupova naziva prepoznaje: 3 skupa (osoba, lokacija, organizacija), 4 skupa (osoba, lokacija, organizacija, razno), 7 skupova (osoba, lokacija, organizacija, vrijeme, datum, novac, postotak).

Za izgradnju modela koristilo se korpusom s CoNLL-a, MUC-6, MUC-7 i ACE-a, redom konferencija na kojima su se predstavljali dosezi i usuglašavale buduće smjernice. Korištenjem tih korpusa s više konferencija postigla se robusnost modela nad domenama. Osim za engleski, sustav je prilagođen i za njemački jezik.

[4] http://nlp.stanford.edu/software/CRF-NER.shtml

4. Podaci za učenje

4.1 Priprema podataka

Za potrebe učenja i provjere algoritma pripremljeni su članci novinskog žanra iz hrvatskog web korpusa hrWaCa prikupljeni u razdoblju od siječnja do ožujka 2011. hrWaC je korpus tekstova prikupljenih s .hr domene koji obuhvaća oko 1.2 milijarde pojavnica uključujući interpunkcije (Ljubešić, Erjavec, 2011).

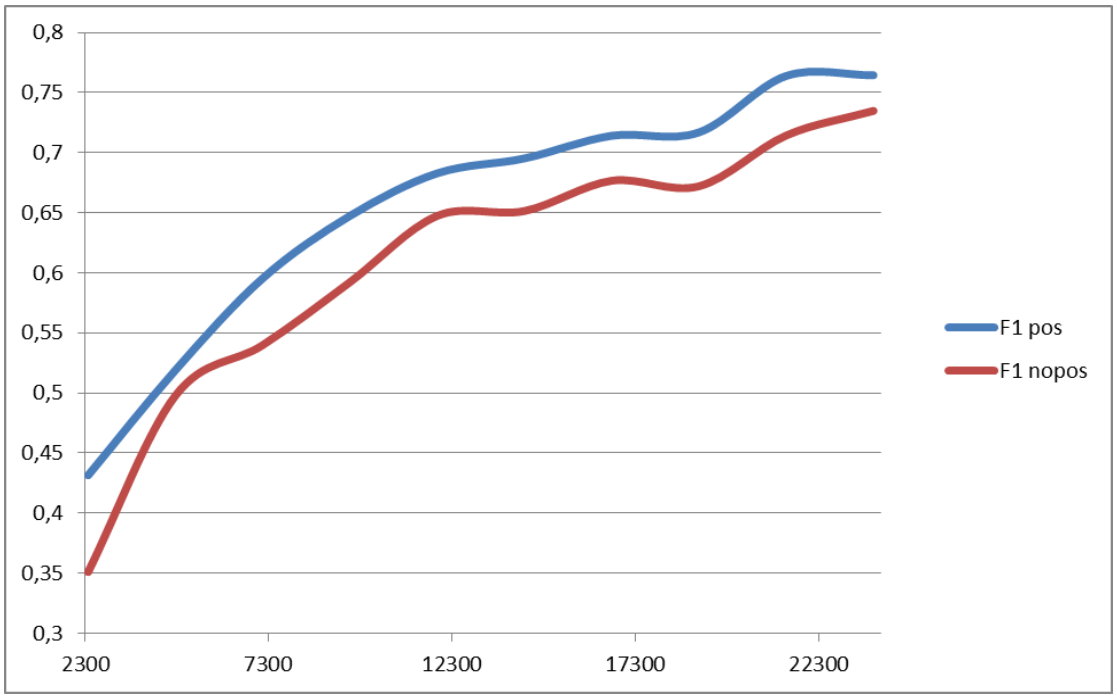
Sa ciljem izrade sustava za prepoznavanje naziva koji bi bio primjenjiv nad tekstovima ne samo jedne nego više različitih tematika, promatrano je više domena. Uzimajući u obzir domene s najvećim brojem pojavnica u hrWaCu, odabrani su tekstovi četiriju domena koji pokrivaju tri tematike: tehnološku (IT) – bug.hr, poslovnu – poslovni.hr i opću – nacional.hr i jutarnji.hr.

Da bi se tijekom označavanja procijenila konačna količina podataka u korpusu i korisnost oznaka vrste riječi (engl. *part-of-speech* - POS oznake) kojima su označavane njegove pojavnice, izrađena je krivulja učenja nad dijelom označenih podataka. Krivulja je dobivena obradom nečistih podataka, točnije podataka čije oznake nisu naknadno provjerene. Broj do tada označenih redova iznosio je 23,800, pri čemu je svaka pojavnica sa svojom oznakom vrste riječi pisana u novom redu. Za ekstrakciju značajki korištene su pretpostavljene vrijednosti parametara. Skup podataka za provjeru bio je stalan i obuhvaćao je 10% korpusa, a skup za učenje 90%. Podskupovi za učenje veličine 2,380 do 23,800 redova građeni su iz skupa za učenje slučajnim uzorkovanjem. Svako uzorkovanje ponovljeno je deset puta.

Krivulja učenja (Slika 1) dala je autorima uvid u brzinu i način učenja naziva. Učenje naziva označenih oznakama vrste riječi prikazano je „F1 pos“ krivuljom koja je u fazi mirnog rasta do otprilike 17,000 pojavnica, nakon čega počinje biti nemirna, ali je zapažen trend rasta – od otprilike 19,000 pojavnica do 24,000 pojavnica dobiveno je 6% u F1. Krivulja „F1 nopos“ dobivena računanjem F1 vrijednosti bez oznaka za vrstu riječi pokazala je niže vrijednosti F1 mjere od „F1 pos“ krivulje za istu veličinu korpusa. Također se pokazala nemirnijom, čime se da zaključiti da je manje izvjesnosti pri označavanju naziva bez oznaka vrste riječi.

Odlučeno je i dalje označavati oznake vrste riječi, te označiti još jednom količinu podataka približno istovjetnu već označenoj, pri čemu je očekivano skoro smirivanje rasta krivulje. Pri

opreznijem oblikovanju sustava moguće je da će napredak biti vidljiv i na većim količinama podataka, no moguće je i suprotno, da se potrebni uzorci nauče ranije.



Slika 1: Krivulja učenja naziva na prvoj polovici podataka gdje „F1 pos“ prikazuje krivulju F1 vrijednosti dobivenu uz oznake vrste riječi, a „F1 nopos“ bez tih oznaka

Konačni korpus opsega 59,143 pojavnice podijeljen je na skup za učenje (90%) sastavljen od 53,377 pojavnica i skup za provjeru (10%) algoritma sastavljen od 5,766 pojavnica.

4.2 Označavanje uzorka

Pojavnice obuhvaćene navedenim domenama ručno su označene od strane autora u dva navrata kako bi se osiguralo od nenamjernih pogrešaka nastalih pri označavanju. U opojavničenim tekstovima svaki novi redak predstavlja novu pojavnicu, a kraj rečenice predstavlja jedna prazna linija koja dolazi nakon završnog interpunkcijskog znaka. Svakoj pojavnici pridodana je oznaka vrste riječi, oznaka tipa klase kojoj pripada i oznaka pripadnosti pojavnice nazivu zapisana IOB2 formatom. Oznaka pripadnosti pojavnice nazivu crticom je odvojena od oznake tipa klase.

Oznake vrsta riječi

Za označavanje hrvatskih i stranih pojavnica korištene su sljedeće oznake vrste riječi opisane u (Tadić, 1998): imenica – N, glagol – V, pridjev – A, zamjenica – P, broj – M, član – D, članak – T, prilog – R, prijedlog – S, veznik – C, uzvik – I, čestica – Q, skraćenica – Y, preostalo – X, te interpunkcija – Z, koja nije navedena prema Tadiću, no redovito se koristi u praksi za označavanje interpunkcijskih znakova.

Pri označavanju je postignut dogovor između autora o svrstavanju vrsta riječi. Imenicama su označena sva opća i vlastita imena na hrvatskom ili nekom drugom jeziku. Prvi sporni primjer bilo je označavanje glagolskih pridjeva trpnih koji inače služe za tvorbu pasiva i pridjeva. Drugi sporni primjer bilo je označavanje glagolskih priloga sadašnjih i prošlih. Glagolski pridjevi označeni su kao pridjevi, glagolski prilozi kao prilozi, a ostali glagolski oblici kao glagoli. Nadalje, pridjevi tvoreni od skraćenica (npr. „ICM-ova“), kao i pridjevi tvoreni od vlastitih imena, označeni su kao pridjevi, dok je primjerice pri označavanju pojavnice „drugi“ trebalo voditi računa o kontekstu u kojem se pojavnica nalazi i prema tome odrediti je li ona broj ili pridjev. Također, trebalo je paziti na kontekst ostalih pojavnica jer su neke od njih višeznačne. Što se interpunkcijskih znakova tiče, točke koje su slijedile iza skraćenica, rednih brojeva i inicijala, znak postotka nakon broja, crtica u prezimenu i u nekim drugim slučajevima, promatrani su skupa s pojavnicama koje su im prethodile kao jedna pojavnica, a ostali interpunkcijski znakovi označeni su pripadnom oznakom. Kategorija nesvrstanih pojavnica X, dodatna je kategorija za koju su se autori odlučili nakon što je uočen broj pojavnica koji se ne može svrstati ni u jednu ranije definiranu kategoriju. Primjeri takvih nesvrstanih pojavnica su internetske domene, ekstenzije računalnih datoteka, nazivi serija proizvoda itd. od kojih su neki prikazani u Tablici 1 kojom je predočen dio rečenice iz označenog teksta.

Tablica 1: Primjer označenog teksta u našem skupu podataka

Serveri	N	O
iz	S	O
obitelji	N	O
P4300	X	B-MISC
koji	P	O
su	V	O
bili	V	O
testirani	A	O
sadrže	V	O
hardverski	A	O

RAID	Y	O
kontroler	N	O
(	Z	O
uobičajeni	A	O
HP-ov	A	B-ORG
SmartArray	N	B-MISC
P410	X	I-MISC
cciss	X	I-MISC
)	Z	O

Tipovi klasa

Četiri klase označene u korpusu su PERS, ORG, LOC i MISC. Pod klasu PERS spadaju osobe, ORG organizacije, LOC lokacije, te MISC ostali nazivi. Ukoliko pojavnica ne pripada nazivu označena je neutralnom oznakom O (engl. *other*). Prema smjernicama s MUC-6 konferencije za ENAMEX grupu[5] označene su tri osnovne klase: PERS, ORG i LOC. Sve što može biti jednoznačno identificirano kao naziv, a ne spada pod prethodne tri klase, niti u NUMEX/TIMEX klase, uključuje dodatna klasa MISC, a smjernice prema kojima je označavana u tekstovima definirane su na CoNLL konferenciji 2003. godine prema (De Meulder, 2003).

Smjernice za označavanje naziva klase PERS:

- nazivi kao „g.“ („gospodin“, engl. „Mr.“) i titule poput „predsjednik“ (engl. „President“) ne smatraju se dijelom naziva, no apozitivi kao „mlađi“ (engl. „Jr.“) smatraju se dijelom imena osobe
- imena obitelji označavaju se oznakom ove klase, npr. „obitelj Kennedy“ (engl. „the Kennedy family“) gdje će biti označena samo riječ „Kennedy“
- razne vrste vlastitih imena se ne označavaju oznakom PERS uključujući političke grupe, zakone nazvane po ljudima, bolesti/nagrade nazvane po ljudima, sveci

Smjernice prema kojima je označavana klasa ORG:

- oznake za korporacije poput engl. „Co.“ smatraju se dijelom naziva organizacije
- razne vrste vlastitih imena koja se označavaju klasom ORG uključuju: sportske timove, burze, multinacionalne organizacije, političke stranke, orkestre, unije, vladajuća tijela bilo koje razine važnosti
- vlastita imena koja se odnose na objekte (npr. tvornice, hoteli, sveučilišta, zračne luke, bolnice, crkve) treba označiti oznakom ORG osim ako ime očito označava građevinu/mjesto, ne organizaciju, ili ako je riječ o imenu organizacije, ali se koristi samo kao referenca na objekt kao strukturu/mjesto; neki navedeni nizovi mogu se označiti opcionalno ovisno o odluci označivača označava li pojavnica organizaciju

[5] http://www.cs.nyu.edu/cs/faculty/grishman/NEtask20.book_6.html

(npr. izraz „Bijela kuća“ nije morao biti označen, ali su takvi primjeri označavani kao ORG)

Smjernice prema kojima su označavani nazivi klase LOC:

- složeni izrazi u kojima su imena mjesta odvojena zarezom označavaju se kao zasebne instance klase LOC, npr. u nizu „Zagreb, Hrvatska“ svaka je riječ posebno označena
- razni nizovi povezani s lokacijama koji se ne označavaju ovom klasom uključuju: adrese ulica, imena ulica, pridjevske oblike imena lokacija (npr. „zagrebački gradonačelnik“)

MISC klasom označavaju se:

- religije, političke ideologije, nacionalnosti, jezici, programi, ratovi, događaji (konferencije, festivali, koncerti itd.), imena vezana uz sport (liga, kup, tablica lige), naslovi (knjiga, pjesama, filmova, albuma, TV programa itd.), slogani, vremenske ere, tipovi (ne brendovi) objekata (tipovi automobila, aviona, motora)

Smjernice su prilagođene prema internom dogovoru autora s obzirom na hrvatski jezik i prema nahođenju kakvi bi trebali biti rezultati dobiveni klasifikatorom. Nazivi osobnih imena i skraćenica pisani kao posvojni pridjevi označavani su kao i njima pripadni osnovni oblici. Npr. u nizu pojavnica „Konzumov honorar dovoljan je“, gdje je „Konzumov“ posvojni pridjev koji označava pripadnost honorara organizaciji „Konzum“, označena je pojavnica „Konzumov“ kao organizacija.

IO oznake

Ideja za označavanjem početaka i krajeva naziva prefiksima krenula je od podjele rečenica na segmente koji nisu rekurzivni i ne preklapaju se (engl. *text chunking*), prema (Ramshaw, Marcus, 1995). Oni u svojem radu navode da je u Brillovom parseru iz 1993. godine uz oznake za vrstu riječi korišten skup dodatnih oznaka {I, O, B} koje su trebale pripomoći sintaksnoj analizi. Dodatne oznake IOB predstavljale su dio unutar imenskog skupa – I (engl. *inside*), dio izvan imenskog skupa, odnosno ono što mu ne pripada – O (engl. *outside*), te njegov početak – B (engl. *begin*). Kasnije se primjena takve podjele našla i u prepoznavanju naziva.

Na CoNLL konferenciji održanoj 2002. godine primijenjen je takav format označavanja pod nazivom IOB2. Sljedeće godine, na CoNLL konferenciji 2003., primijenjen je IOB1 format označavanja. Za IOB1 i IOB2 formate, oznaka B-XXX označava prvu riječ naziva u klasi XXX, I-XXX oznakom označene su riječi koje slijede u nazivu i nalaze se u sklopu njega, dok su riječi izvan naziva označene oznakom O (Curran, Clark, 2003). Razlika IOB1 i IOB2 formata je u tome da se u IOB2 formatu oznakom B označava početak svakog novog naziva, dok se u IOB1 formatu oznaka B koristi samo ako je potrebno izbjeći dvoznačnost; inače se koristi oznaka I čak i na počecima naziva (Siefkes, 2006).

I(nside)/O(utside)/B(egin) format dodatno je proširen novim konvencijama koje uključuju S(tart)/E(nd) oznake. Tjong Kim Sang 1997. uvodi tri verzije označavanja: IOB2, IOE1 i IOE2 (Kudo, Matsumoto, 2001). Osim pojašnjenih IOB1 i IOB2 formata označavanja koriste se prema (Krishnan, Ganapathy, 2005):

- IOE1: oznakom E označava se zadnja pojavnica bloka naziva, kojoj prethodi drugi blok istog naziva,
- IOE2: isti format kao IOE1, osim što se oznaka E pridodaje svakoj pojavnici koja se nalazi na kraju bloka naziva,
- START/END: sastoji se od oznaka B, E, I, S ili O gdje S predstavlja blok koji sadrži samo jednu pojavnicu; blokovi naziva koji imaju dvije ili više od dvije pojavnice uvijek započinju oznakom B i završavaju oznakom E,
- IO: koriste se samo oznake I te O, pa se zato ne mogu razlikovati okolni blokovi nekog naziva.

Stanford NER koji koristimo za prepoznavanje naziva podržava konvencije IO, IOB2, IOB1, IOE1, IOE2, BILOU i SBEIO, koje su navedene unutar njegove klase „LabeledChunkIdentifier“. Primjeri označavanja svakom od njih prikazani su u Tablici 2.[6]

[6] http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/pipeline/LabeledChunkIdentifier.html

Tablica 1: Primjer označavanja IO formatima iz klase "LabeledChunkIdentified" Stanford NER-a

	Bill	works	for	Bank	of	America
IO:	I-PER	O	O	I-ORG	I-ORG	I-ORG
IOB2:	B-PER	O	O	B-ORG	I-ORG	I-ORG
IOB1:	I-PER	O	O	B-ORG	I-ORG	I-ORG
IOE1:	E-PER	O	O	I-ORG	I-ORG	E-ORG
IOE2:	I-PER	O	O	I-ORG	I-ORG	E-ORG
BILOU:	U-PER	O	O	B-ORG	I-ORG	L-ORG
SBEIO:	S-PER	O	O	B-ORG	I-ORG	E-ORG

4.3 Analiza uzorka

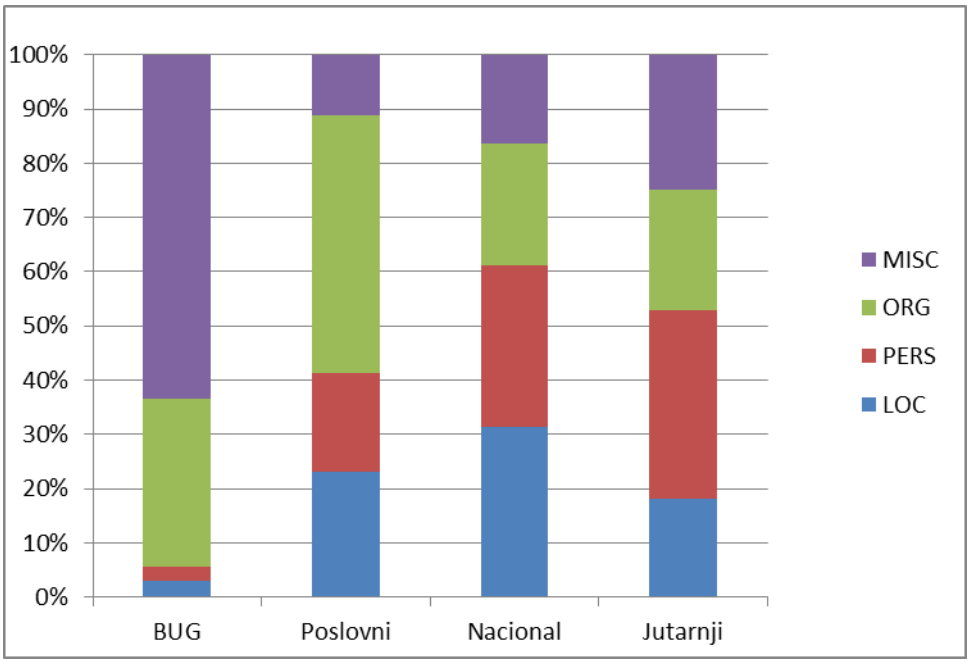
BUG-ova tehnološka domena zastupljena je s 9,608 pojavnica, poslovna domena Poslovnog s 19,256 pojavnica, a domene opće tematike čine polovicu korpusa s 30,279 pojavnica, točnije Nacional s 20,521 i Jutarnji s 9,758 pojavnica. Domene opće tematike tvore polovicu korpusa jer je bilo u cilju napraviti klasifikator koji pokriva što više različitih tematika, a priroda općih domena je da sadrže informativne tekstove s raznih područja. Također, postotak naziva u tekstu veći je pod domenama opće tematike nego pod ostale dvije domene. Nacional prednjači u postotku naziva s 5.06% naziva po tekstu, a slijedi ga Jutarnji s 4.50% (Tablica 3).

Tablica 3: Broj naziva prema domenama, klasama naziva i postotak njihove zastupljenosti u tekstu ovisno o broju pojavnica za svaku domenu

	PERS	**ORG**	**LOC**	**MISC**	**Postotak u tekstu (%)**	**Pojavnica u tekstu**
BUG	11	133	13	271	4,45	9,608
Poslovni	134	353	172	83	3,85	19,256
Nacional	309	234	326	169	5,06	20,521
Jutarnji	152	98	80	109	4,50	9,758
Ukupno	606	818	591	632	4,47	59,143

Tablicom 3 prikazuje se i ukupan broj naziva za svaki tip klase. Najveći broj naziva u označenom uzorku, njih 818, imaju organizacije, dok su osobna imena, lokacije i ostali nazivi podjednako zastupljeni u korpusu te svaki od njih ima oko 600 naziva.

Postotak naziva zastupljen jednom klasom u pojedinoj domeni veći je za domene specifične tematike (Slika 2), vidljivo na primjerima BUG-a i Poslovnog. Tehnološka domena obuhvaća više od 60% naziva klasom MISC, što se da objasniti činjenicom da ta domena pokriva nazive softverskih i hardverskih proizvoda, virusa, crva i sl. Na poslovnoj domeni, čija je klasa ORG najzatupljenija s oko 50% naziva, većinom se piše o tvrtkama i organizacijama. Distribucija naziva prema klasama ravnomjernija je za domene opće tematike.



Slika 1: Postotak zastupljenosti klasa naziva prema domenama

5. Značajke

Kod sustava za prepoznavanje i klasifikaciju naziva koji koriste metode strojnog učenja, važan korak predstavlja odabir i ekstrakcija značajki. Značajke predstavljaju karakteristična obilježja teksta na jezičnoj i izvanjezičnoj razini, te se koriste kako bi se identificirali i klasificirali željeni nazivi u tekstu.

Prema (McDonald, 1996), značajke se dijele na unutarnje i vanjske, gdje su unutarnje značajke one koje se odnose na niz riječi koji sačinjava određeni naziv (npr. prisutnost niza znakova koji ukazuje na organizaciju – „d.d.“ i „d.o.o.“, ili skraćenice koje mogu ukazivati na osobna imena – „g.“, „gđa“ i sl.). Vanjske su značajke one koje se odnose na karakteristična obilježja konteksta unutar kojeg se nazivi pojavljuju – prisutnost takvih obilježja u tekstu u neposrednoj blizini određenog tipa naziva može se iskoristiti za provjeru, odnosno potvrdu naziva.

Nadieau i Sekine (Nadeau, Sekine, 2007) dijele značajke na tri razine – značajke na razini riječi, značajke koje uključuju liste za provjeru (engl. *list lookup*) i značajke na razini dokumenta. Značajke na razini riječi odnose se na karakteristike znakova koji sačinjavaju pojedinu riječ; one opisuju interpunkcijske znakove, brojčane vrijednosti, oznake vrste riječi, morfološka obilježja kao što su prefiksi i sufiksi, je li riječ pisana velikim ili malim slovom i sadrži li posebne znakove. Značajke koje uključuju liste za provjeru mogu ukazati na pripadnost nekog naziva nekoj klasi ukoliko se naziv nalazi unutar određenog popisa. To mogu biti generalni popisi (npr. popisi stop riječi, rječnici, popisi skraćenica i sl.), popisi naziva (npr. za organizacije, nebeska tijela, lokacije) ili popisi riječi koje bi mogle ukazivati na nazive u tekstu (npr. titule). Značajke na razini dokumenta definiraju se na razini sadržaja ili strukture dokumenta; to mogu biti informacije o koreferenciji, poziciji naziva u rečenici, odjeljku ili dokumentu, metainformacije o dokumentu, te frekvencija pojavljivanja riječi i njihovo supojavljivanje.

Unutar nekog modela koji ih koristi, značajke mogu biti predstavljene vektorima - obično za svaku pojedinu riječ u tekstu. Atributi vrijednosti značajki koje mogu biti sadržane unutar vektora mogu biti logički (npr. počinje li riječ prvim velikim slovom), numeričke (npr. duljina riječi u znakovima) i nominalne (npr. oblik riječi pisan malim slovima). Tako bi za rečenicu “The president of Apple eats an apple.”, odgovarajući vektor značajki s obzirom na primjere

atributa vrijednosti bio: <true, 3, “the”>, <false, 9, “president”>, <false, 2, “of”>, <true, 5, “apple”>, <false, 4, “eats”>, <false, 2, “an”>, <false, 5, “apple”>. (Nadeau, Sekine, 2007)

5.1 Najčešće korištene značajke

Općenito, najčešće korištene značajke pri rješavanju problema prepoznavanja naziva u tekstu su značajke na leksičkoj (velika ili mala slova, afiksi i sl.), sintaktičkoj razini (part-of-speech oznake), te značajke popisa.

Ratinov i Roth (Ratinov, Roth, 2009) navode i korištenje nelokalnih značajki koje se šire izvan granica dokumenta kako bi se osigurala veća robusnost sustava. Pristupima tog tipa nastoji se razriješiti problem dodjeljivanja oznaka identičnim pojavnicama koje u različitim kontekstima mogu imati različite oznake klase. Osim toga, značajke koje koriste sastoje se od prethodna dva predviđanja, trenutna riječ, tip riječi (velika i mala slova, brojevi i sl.), prefiksi i sufiksi trenutne riječi, pojavnice u određenom okviru, velika i mala slova unutar određenog okvira oko trenutne riječi, te veza između okvira i predviđanja.

U drugom pristupu prepoznavanja naziva (Krishnan, Ganapathy, 2005), kao glavne značajke koriste se okvir od dvije prethodne i sljedeće riječi, njihovi prefiksi i sufiksi, oznake vrste riječi i oznake klase, dok u pristupu (Finkel, 2007) koriste se značajke riječi kao što su trenutna riječ, prethodna riječ, sljedeća riječ, sve riječi unutar određenog okvira, različite pravopisne značajke, afiksi do određene duljine (ngrami početka i kraja riječi), oznake klase te značajke distribucijske sličnosti.

U pristupu prepoznavanja naziva kreiranjem vrlo velike količine značajki (Mayfield et al, 2003) kao glavne koriste se značajke riječi (u okviru od 7 riječi), trigrami i kvadrigrami, uzorci riječi i brojeva, sadrži li riječ crticu, nalazi li se riječ unutar navodnika, pozicija riječi u rečenici, duljina riječi, POS, dijelovi teksta (engl. *chunks*), leme, je li riječ dio naziva (IO), vjerojatnost da se riječ nalazi unutar određene klase i dodatna usporedba s prijašnjim rezultatima istog označivača.

5.2 Problem odabira značajki

Smatra se kako je u procesu izgradnje dobrog sustava za prepoznavanje naziva odabir značajki jednako važan kao i odabir algoritma za strojno učenje. Pritom se ne koriste sve značajke dostupne u tekstu, već se teži pronalaženju optimalnog skupa značajki (engl. *set of*

features) koji će osigurati najveću točnost sustava uz najmanju moguću kompleksnost izračuna pri izgradnji klasifikatora.

Poznati su pristupi rješavanja problematike prepoznavanja naziva koji koriste vrlo velik broj značajki (Mayfield et al, 2003), no prema (Oliveira et al, 2003) pokazalo se kako uključivanje dodatnih značajki nakon određene točke može čak i dovesti do lošijih rezultata. U slučaju izgradnje jezično neovisnog sustava za prepoznavanje naziva problem odabira značajki još je teži s obzirom da značajke koje se odabiru moraju biti efikasne i generičke (tj. primjenjive na više jezika). Odabir značajki može utjecati na aspekte sustava za prepoznavanje naziva kao što su točnost, vrijeme potrebno za učenje algoritma i optimalna veličina skupa podataka za učenje.

Dva glavna problema odabira skupa značajki prema (Ktoogo, Baryamureeba, 2007) su zalihost (određene značajke mogu biti u uzajamnoj vezi tako da neke od njih nije potrebno uključiti) i međuovisnost značajki (određene značajke zajedno mogu otkriti važne informacije koje zasebno ne bi mogle).

Prema (Ktoogo, Baryamureeba, 2007) pristupi odabira skupa značajki mogu se podijeliti u tri grupe – potpuna, nasumična i heuristička pretraga. Kod potpune pretrage značajke se ručno kreiraju i koriste u cijelosti, bez odabira optimalnog skupa, ili se odbacivanje značajki koje nisu korisne prepusti algoritmu za učenje. Kod nasumične pretrage koriste se probabilistički modeli algoritama ili procesi uzorkovanja (engl. *sampling processes*). Pristupi ovog tipa uključuju iterativno isključivanje značajki s najnižom težinom, izbacivanje značajki na temelju postavljanja praga pojavljivanja u skupu za učenje i sl. Heuristička pretraga uključuje različite pristupe koji kombiniraju tzv. branch-and-bound metode. Najčešće korišteni algoritmi tog tipa koji se koriste za odabir skupa značajki su odabir prema naprijed (engl. *forward selection*) gdje se parametri uključuju u model jedan po jedan ako su statistički značajni, i eliminacija prema unatrag (engl. *backward elimination*) gdje se počinje testiranjem svih kandidata parametara, te se izbacuju oni koji nisu statistički značajni za model.

Neki primjeri pristupa odabira značajki su korištenje algoritma koji procjenjuje točnost i važnost alternativnih skupova značajki te na taj način odabire onaj najbolji (Ktoogo, Baryamureeba, 2007), postavljanje praga na broj pojavljivanja ili težinu određene značajke (Finkel, 2007) i ekstrakciju velikog broja značajki u skupu (u tom slučaju važan faktor predstavlja računalna snaga) (Mayfield et al, 2003).

5.3 Stanford NER značajke

Sustav Stanford NER nudi predefinirane parametre za značajke koje se mogu koristiti i kombinirati u svrhu poboljšavanja rezultata pri izgradnji klasifikatora. Parametri se odnose na ekstrakciju značajki riječi (trenutna riječ, prethodna, sljedeća riječ, sve riječi unutar nekog okvira i sl.), pravopisa (velika slova, brojevi), prefikse i sufikse, sekvence oznaka, popise naziva itd., te ih je moguće i kombinirati i spajati.

U Prilogu 1 navedeni su svi parametri za ekstrakciju značajki (prema dostupnoj dokumentaciji) koje je moguće koristiti, a definirani su u klasi NERFeatureFactory[7] koja sadrži metode za izvođenje svakog parametra. Odabrani parametri za ekstrakciju značajki definiraju se u datoteci za mogućnosti (s ekstenzijom „.prop“), logičkim, numeričkim ili nominalnim vrijednostima, te također imaju postavljene inicijalne vrijednosti. Moguće ih je i naknadno definirati pri pokretanju naredbe za treniranje klasifikatora na samoj komandnoj liniji.

[7] http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/ie/NERFeatureFactory.html

6. Pristupi odabira značajki

Za odabir optimalnog skupa značajki za izgradnju modela za hrvatski jezik u ovom radu koriste se dva pristupa - pristup odozdo prema gore (engl. *bottom-up*) i pohlepna optimizacija (engl. *greedy optimisation*) - gdje na različite načine ispitujemo važnost dostupnih Stanford NER parametara za značajke. S obzirom da nije bilo promjena u kodu sustava s naše strane, ne koristimo značajke koje se tiču isključivo engleskog jezika, kao ni one koje se tiču popisa naziva i lema. U nastavku teksta bit će pobliže objašnjeni pristupi za odabir.

6.1 Pristup odozdo prema gore

U ovom pristupu polazilo se od pretpostavke da su značajke neovisne ili da mogu biti u kombinaciji sa srodnim značajkama. To je tip pristupa odabira prema naprijed. Za ispitivanje tih pretpostavki ručno je pripremljen niz datoteka koje definiraju parametre (ukupno 86 datoteka) s „uključenim“ samo jednim parametrom za ekstrakciju značajki ili povezanom grupom parametara, i te su datoteke pokrenute na poslužitelju nad skupom podataka za učenje kako bi se istražio utjecaj pojedine značajke na krajnji rezultat. Važno je pritom napomenuti kako Stanford NER koristi parametre za značajke „useWord“ i „useSequences“ kao inicijalne aktivne značajke. Za sve ostale parametre kreirane su zasebne datoteke. Parametri koji definiraju značajke najčešće su logičke vrijednosti, dok je za parametre s numeričkim (npr. duljina ngrama) ili nominalnim (npr. oblik riječi) vrijednostima bilo potrebno ispitati više vrijednosti. Petodjelnom unakrsnom provjerom na skupu za učenje pokušalo se dobiti uvid u značajnost parametara te njihovih optimalnih vrijednosti

Datoteka s uključenim samo inicijalnim vrijednostima („useWord“ i „useSequences“) imala je rezultat F1 mjere 0.54. Velik dio parametara za značajke dao je identičan ili čak lošiji rezultat. Zaključak je kako je većina značajki međuovisna, te ih je potrebno kombinirati s ostalima.

Rezultati su također pokazali kako značajke ngrama i riječi imaju očekivano visoke rezultate. Za ngrame, najbolji rezultat (F1 0.723) dao je parametar nominalne vrijednosti maksimalne duljine ngrama 5. Za riječi, kombinacija svih značajki koje se tiču riječi i parova riječi dala je isti rezultat - F1 0.556. Datoteka s uključenim svim značajkama dala je ipak najbolje rezultate (F1 0.74). U načelu su parametri dali uvijek bolji rezultat za preciznost nego odaziv. Iznimka su parametri koji određuju oblik riječi („wordShape“) i oni koji kreiraju značajke već

promatranih sekvenci („useObservedSequencesOnly“). Rezultati svih datoteka dani su u Prilogu 2.

6.2 Pohlepna optimizacija

Pristup pohlepne optimizacije temelji se na lokalnom odabiru optimalnih parametara za ekstrakciju značajki kako bi se dobio jednostavniji model značajki. Kod ovog pristupa datoteka s parametrima definira se dodavanjem one skupine značajki koja nas trenutno lokalno vodi najboljem rezultatu – to je nadgledan proces koji prolazi kroz više iteracija nad skupom za učenje. Parametri za ekstrakciju značajki grupirani su s obzirom na vrstu značajki na koje se odnose (riječ, ngram i sl.). Koraci ovakvog pristupa su:

1. korištenje parametra za značajke za koji je u prethodnom eksperimentu zaključeno kako daje najbolji rezultat – u ovom slučaju, to je grupa parametara koja određuje korištenje ngrama duljine do 5 znakova te susjednih ngrama iste duljine

2. u sljedećem koraku u svim drugim datotekama za definiranje parametara uključuje se ngrams5 grupa parametara za značajke, te se odabire ona koja u tom koraku daje najbolji rezultat

3. ponavljanje koraka 1 i 2 za trenutni parametar za ekstrakciju značajki sve dok se primjećuje napredak

Za navedeni pristup korišteno je 20 grupa parametara za ekstrakciju značajki. Već na petoj iteraciji uočeno je kako uključivanje pojedinih značajki više ne čini zamjetnu razliku u oblikovanju krajnjeg rezultata. Iteracijama su izlučeni parametri za ekstrakciju značajki koji su dali najbolje rezultate. Parametri za ekstrakciju značajki s najboljim rezultatima prve 4 iteracije su useNGrams (F1 0.714), useSequences (F1 0.731), usePrev (F1 0.737) i sigma (F1 0.756). Ovakvom optimizacijom se krajnji rezultat s 0.714 povećao na 0.756, a također je vidljivo da je preciznost u većini slučajeva viša od potpunosti. Rezultati prve četiri iteracije dani su u Prilogu 3.

7. Rezultati

Za provjeru učinkovitosti odabranog skupa parametara za ekstrakciju značajki, cijeli korpus označenih podataka na početku je podijeljen na skup za učenje i testiranje. Skup za testiranje je konstantan, te sačinjava 10 posto podataka iz korpusa koji nisu ulazili u skup za učenje.

Kako bi se pratio napredak u učenju algoritma, skup za učenje uvijek je podijeljen na 20 dijelova, te su na temelju toga načinjene krivulje učenja za sve pristupe na način da se svaka krivulja učenja računala upravo u 20 koraka (dodavanjem slijedećeg dijela skupa za učenje u svakom slijedećem koraku). Svaka podatkovna točka određenog koraka izračunavala se osam puta slučajnim odabirom podskupa podataka, te se kao krajnji rezultat prikazala harmonijska sredina dobivenih vrijednosti. U nastavku teksta bit će prikazani rezultati eksperimenata sa skupovima parametara za ekstrakciju značajki nad cijelim korpusom te potkorpusom općih novinskih tekstova.

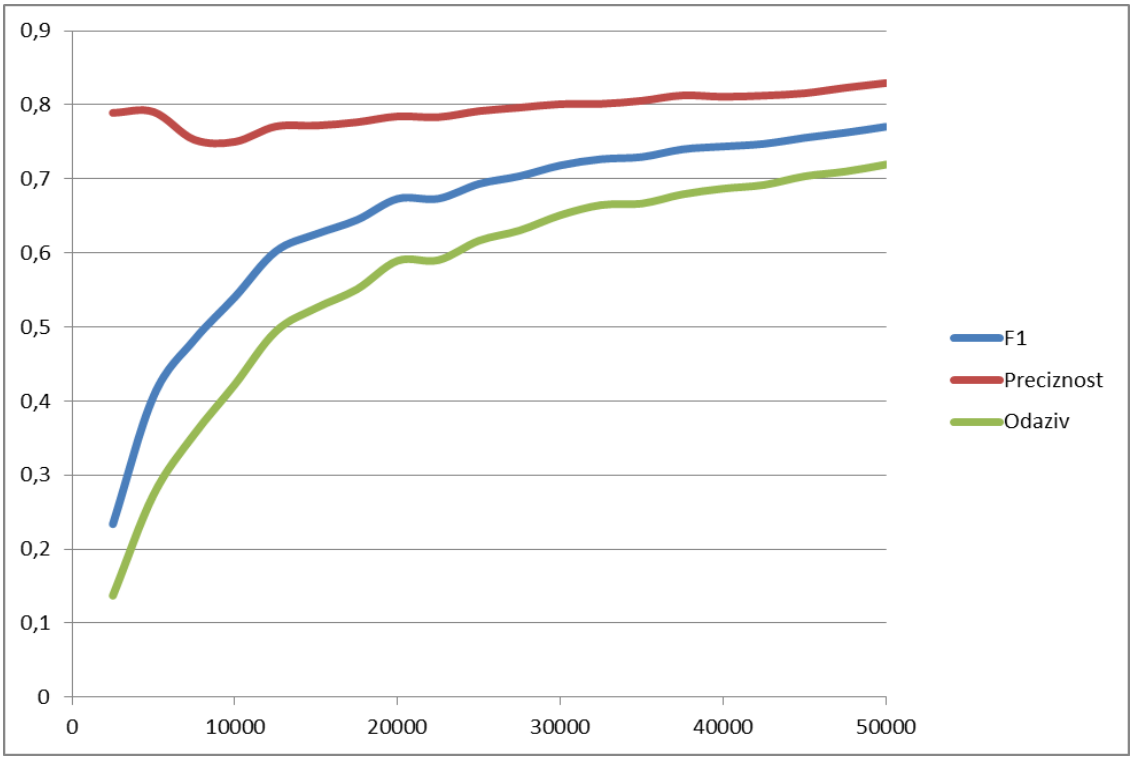
7.1 Cijeli korpus

Skup za učenje cijelog korpusa iznosi 53.377, a skup za testiranje 5.766 pojavnica. Evaluacija je vršena nad skupom značajki dobivenih pohlepnom optimizacijom, te skupom svih značajki. U tablicama 4 i 5 prikazani su rezultati testiranja korištenjem svih parametara za ekstrakciju značajki i korištenjem parametara dobivenih pohlepnom optimizacijom. Općenito, pristup pohlepnom optimizacijom dao je bolje rezultate od pristupa svih značajki na manjim količinama podataka.

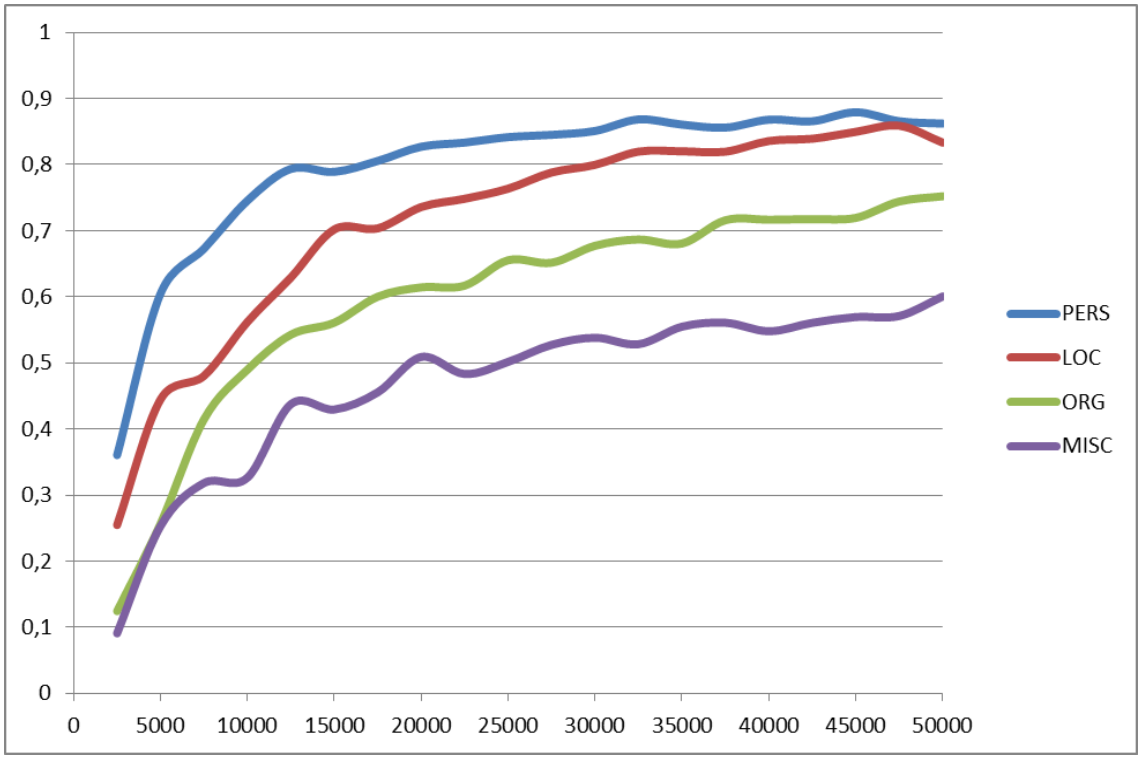
Rezultati evaluacije skupa značajki dobivenih pohlepnom optimizacijom (Slika 3) pokazuju kako se rast krivulje učenja ne prekida, već nastavlja (F1 0,77). Iz toga je vidljivo kako i dalje postoji prostor za napredak u vidu većeg skupa podataka za učenje. Rezultati evaluacije skupa svih značajki (Slika 5) pokazuju istu tendenciju s F1 0,759.

Pohlepna optimizacija bez oznaka vrste riječi dala je nešto niže rezultate, s F1 0,752. Uklanjanjem MISC klase rezultati se penju na F1 0,818, iz čega se da zaključiti kako je ta klasa najteža za učenje, što je vidljivo i iz distribucije po klasama (Slika 4) – najlakše se uče nazivi osoba, sljedeće su lokacije (na većim količinama podataka daju vrlo slične rezultate kao i za osobe), organizacije i na kraju klasa ostalo (MISC). Bez klasifikacije klasa naziva,

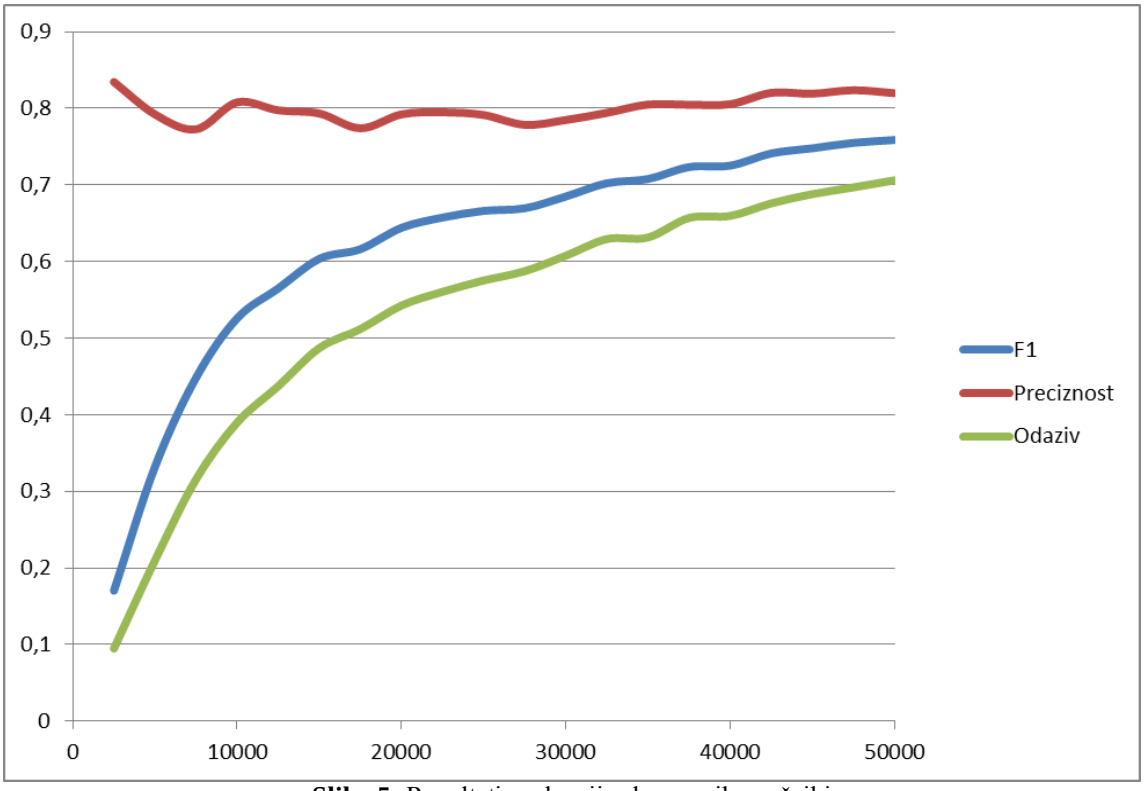
dakle samo njihovim prepoznavanjem, rezultati se penju na F1 0,903, te se pokazalo kako preciznost i odaziv teže prema istom rezultatu.



Slika 3: Rezultati evaluacije skupa značajki dobivenih pohlepnom optimizacijom



Slika 4: Krivulja učenja za pojedine klase u cijelom korpusu



Slika 5: Rezultati evaluacije skupa svih značajki

7.2 Podskup novinskih tekstova

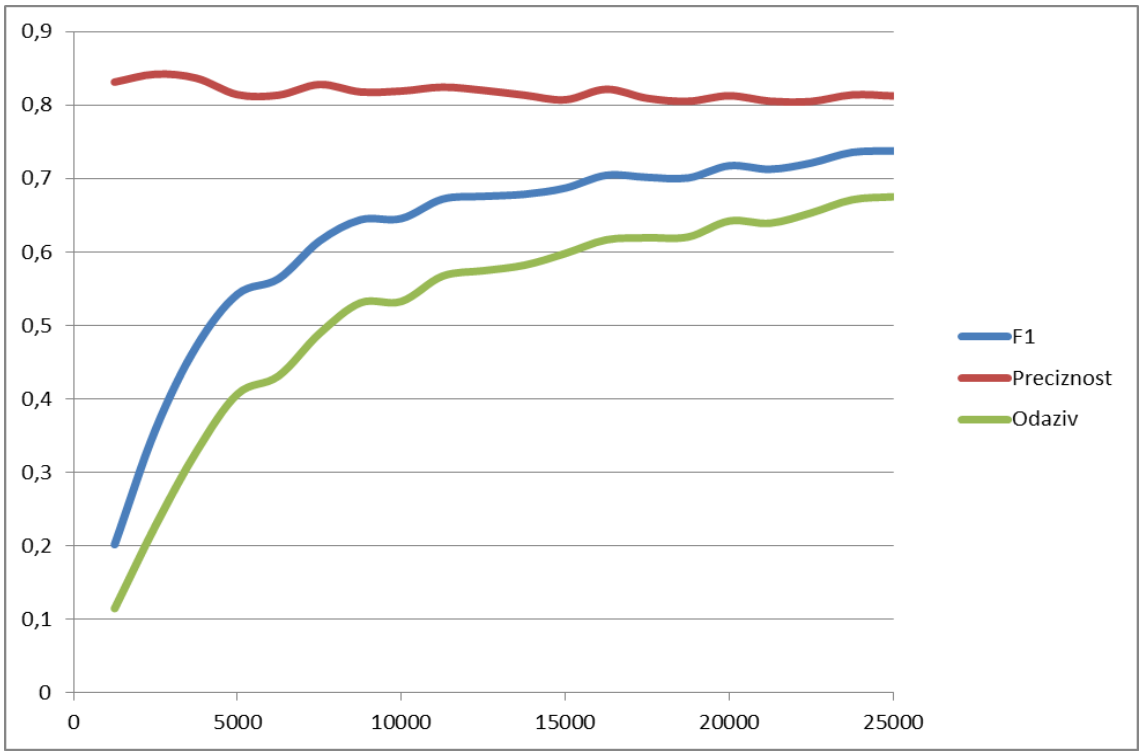
Kako bi se provjerila prilagodljivost odabranog skupa značajki pohlepnom optimizacijom, iz korpusa je stvoren potkorpus iz domene općih novinskih tekstova (Nacional i Jutarnji). Za potrebe evaluacije, taj skup je podijeljen na skup za učenje s 28577 pojavnica i konstantan skup za provjeru od 3210 pojavnica.

Općenito, potkorpus općih novinskih tekstova maloj količini podataka daje rezultate koji su mjerljivi s onima glavnog korpusa (F1 potkorpusa sa značajkama pohlepne optimizacije iznosi 0,738, a kod cijelog korpusa 0,77).

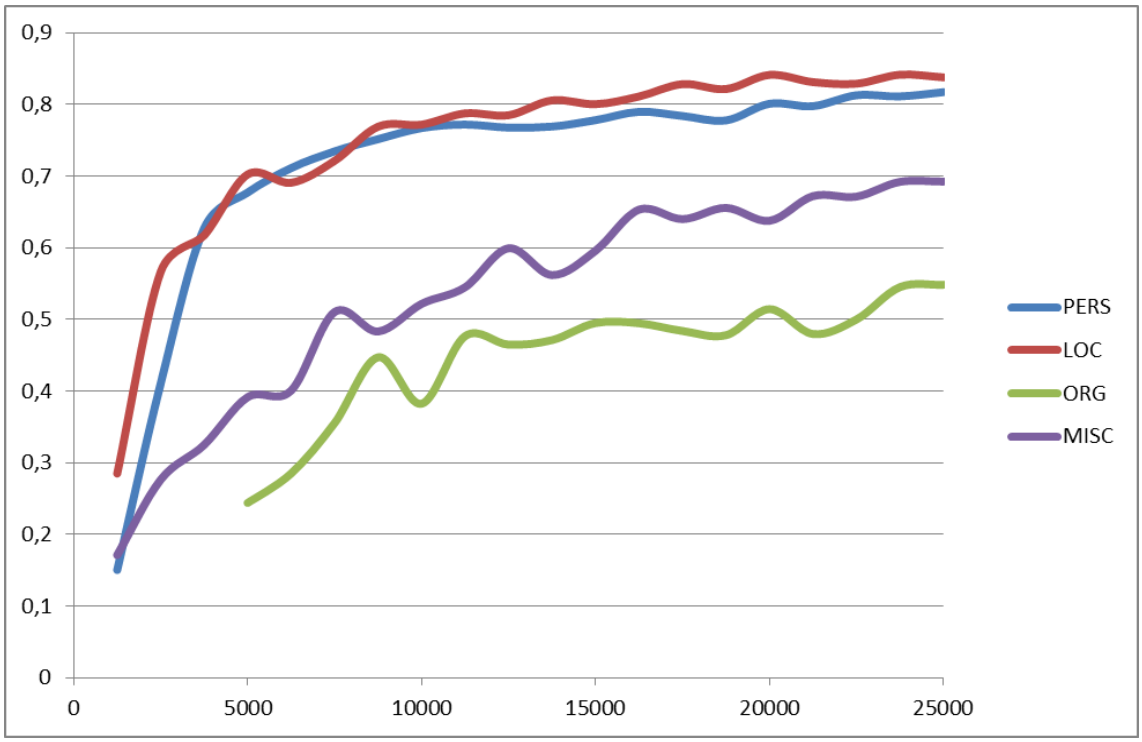
Preciznost je u slučaju uzimanja klasa u obzir (Slika 6) u stagnaciji ili čak vrlo blagom opadanju, s F1 0,738, dok se micanjem klasa naziva blago penje s F1 0,886. Općenito, bez uzimanja klasa naziva u obzir, krivulja učenja je u stanju blagog rasta.

Zanimljiv je obrat u kontinuitetu učenja pojedinih klasa (Slika 7) u usporedbi s cjelokupnim korpusom – u potkorpusu novinskih tekstova najteže se uče organizacije, dok klase koje označavaju osobe i lokacije karakterizira približno jednaka lakoća učenja. Klasa koja označava organizacije nema na svakoj razini podatak iz razloga što u poddijelovima skupa za učenje nije bilo te kategorije.

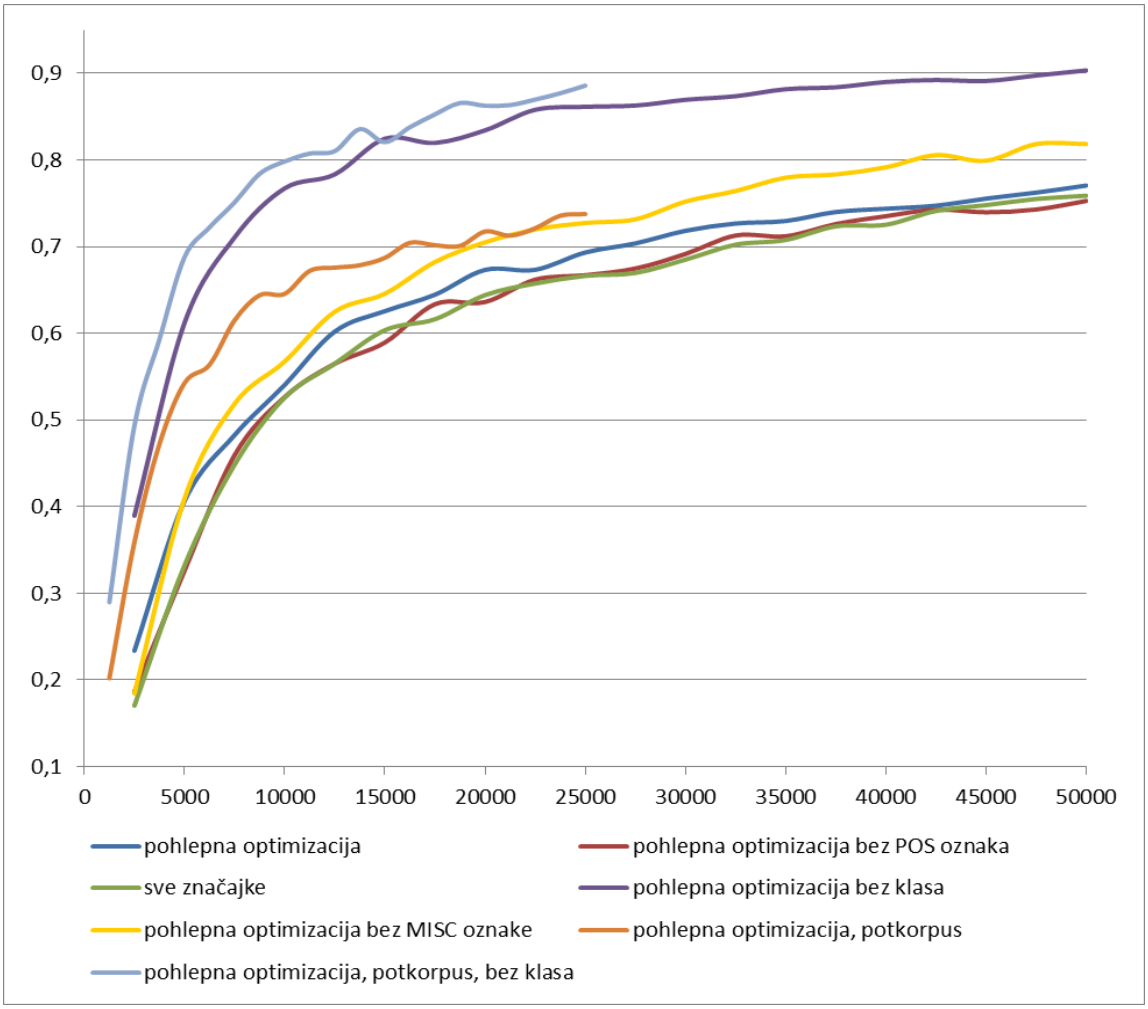
Na slici 8 prikazan je odnos svih navedenih pristupa nad glavnim korpusom i potkorpusom općih novinskih tekstova pomoću F1 mjere. Očekivano je razlika između temeljnog pristupa pohlepnom optimizacijom nad cijelim korpusom i nad potkorpusom veća od one između pristupa pohlepnom optimizacijom bez prepoznavanja klasa nad cijelim korpusom ili potkorpusom.



Slika 6: Rezultati evaluacije učenjem s potkorpusom novinskih tekstova (Nacional i Jutarnji)



Slika 7: Krivulje učenja s F1 vrijednostima za pojedine klase u potkorpusu novinskih tekstova (Nacional i Jutarnji)



Slika 8: Krivulje učenja s F1 vrijednostima za pojedine pristupe – pohlepna optimizacija, pohlepna optimizacija bez oznaka vrste riječi, svi parametri za ekstrakciju značajki, pohlepna optimizacija bez kategorija, pohlepna optimizacija bez klase MISC, pohlepna optimizacija nad potkorpusom novinskih tekstova (Nacional i Jutarnji) te pohlepna optimizacija nad potkorpusom novinskih tekstova (Nacional i Jutarnji) bez kategorija

Tablica 4: Evaluacija korištenjem svih parametara za ekstrakciju značajki

Klasa	P	R	F1	TP	FP	FN
LOC	0,8772	0,8065	0,8403	50	7	12
MISC	0,7692	0,4348	0,5556	30	9	39
ORG	0,7681	0,7571	0,7626	53	16	17
PERS	0,9107	0,9107	0,9107	51	5	5
Totals	0,8326	0,7160	0,7699	184	37	73

Tablica 5: Evaluacija parametara dobivenih pohlepnom optimizacijom

Klasa	P	R	F1	TP	FP	FN
LOC	0,9259	0,8065	0,8621	50	4	12
MISC	0,8049	0,4783	0,6000	33	8	36
ORG	0,7647	0,7429	0,7536	52	16	18
PERS	0,8333	0,8929	0,8621	50	10	6
Totals	0,8296	0,7198	0,7708	185	38	72

8. Zaključak

Prepoznavanje naziva vrlo je zanimljiv problem u domeni jezičnih tehnologija koji se pojavljuje kao dio procesa ekstrakcije informacija. U ovom radu je, uz postavljanje problematike prepoznavanja naziva u odgovarajući kontekst, predstavljen i jednostavan pristup stvaranja sustava za prepoznavanje naziva u tekstovima pisanima na hrvatskom jeziku, i to korištenjem mogućnosti koje pruža alat Stanford NER. Kombinirajući mogućnosti sustava sa skupom podataka na hrvatskom jeziku, pokušalo se dobiti što bolji model za prepoznavanje naziva.

Iz provedenih eksperimenata zaključeno je kako statistički modeli u problemu prepoznavanja naziva funkcioniraju vrlo dobro, te se mogu prilagoditi za druge jezike. Velik problem predstavlja odabir značajki koje će s obzirom na specifičnost jezika dati najbolje rezultate. Pristup pohlepne optimizacije opisan u ovom radu naivan je pristup odabira značajki. Moguća je još veća prilagodba sustava hrvatskom jeziku, na način da se intervenira u sam kod sustava, te da se ostvare značajke specifične hrvatskom jeziku. Trenutnim pristupom koristile su se uglavnom generičke značajke. Vidljivo je također kako krivulja učenja sustava pokazuje tendenciju rasta, tako da se može pretpostaviti kako bi veća količina podataka za učenje dala još bolje rezultate.

Odabirom značajki pohlepnom optimizacijom, te korištenjem svih značajki dostupnih prema Stanford NER dokumentaciji, serijalizirana su dva klasifikatora za prepoznavanje naziva. Veličina klasifikatora dobivenog pohlepnom optimizacijom je 14 Mb, dok je veličina klasifikatora dobivenog korištenjem svih značajki 16 Mb. Brzina označavanja klasifikatora dobivenog pohlepnom optimizacijom je 4528 riječi po sekundi, dok za klasifikator dobiven korištenjem svih značajki iznosi 3910 riječi po sekundi.

Pristupom pohlepne optimizacije dobiven je bolji uvid u funkcije parametara za ekstrakciju značajki, te njihovu međusobnu povezanost, a također je i smanjena veličina modela za 12%, dok je postignuto ubrzanje označavanja za 16%.

Završni model dobiven pristupom pohlepne optimizacije planira se ustupiti istraživačkoj zajednici za slobodno korištenje.

Prilozi

Prilog 1. Popis korištenih Stanford NER mogućnosti za značajke (prema NERFeatureFactory klasi)

Naziv mogućnosti	Inicijalna vrijednost	Opis
cacheNGrams	false	spremanje značajki ngrama za određeni niz znakova kako bi se te vrijednosti mogle ponovno iskoristiti umjesto da se nanovo izračunavaju za isti niz znakova
conjoinShapeNGrams	false	spajanje značajki oblika riječi i ngrama
dehyphenateNGrams	false	micanje crtica iz teksta prije stvaranja značajki ngrama
disjunctionWidth	4	broj riječi sa svake strane trenutne riječi koje se ubrajaju u značajke razdvajanja (disjunction)
dontExtendTaggy	false	za ograničavanje dometa značajki koje nastaju pomoću useTaggySequences kada se povećava maxLeft
entitySubclassification	"IO"	pretvaranje označavanja klasa u neki od alternativnih formata zapisa (IO, IOB1, IOB2, IOE1, IOE2, SBIEO, gdje su S(ingle), B(eginning), E(nding) i I(nside) prepoznati samo pomoću CoNLLDocumentIteratorFactory klase
gazette	null	vrijednost može biti jedna ili više datoteka u kojima su nazivi odvojeni prazninom, zarezom ili točka-zarezom. Svaki redak mora sadržavati klasu naziva odvojenu razmakom od samog naziva (naziv može biti jedna ili više riječi)
intern, intern2	false	omogućava ekstrakciju značajki podnizova klasa i sl.
lowercaseNGrams	false	značajke iz ngrama slova (znakova) gdje su svi pisani malim slovom
maxLeft	2	maksimalni kontekst kojeg koriste značajke klase (potrebno za Viterbijev algoritam)
maxNGramLeng	-1	ngrami iznad navedene vrijednosti neće biti korišteni u izgradnji modela
mergeTags	false	za spajanje B- i I- oznaka.
noMidNGrams	false	izbacivanje ngrama koji ne sadrže početak ili kraj riječi (za dobivanje afiksa)
normalizeTerms	false	normalizacija određenih riječi u engleskom jeziku
normalizeTimex	false	normalizacija velikih i malih slova u nazivima mjeseca i dana
retainEntitySubclassification	false	ostavljanje podtipova oznaka naziva u rezultatu (IOB i slično)
serializeTo	n/a	putanja za serijalizaciju klasifikatora

sloppyGazette	false	značajka za popis naziva pokrene se ako bilo koja pojavnica unutar popisa odgovara promatranoj riječi
strictlyFirstOrder	false	brisanje svih značajki koje ne spadaju pod C i CpC pri izgradnji klasifikatora
testFile	n/a	putanja do datoteke koja sadrži podatke za testiranje
tolerance	1,00e-04	usklađivanje tolerancije pri optimizaciji
trainFile	n/a	putanja do datoteke koja sadrži podatke za učenje
useBeginSent	false	korištenje kombinacije inicijalne pozicije u rečenici i klase (zajedno s oblikom riječi) kao značajke
useBoundarySequences	false	korištenje dodatnih značajki klase sekvenci ukoliko se radi o CoNLL granicama, tako da se naziv može širiti izvan njih
useClassFeature	false	ekstrakcija značajki klase (za granice)
useDisjShape	false	značajke koje označavaju razdvajanje oblika riječi s lijeve ili desne strane riječi s obzirom na disjunctionWidth (zadržava se smjer a ne pozicija)
useDisjunctive	false	uključivanje u značajke razdvajanja riječi s lijeve ili desne strane riječi definiranih prema disjunctionWidth (zadržava se smjer, ali ne pozicija)
useExtraTaggySequences	false	dodavanje sekvenci tagova samo s trenutnim značajkama klase
useGazettes	false	omogućavanje korištenja popisa naziva koji se definiraju dodatnim varijablama (ispod)
useLastRealWord	false	ukoliko je prethodna riječ kraća od tri znaka, dodaje se posebna značajka koja sagledava trenutnu riječ i onu prije prethodne
useLemmaAsWord	false	lema riječi kao značajka
useLemmas	false	lema riječi kao značajka
useLongSequences	false	korištenje sekvence stanja višeg reda kako bi se smanjila duljina maxLefta
useNext	false	ekstrakcija značajki za sljedeću riječ i klasu, te zajdeno s drugim mogućnostima omogućava neke druge sljedeće značajke (npr. sljedeći tag i klasu s useTags)
useNextRealWord	false	ukoliko je sljedeća riječ kraća od tri znaka, dodaje se posebna značajka koja sagledava trenutnu riječ i onu nakon sljedeće
useNextSequences	false	korištenje sljedećih sekvenci
useNGrams	false	značajke iz ngrama slova (znakova)

useOccurrencePatterns	false	za hvatanje višestrukih pojava istog naziva u tekstu s obzirom na velika i mala slova, i to u prozoru od 150 riječi
usePosition	false	značajka kombinacije pozicije u rečenici i klase
usePrev	false	ekstrakcija značajki za prethodnu riječ i klasu, te zajdeno s drugim mogućnostima omogućava neke druge prethodne značajke (npr. prethodni tag i klasu s useTags)
usePrevNextLemmas	false	prethodna/sljedeća lema riječi kao značajka
usePrevSequences	false	korištenje prethodnih sekvenci
useReverse	false	ekstrakcija značajki sekvenci s desna na lijevo (uz sve vezane parametre)
useSequences	true	korištenje sekvenci
useShapeConjunctions	false	sjedinjavanje oblika s tagom ili pozicijom
useSymTags	false	ekstrakcija značajki za prethodni tag, tag, sljedeći tag i klasu; tag, sljedeći tag i klasu; prethodni tag, tag i klasu
useSymWordPairs	false	ekstrakcija zna Z koja nije navedena prema Tadiću, ali ju također gledamo kao pojavnicučajki za prethodnu riječ, sljedeću riječ i klasu
useTaggySequences	false	korištenje značajki sekvenca tagova i klasa prvog, drugog i trećeg reda
useTaggySequencesShapeInteraction	false	dodatno spajanje sekvenci od dva ili tri taga s trenutnim oblikom
useTags	false	ekstrakcija značajki za tag i klasu (s usePrev ili useNext također i prethodni/sljedeći tag i klasu)
useTitle	false	korištenje liste titula (za osobna imena)
useTypeSeqs	false	korištenje značajki oblika riječi osnovnog reda
useTypeSeqs2	false	dodatne značajke prvog i drugog reda za oblik riječi
useTypeSeqs3	false	dodatna značajka oblika sekvence prvog reda
useTypeySequences	false	uzorci oblika riječi prvog reda
useWord	true	značajke za riječ
useWordPairs	false	ekstrakcija značajki za prethodnu riječ, riječ i klasu, te riječ, sljedeću riječ i klasu
useWordTag	false	značajke za parove riječi i tagova (oznaka)
wideDisjunctionWidth	4	broj riječi sa svake strane trenutne riječi koje se ubrajaju u značajke razdvajanja (disjunction)

wordShape	"none"	značajke oblika riječi - vrijednost može biti "none" za nikakav oblik, ili neki od naziva funkcija za prepoznavanje oblika riječi definiranih prema WordShapeClassifier klasi

Prilog 2. Rezultati pristupa odozdo prema gore

F1	Preciznost	Odaziv	Parametar za ekstrakciju značajki
0.740432218685	0.76573462685	0.716739789326	svi parametri
0.723373490587	0.784861731886	0.67079849473	ngrams5
0.713568540468	0.780886131606	0.656933130496	ngram
0.61480951416	0.609508050618	0.620194570267	ngrams4
0.613126274419	0.617846659923	0.608518769394	6gram
0.612387040878	0.605874537211	0.619005146328	4gram
0.611859399609	0.610003173236	0.613715852791	ngrams6
0.611859399609	0.610003173236	0.613715852791	ngrams1
0.610376931561	0.609689753373	0.611049598909	ngrams3
0.605173998421	0.593924539	0.616814399426	3gram
0.591571406667	0.621722159641	0.564167528017	ngrams2
0.591521365692	0.595468704774	0.587608474628	ngrams7
0.565817339539	0.562577944861	0.569141339061	2gram
0.558647448954	0.546286754865	0.571570501133	words
0.556617063703	0.709511509516	0.457966881408	useWordPairs
0.5538440373	0.700046408938	0.458157258646	useWordTag
0.543613909703	0.693576770641	0.446960095207	useReverse
0.540617270349	0.689258299064	0.444708206088	wsdan1useLC
0.540617270349	0.689258299064	0.444708206088	useWord
0.540617270349	0.689258299064	0.444708206088	useTypeSeqs3
0.540617270349	0.689258299064	0.444708206088	useTypeSeqs23
0.540617270349	0.689258299064	0.444708206088	useTypeSeqs2
0.540617270349	0.689258299064	0.444708206088	useTypeSeqs13
0.540617270349	0.689258299064	0.444708206088	useTypeSeqs123
0.540617270349	0.689258299064	0.444708206088	useTypeSeqs12
0.540617270349	0.689258299064	0.444708206088	useTypeSeqs
0.540617270349	0.689258299064	0.444708206088	useTitle
0.540617270349	0.689258299064	0.444708206088	useTaggySequencesShapeInteraction
0.540617270349	0.689258299064	0.444708206088	useTaggySequences
0.540617270349	0.689258299064	0.444708206088	useTaggyExtraTaggySequences
0.540617270349	0.689258299064	0.444708206088	useSymTags
0.540617270349	0.689258299064	0.444708206088	useSum
0.540617270349	0.689258299064	0.444708206088	useSequences
0.540617270349	0.689258299064	0.444708206088	usePrevSequences
0.540617270349	0.689258299064	0.444708206088	useNextSequences
0.540617270349	0.689258299064	0.444708206088	useNB
0.540617270349	0.689258299064	0.444708206088	useExtraTaggySequences

0.540617270349	0.689258299064	0.444708206088	useClassFeatureuseDisjShape
0.540617270349	0.689258299064	0.444708206088	useBinnedLength
0.540617270349	0.689258299064	0.444708206088	taggySequences
0.540617270349	0.689258299064	0.444708206088	strictlyFirstOrder
0.540617270349	0.689258299064	0.444708206088	normalizeTimex
0.540617270349	0.689258299064	0.444708206088	normalizeTermsTimex
0.540617270349	0.689258299064	0.444708206088	normalizeTerms
0.540617270349	0.689258299064	0.444708206088	maxLeftTaggy
0.540617270349	0.689258299064	0.444708206088	maxLeft
0.540617270349	0.689258299064	0.444708206088	intern
0.540617270349	0.689258299064	0.444708206088	flat
0.529029874435	0.673669102977	0.4355475554	wsjenny1
0.528970371636	0.674390675542	0.435143767524	wsdan2
0.521660975972	0.675416234801	0.424941557691	wschris2
0.517215783486	0.478867362756	0.562216860695	wsdan2useLC
0.516105243898	0.477979207052	0.560835544989	wsjenny1useLC
0.51504713526	0.477573142916	0.558835001825	wschris2useLC
0.509898669259	0.674903969268	0.409753530247	useBoundarySequences
0.496116165025	0.654345289352	0.399522684971	useSymWordPairs
0.496116165025	0.654345289352	0.399522684971	symTagsWords
0.485489506922	0.621000055627	0.398494698174	useLastRealWord
0.466605299202	0.710218978156	0.347419898831	wsdan1
0.449254054727	0.46290366351	0.436390429509	useShapeConjunctions
0.440630400485	0.452859978715	0.429035272804	usePosition
0.435854224359	0.582181080481	0.348323893591	useNextRealWord
0.396501808847	0.533504374365	0.315497611853	realWord
0.38999906029	0.566666384822	0.297327271718	usePrev
0.344211575897	0.684796873303	0.229887296561	useTags
0.337748138756	0.658685420638	0.227104593507	useOccurrencePatterns
0.319215023266	0.511826867255	0.23193537316	usePrevNext
0.312150449423	0.484946772265	0.230132674237	useNext
0.30550906643	0.225958637718	0.471530995012	useObservedSequencesOnly
0.269420370604	0.519276065402	0.181885371526	useDisjunctive2
0.264192922622	0.740660796515	0.160781014269	useBeginSent
0.258595991946	0.904470680316	0.150873274121	useTypeySequences
0.258595991946	0.904470680316	0.150873274121	useTypeSeqs123Typey
0.24988341686	0.890218138614	0.145358390998	sequences
0.243841027927	0.887045836067	0.141340082398	useLongSequences
0.243286852826	0.539303591197	0.157076550844	useDisjunctive3
0.213482319266	0.552299738983	0.132299790533	useDisjunctive4
0.195088872228	0.585039256168	0.117084625775	useDisjunctive5
0.195088872228	0.585039256168	0.117084625775	disjunction45
0.195088872228	0.585039256168	0.117084625775	disjunction44
0.195088872228	0.585039256168	0.117084625775	disjunction43
0.195088872228	0.585039256168	0.117084625775	disjunction42
0.18295446797	0.595533935867	0.108093585301	useDisjunctive6
0.171825189936	0.718738205778	0.0975455430828	disjunction5
0.162584533126	0.487214116266	0.0975958909236	disjunction2
0.143514222855	0.64887599049	0.0806885472655	disjunction
0.128696458499	0.594082866523	0.0721700226419	disjunction3

Prilog 3. Rezultati prve četiri iteracije pristupa pohlepne optimizacije

Prva iteracija

F1	Preciznost	Odaziv	Parametar za ekstrakciju značajki
0.713568540468	0.780886131606	0.656933130496	useNGrams
0.543613909703	0.693576770641	0.446960095207	useReverse
0.540617270349	0.689258299064	0.444708206088	wordShape
0.540617270349	0.689258299064	0.444708206088	useTypeSeqs
0.540617270349	0.689258299064	0.444708206088	useTitle
0.540617270349	0.689258299064	0.444708206088	useTaggySequences
0.540617270349	0.689258299064	0.444708206088	useNB
0.540617270349	0.689258299064	0.444708206088	useHuber
0.540617270349	0.689258299064	0.444708206088	normalizeTerms
0.540617270349	0.689258299064	0.444708206088	justify
0.540617270349	0.689258299064	0.444708206088	featureDiffThresh
0.540617270349	0.689258299064	0.444708206088	entitySubclassification
0.496116165025	0.654345289352	0.399522684971	strictlyFirstOrder
0.440630400485	0.452859978715	0.429035272804	usePosition
0.371186718775	0.684576992722	0.254644302359	usePrev
0.31601788272	0.876266932721	0.192764193449	useOccurrencePatterns
0.30550906643	0.225958637718	0.471530995012	useObservedSequencesOnly
0.291908336608	0.700484532954	0.184345008398	useBeginSent
0.24988341686	0.890218138614	0.145358390998	useSequences
0.143514222855	0.64887599049	0.0806885472655	useDisjunctive

Druga iteracija

F1	Preciznost	Odaziv	Parametar za ekstrakciju značajki
0.731410838158	0.81449794861	0.663658859867	useSequences
0.727899118285	0.789920289584	0.674936094081	usePrev
0.714736743655	0.781939189366	0.658163573581	normalizeTerms
0.714475872774	0.781649656233	0.657908833516	useReverse
0.713568540468	0.780886131606	0.656933130496	wordShape
0.713568540468	0.780886131606	0.656933130496	useTypeSeqs
0.713568540468	0.780886131606	0.656933130496	useTitle
0.713568540468	0.780886131606	0.656933130496	useTaggySequences
0.713568540468	0.780886131606	0.656933130496	useNB
0.713568540468	0.780886131606	0.656933130496	useHuber
0.713568540468	0.780886131606	0.656933130496	justify
0.713568540468	0.780886131606	0.656933130496	entitySubclassification
0.712032262894	0.781310566599	0.654060957797	usePosition
0.711490667305	0.778564794615	0.655060602512	featureDiffThresh
0.710999064589	0.760038862417	0.667926387379	sigma
0.710708029313	0.784467115308	0.649618183778	useOccurrencePatterns
0.710534266324	0.773766800751	0.656872217154	useObservedSequencesOnly
0.709259202025	0.778068716477	0.651674000146	strictlyFirstOrder
0.702743439356	0.782064680372	0.638079067586	useBeginSent
0.680441630846	0.775748985726	0.606033600521	useDisjunctive

Treća iteracija

F1	Preciznost	Odaziv	Parametar za ekstrakciju značajki
0.737278748933	0.831176332579	0.662452123236	usePrev
0.733156509371	0.818825925099	0.663713721258	strictlyFirstOrder
0.732750604048	0.817832883791	0.663767606276	usePosition
0.732459355333	0.815574451654	0.664691944605	normalizeTerms
0.731410838158	0.81449794861	0.663658859867	wordShape
0.731410838158	0.81449794861	0.663658859867	useTypeSeqs
0.731410838158	0.81449794861	0.663658859867	useTitle
0.731410838158	0.81449794861	0.663658859867	useTaggySequences
0.731410838158	0.81449794861	0.663658859867	useNB
0.731410838158	0.81449794861	0.663658859867	useHuber
0.731410838158	0.81449794861	0.663658859867	justify
0.731410838158	0.81449794861	0.663658859867	entitySubclassification
0.731348711075	0.813635247614	0.664157830448	useObservedSequencesOnly
0.730713608537	0.815236232487	0.662073615064	featureDiffThresh
0.729991424008	0.812975420367	0.662387738586	useReverse
0.729814753737	0.824175484556	0.65485633758	useBeginSent
0.727000772848	0.786676415974	0.675749840735	sigma
0.722712273343	0.799171745495	0.659622610762	useOccurrencePatterns
0.699682059992	0.806461923214	0.617842677382	useDisjunctive

Četvrta iteracija

F1	Preciznost	Odaziv	Parametar za ekstrakciju značajki
0.756245021674	0.823791691598	0.698943335587	sigma
0.738367593755	0.831438041541	0.664014918327	useObservedSequencesOnly
0.738242735212	0.831905122662	0.663509714071	useReverse
0.737780409777	0.831739925765	0.662929377312	normalizeTerms
0.737640298799	0.832133194615	0.662452123236	useTitle
0.73734516642	0.834737614007	0.660306398132	usePosition
0.737278748933	0.831176332579	0.662452123236	wordShape
0.737278748933	0.831176332579	0.662452123236	useTypeSeqs
0.737278748933	0.831176332579	0.662452123236	useNB
0.737278748933	0.831176332579	0.662452123236	useHuber
0.737278748933	0.831176332579	0.662452123236	justify
0.737278748933	0.831176332579	0.662452123236	entitySubclassification
0.737112527511	0.833971853862	0.6604290207	featureDiffThresh
0.737106735132	0.837824672592	0.657988500779	useBeginSent
0.737093044956	0.821845405335	0.66822257171	useOccurrencePatterns
0.735604222917	0.833920450704	0.658055586212	useTaggySequences
0.734561125338	0.829575066581	0.659100264029	strictlyFirstOrder
0.721275377761	0.835712711696	0.634417467323	useDisjunctive

Prilog 4. Rezultati testiranja značajki po koracima (korak označava broj pojavnica uzetih u podskup za treniranje iz cijelog skupa)

a) Pohlepna optimizacija

Korak	F1	Preciznost	Odaziv
2500	0,23375585	0,78923064	0,13717401
5000	0,4065002	0,79059374	0,27356158
7500	0,48263097	0,75301802	0,35512528
10000	0,54056484	0,75021705	0,42248808
12500	0,60211382	0,77086891	0,49396173
15000	0,6256122	0,77203681	0,5258206
17500	0,64465123	0,77665649	0,55098544
20000	0,67321434	0,78449843	0,58955166
22500	0,6733224	0,78350341	0,59031415
25000	0,69331926	0,79172597	0,61665612
27500	0,70378114	0,79636579	0,63044427
30000	0,71833368	0,80103594	0,65108499
32500	0,72669146	0,80139716	0,66475478
35000	0,72965611	0,8056618	0,66673354
37500	0,74002812	0,8127939	0,67920019
40000	0,74379736	0,81100568	0,68689705
42500	0,74734967	0,81251341	0,69188295
45000	0,75550364	0,81572541	0,70356778
47500	0,76238711	0,82314343	0,7100009
50000	0,77054927	0,82951993	0,71941819

b) Sve značajke

Korak	F1	Preciznost	Odaziv
2500	0,170402986	0,834473437	0,094849289
5000	0,332260673	0,792145306	0,210214833
7500	0,448450409	0,772795828	0,315867918
10000	0,525952642	0,807921494	0,389891986
12500	0,56503741	0,797536807	0,437497049
15000	0,603659671	0,793369177	0,487161436
17500	0,616346055	0,774114663	0,511999715
20000	0,644134747	0,792215397	0,542694645
22500	0,65752721	0,794977762	0,560595947
25000	0,666271518	0,791385037	0,575315402
27500	0,669674154	0,778578494	0,587508968
30000	0,685030178	0,784831648	0,607740277
32500	0,702315311	0,794236116	0,629468693
35000	0,707791119	0,805023401	0,631510748
37500	0,723481267	0,8047555	0,657097032
40000	0,725356977	0,805547037	0,659673853
42500	0,741202099	0,820274967	0,676037787
45000	0,74797639	0,819221935	0,688099515
47500	0,755057252	0,823804995	0,696885772
50000	0,758769387	0,819831288	0,7061962

Popis literature

1. Bekavac, B. Strojno prepoznavanje naziva u suvremenim hrvatskim tekstovima. Doktorska disertacija. Zagreb : Sveučilište u Zagrebu : Filozofski fakultet. 2005.

2. Bošnjak, M. Strojno prepoznavanje naziva tehnikama strojnog učenja. Diplomski rad. Zagreb : Sveučilište u Zagrebu : Fakultet elektrotehnike i računarstva. 2007. Dostupno na: http://bib.irb.hr/datoteka/306732.Diplomski_rad_1679_-_Matko_Bosnjak.pdf (19.4.2012).

3. Curran, J. R. ; Clark, S. Language Independent NER using a Maximum Entropy Tagger. // Proceedings ofthe Seventh Conference on Natural Language Learning (CoNLL-03), Edmonton, 2003. Str. 164–167. Dostupno na: http://acl.ldc.upenn.edu/W/W03/W03-0424.pdf (25.3.2012).

4. De Meulder, F. CONLL-2003 : list of tags with associated categories of names. Dostupno na: http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt (25.3.2012).

5. Enamex : Specific guidelines. // Named Entity Task Definition : Appendix A. Dostupno na: http://www.cs.nyu.edu/cs/faculty/grishman/NEtask20.book_6.html (25.3.2012).

6. Finkel, J. R. Named Entity Recognition and the Stanford NER Software. 2007. Dostupno na: http://nlp.stanford.edu/software/jenny-ner-2007.pdf (25.3.2012).

7. Grishman, R.; Sundheim, B. Message Understanding Conference – 6 : A Brief History. // Proceedings of the International Conference on Computational Linguistics. 1996. Dostupno na: http://acl.ldc.upenn.edu/C/C96/C96-1079.pdf (19.4.2012).

8. Klinger, R. ; Friedrich, C. M. Feature Subset Selection in Conditional Random Fields for Named Entity Recognition. // Proceedings of Recent Advances in Natural Language Processing (RANLP). 2009. Str. 185-191. Dostupno na: http://aclweb.org/anthology-new/R/R09/R09-1035.pdf (13.4.2012).

9. Krishnan, V. ; Ganapathy, V. Named Entity Recognition. 2005. Dostupno na: http://cs229.stanford.edu/proj2005/KrishnanGanapathy-NamedEntityRecognition.pdf (13.4.2012).

10. Ktoogo, F. E. ; Baryamureeba, V. A Methodology for Feature Selection in Named Entity Recognition. // IJCIR 1, 1(2007), str. 88-100. Dostupno na: http://cit.mak.ac.ug/iccir/downloads/SREC_07/Fredrick%20Edward%20Kitoogo%20and%20Venansius%20Baryamureeba_07.pdf (13.4.2012).

11. Kudo, T. ; Matsumoto, Y. Chunking with Support Vector Machines. // Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. 2001. Str. 192-199. Dostupno na: http://acl.ldc.upenn.edu/N/N01/N01-1025.pdf (25.3.2012).

12. Ljubešić, N. ; Erjavec, T. hrWaC and slWaC: Compiling Web Corpora for Croatian and Slovene. // TSD'11 Proceedings of the 14th international conference on Text, speech and dialogue. 2011. Dostupno na: http://www.nljubesic.net/projects/hrWaC_files/ljubesic11-hrwac.pdf (25.3.2012).

13. Mayfield, J. ; McNamee, P. ; Piatko, C. Named Entity Recognition using Hundreds of Thousands of Features. // Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003. 4 (2003). Str. 184-187. Dostupno na: http://acl.ldc.upenn.edu/W/W03/W03-0429.pdf (13.4.2012).

14. McDonald, D. Internal and external evidence in the identification and semantic categorization of proper names. Corpus Processing for Lexical Acquisition. 1996. Str. 21-39.

15. Minkov, Einat; Wang, R.; Cohen, W. Extracting Personal Names from Email : Applying Named Entity Recognition to Informal Text. // Proceedings of the Human Language Technology and Conference Conference on Empirical Methods in Natural Language Processing. 2005. Dostupno na: http://delivery.acm.org/10.1145/1230000/1220631/p443-minkov.pdf?ip=193.198.212.70&acc=OPEN&CFID=100409463&CFTOKEN=72294369&__acm__=1335946591_1720fcca98d2eecbc9469acfbb960dc6 (19.4.2012).

16. Nadeau, D. ; Sekine, S. A survey of named entity recognition and classification. Linguisticae Investigationes 30, 1(2007). Str. 3–26. Dostupno na: http://nlp.cs.nyu.edu/sekine/papers/li07.pdf (13.4.2012).

17. Oliveira, L. S. et al. A Methodology for Feature Selection Using Multi-Objective Genetic Algorithms for Handwritten Digit String Recognition. // International Journal of Pattern Recognition and Artificial Intelligence, 17(2003). Str. 903-929. Dostupno na: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.8488 (13.4.2012).

18. Ramshaw, L. A. ; Marcus, M. P. Text Chunking using Transformation-Based Learning. // Proceedings of the Third Workshop on Very Large Corpora. Cambridge, 1995. Str. 82-94. Dostupno na: http://acl.ldc.upenn.edu/W/W95/W95-0107.pdf (25.3.2012).

19. Ratinov, L. ; Roth, D. Design Challenges and Misconceptions in Named Entity Recognition. // Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL). Boulder, Colorado : 2009. Str 147–155. Dostupno na: http://aclweb.org/anthology-new/W/W09/W09-1119.pdf (13.4.2012).

20. Sekine, S.; Nobata, C. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. // Proceedings of the Conference on Language Resources and Evaluation. 2004. Dostupno na: http://cs.nyu.edu/~sekine/papers/lrec04-65.pdf (19.4.2012).

21. Siefkes, C. A Comparison of Tagging Strategies for Statistical Information Extraction. Proceedings of the Human Language Technology Conference of the North American

Chapter of the ACL. 2006. Str. 149-152. Dostupno na: http://aclweb.org/anthology-new/N/N06/N06-2038.pdf (25.3.2012).

22. Tadić, M. Croatian Specifications. MULTEXT-East Morphosyntactic Specifications, Version 4. 12.5.2010. Dostupno na: http://nl.ijs.si/ME/V4/msd/html/msd-hr.html (13.4.2012).

23. Vitas, D.; Pavlović-Lažetić, G. Resources and Methods for Named Entity Reco gnition in Serbian. INFOTHECA : Journal of Informatics and Librarianship. 9, 1-2(2008). Dostupno na: http://infoteka.bg.ac.rs/PDF/Eng/2008/INFOTHECA_IX_1-2_May2008_35a-42a.pdf (19.4.2012).

Sažetak

Lobel Filipić, Tereza Jurić, Marija Stupar
Strojno prepoznavanje naziva u tekstovima pisanima hrvatskim jezikom

U radu je opisano oblikovanje sustava za prepoznavanje i klasifikaciju naziva teksta pisanog hrvatskim jezikom na temelju alata Stanford NER koji je zasnovan na CRF klasifikatoru i nizu ekstraktora značajki. Podaci za označavanje preuzeti su iz korpusa hrWaC, i to iz tri različite domene: opće, poslovne i tehnološke. Pojavnice u tekstovima su označene vrstom riječi te klasom naziva. Pri označavanju klasa naziva korištene su smjernice zadane prema MUC-7 i CoNLL konferencijama te su razlikovane sljedeće kategorije: osobe, lokacije, organizacije i ostalo. Označeni skup podataka sadrži 59,143 pojavnice. Za odabir relevantnog skupa značajki korištena je metoda pohlepne optimizacije peterostrukom unakrsnom validacijom nad skupom za učenje. Krajnja je evaluacija vršena nad izdvojenim skupom podataka za provjeru te je postignuta F1 mjera od 0.77. Dodatno je istražen utjecaj oznake vrsti riječi, klase te sadržajne domene s obzirom na količinu dostupnih podataka.

Ključne riječi: prepoznavanje naziva, uvjetna nasumična polja, ekstrakcija informacija, hrvatski jezik, hrWaC

Summary

Lobel Filipić, Tereza Jurić, Marija Stupar
Named Entity Recognition in Croatian Texts

This paper describes a system for named entity recognition and classification for Croatian with the Stanford NER tool which is based on the CRF classifier and a series of feature extractors. The data for annotation is taken from the hrWaC corpus from three different domains: general, business and technology. Each token in the text is annotated with its part-of-speech and named entity class. During the annotation of named entity classes the guidelines defined for MUC-7 and CoNLL conferences were followed by annotating the following classes: person, location, organization and miscellaneous. The whole annotated dataset consists of 59,143 tokens. For building the optimal feature set a greedy optimization method was used via five-fold cross-validation on the training set. The final evaluation was performed on the test set yielding a F1 measure of 0.77. The impact of part-of-speech tags, named entity classes and content domains regarding the amount of available training data is presented in the paper as well.

Key words: named entity recognition, conditional random fields, information extraction, croatian language, hrWaC