RAZPOZNAVANJE IMENSKIH ENTITET V SLOVENSKEM BESEDILU

Tadej ŠTAJNER
Institut "Jožef Stefan", Laboratorij za umetno inteligenco
Mednarodna podiplomska šola Jožefa Stefana

Tomaž ERJAVEC
Institut "Jožef Stefan", Odsek za tehnologije znanja
Mednarodna podiplomska šola Jožefa Stefana

Simon KREK
Institut "Jožef Stefan", Laboratorij za umetno inteligenco
Univerza v Ljubljani, Fakulteta za družbene vede

Štajner, T., Erjavec, T., Krek, S. (2013): Razpoznavanje imenskih entitet v slovenskem besedilu. Slovenščina 2.0, 1 (2): 58–81.
URL: http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_04.pdf.

Članek predstavlja algoritem in implementacijo programa za razpoznavanje imen v slovenskem jeziku s pomočjo strojnega učenja. Nadzorovani pristop na osnovi pogojnih naključnih polj je naučen na označenem korpusu ssj500k. V korpusu, ki je prosto dostopen pod licenco Creative Commons CC-BY-NC-SA, so pri besednih pojavnicah poleg oblikoskladenjskih oznak in lem označena tudi imena organizacij, osebna, zemljepisna ter stvarna imena. Članek predstavlja vpliv na natančnost razpoznavanja ob uporabi oblikoskladenjskih oznak, leksikonov in konjunkcij sosednjih lastnosti. Ena od ugotovitev raziskave je, da so oblikoskladenjske oznake pri razpoznavanju entitet koristne. V kombinaciji z vsemi ostalimi lastnostmi doseže sistem na testni množici 74% natančnost in 72% priklic, pri čemer so najbolje razpoznana osebna imena, sledijo jim zemljepisna ter organizacijska in nazadnje stvarna imena. Novo spoznanje članka je tudi to, da lahko z delitvijo razreda vseh stvarnih imen na organizacije in preostala stvarna imena dosežemo boljše rezultate prepoznavanja tudi pri drugih razredih. Preizkusi na neodvisno označenih korpusi kažejo dobro posplošenost modela za osebna in zemljepisna imena. Programska oprema, narejena v raziskavi, je prosto dostopna pod licenco Apache 2.0 na naslovu

http://ailab.ijs.si/~tadej/slner.zip, razvojne različice pa so na voljo na naslovu https://github.com/tadejs/slner.

Ključne besede: prepoznavanje lastnih imen, izločanje entitet, procesiranje naravnega jezika

1 UVOD

Članek opisuje sistem, namenjen razpoznavanju imenskih entitet v slovenskih besedilih. Razpoznavanje pojavnih oblik entitet (v angleščini *entity extraction*, *named entity recognition*, *entity identification*) je pomembna naloga pri izločanju informacij iz besedil, saj besede ali besedne zveze, ki predstavljajo imenske entitete, npr. lastno ime osebe, kraja ali organizacije, k vsebini besedila prispevajo več informacij, kot bi bilo moč razbrati zgolj iz posameznih besed. Razpoznavanje entitet obravnava besedilo na drugem nivoju abstrakcije, ker ne govorimo več o posameznih besedah, temveč (največkrat) o dvo- ali večbesednih zvezah. Med imenske entitete pogosto spadajo tudi datumski ali številski izrazi, četudi tehnično ne predstavljajo imen. V časopisni industriji in založništvu je pogosta praksa, da entitete in ključne besede, ki se pojavijo v člankih, indeksirajo, pogosto še vedno ročno. Nekatere časopisne hiše to počnejo že od 19. stoletja, New York Times denimo od leta 1851 (Sandhaus 2008). Razpoznavanje imen oseb, krajev in stvarnih imen se lahko uporablja tudi za povezovanje zgodb v časopisnih prispevkih (Štajner, Grobelnik 2009), pri čemer uporaba entitet (poleg samega besedila) prispeva k natančnejšemu povezovanju različnih člankov v smiselne verige.

V angleško govorečem delu znanstvene skupnosti je tehnologija razpoznavanja entitet doživela hiter razvoj, v veliki meri kot rezultat serije konferenc *Message Understanding Conference* (Grishman, Sundheim 1996), ki se je odvijala v devetdesetih letih, in konference *TREC* (Balog in dr. 2010), ki se v okviru sistemov za priklic informacij odvija še dandanes. V okviru obeh konferenc so bila organizirana odprta tekmovanja v raznih nalogah iz obdelave naravnega jezika, pri čemer je bilo veliko nalog osredotočenih na

razpoznavanje entitet. Najzmogljivejši sistemi uporabljajo predvsem postopke strojnega učenja, natančneje: modele na probablističnih grafih, kot so npr. skriti markovski modeli – *Hidden Markov Models* (Rabiner 1986), ali pogojna naključna polja – *Conditional Random Fields* (Lafferty in dr. 2001), npr. Mallet (McCallum 2002) ali Stanford NER (Finkel in dr. 2005). V praksi so ti sistemi implementirani z nadzorovanim učenjem na besedilu, pri katerem so entitete že označene. V procesu učenja se za vsako besedo generirajo posamezne lastnosti, kot npr. oblikoskladenjske oznake, velike začetnice, prisotnost pomišljaja in podobno, v procesu označevanja pa sistem uporabi model, zgrajen na osnovi teh lastnosti.

Nekateri sistemi uporabljajo tudi eksplicitno predznanje, njihova slabost pa je ta, da ne zaznajo neznanih entitet, če jih nimajo v obstoječem leksikonu. Zato se jih pogosto kombinira s sistemom, osnovanem na strojnem učenju, tako da oba skupaj tvorita hibridni sistem (Cohen, Sarawagi 2004). Nekateri sistemi uporabljajo tudi nenadzorovano izločanje entitet, saj ta pristop ne zahteva vnaprejšnjega učenja (Etzioni in dr. 2005).

Obstoječi prepoznavalniki lastnih imen za slovenščino (Štajner in dr. 2012) zmorejo prepoznavati osebna in zemljepisna imena z visoko natančnostjo ter stvarna imena z nekoliko nižjo natančnostjo. Ta članek nadgrajuje predhodno delo s preverjanjem dodatne hipoteze, ki pravi, da lahko razdelitev razreda stvarnih imen na organizacije ter druga stvarna imena izboljša natančnost prepoznavanja.

Pristopi, ki temeljijo na pogojnih naključnih poljih, so trenutno najširše uporabljeni sistemi za tovrstno učenje, predvsem tisti, osnovani na modelu Stanford NER (Finkel in dr. 2005). Najnovejši preboji na tem področju so predvsem v uporabi večjezičnih korpusov za učenje (Che in dr. 2013). Pristopi s pomočjo medjezičnega učenja omogočajo učenje sistema s pomočjo vzporednih korpusov brez učnih podatkov v ciljnem jeziku (Munro, Manning 2012). Poudarek je tudi na zaznavanju imenskih entitet v neformalnem jeziku,

kot npr. v družabnih medijih (Jung 2012).

Prispevek tega članka je predvsem v prepoznavanju značilnosti slovenskega jezika, ki so koristne za prepoznavanje imenskih entitet. V primerjavi s sorodnimi pristopi v angleškem jeziku je v slovenščini mnogo večji poudarek na uporabi oblikoskladenjskih oznak. Optimalen model je na voljo tudi kot odprtokodni projekt.

V nadaljevanju ima razprava naslednjo strukturo: v 2. razdelku predstavljamo korpus, na katerem je bil sistem naučen in testiran, v 3. razdelku opisujemo razviti razpoznavalnik, v 4. razdelku poskuse, ki smo jih izvedli, nato pa sledijo zaključki.

2 UČNI KORPUS SSJ500K

Za nadzorovano učenje je potreben korpus, v katerem so pojavitve lastnih imen ustrezno označene. Za slovenski jezik do sedaj še nismo imeli tako označenega korpusa, vendar je bil pred kratkim izdelan ročno označeni učni korpus ssj500k, ki je nastal v okviru projekta Sporazumevanje v slovenskem jeziku (SSJ) in temelji na učnih korpusih jos100k in jos1M, izdelanih v okviru projekta JOS (Erjavec in dr. 2010a; Erjavec in dr. 2010b).

Korpus ssj500k sestavljata dva dela: celotni korpus jos100k in dodatnih 400.000 besed iz enomilijonskega korpusa jos1M. Vse jezikoslovne oznake (oblikoskladenjske oznake, leme, skladnja) so bile v korpusu ssj500k še enkrat ročno pregledane, skladenjsko razčlenjeni del pa je bil povečan na 11.411 stavkov. V celoti je bila ročno pregledana in popravljena tudi stavčna segmentacija in tokenizacija, kar omogoča preverjanje uspešnosti označevalnikov ter razčlenjevalnikov tudi pri teh dveh postopkih. Učni korpus ssj500k je prosto dostopen pod licenco "Creative Commons Priznanje avtorstva-Nekomercialno-Deljenje pod enakimi pogoji 2.5 Slovenija" (CC-BY-

NC-SA) na spletnih straneh projekta SSJ.[1]

V delu, ki vsebuje podatke iz korpusa jos100k, so bile dodane tudi informacije o lastnih imenih za potrebe strojnih razpoznavalnikov imenskih entitet. Ta del zajema petino celotnega korpusa ssj500k, podatki zgolj za ta podkorpus (ssj100k) so podani v Tabeli 1.

Elementov	***n***
Besedil	248
Odstavkov	1.599
Stavkov oz. povedi	5.808
Besed	100.135
Ločil in simbolov	18.499
Skladenjsko označenih stavkov	5.808
Skladenjskih povezav	118.635
Stavkov z imenskimi entitetami	2.177
Imenskih entitet	4.397

Tabela 1: Število elementov v podkorpusu ssj500k, označenem s podatki o imenskih entitetah oz. lastnih imenih.

Lastna imena v korpusu so razdeljena v štiri razrede: osebna imena (1.922), zemljepisna imena (1.284), imena organizacij (785) in stvarna imena (406). Lastna imena vsebuje 2.177 oz. 37,48 % vseh stavkov, pri čemer je distribucija lastnih imen po teh stavkih razmeroma neenakomerna. Več kot polovico jih vsebuje eno lastno ime, četrtina dve, desetina tri, temu sledi dolg "rep" do stavka s kar 47 lastnimi imeni.

[1] Domača stran projekta SSJ je http://www.slovenscina.eu/, korpusi in leksikoni projekta pa so dostopni tudi na naslovu http://nl.ijs.si/ssj/, kjer se nahajajo tudi v izvedenih formatih.

3 IMPLEMENTACIJA

V skladu s trenutno prakso obstoječih sistemov za druge jezike implementacija sistema, predstavljenega v članku, uporablja nadzorovano učenje s pogojnimi naključnimi polji (*Conditional Random Fields*), ali krajše CRF, ki temelji na sistemu Mallet (McCallum 2002). Nadzorovano učenje predpostavlja, da so na voljo označeni podatki, iz katerih se statistični model nauči lastnosti besed glede na njihove oznake. Tako naučen sistem je s tem modelom zmožen označevati novo, neoznačeno besedilo.

3.1 Model

Pogost pristop pri modeliranju zaznavanja imenskih entitet pri sodobnih pristopih k analizi imenskih entitet je verižni model, pri katerem besede označujemo zaporedno, pri vsaki odločitvi pa med drugim upoštevamo tudi oznako iz prejšnjega koraka.

V takšnem modelu, kot je npr. sekvenčni CRF, so stanja določena z željenimi oznakami, ki predstavljajo tipe entitet. Množica stanj modela je torej *{osebno, zemljepisno, organizacija, stvarno, brez}*. Naj bo stavek predstavljen kot zaporedje besed. V postopku označevanja vsaki besedi priredimo oznako najverjetnejšega stanja glede na oznako prejšnje besede ter glede na lastnosti trenutne besede. Z drugimi besedami, tak model ima lastnost, da je trenutno stanje odvisno le od lokalnih lastnosti trenutne besede in od razreda predhodne besede.

Model CRF lahko predstavimo kot graf, v katerem stanja predstavlja množica vseh razredov entitet, vključno s prazno vrednostjo. Ker lahko vsako besedo opišemo z vektorjem lastnosti *x*, lahko definiramo verjetnostno porazdelitev *P(X)*, ki opisuje verjetnostno porazdelitev pojavitev lastnosti. Ker pa beseda spada v enega izmed razredov entitet, lahko definiramo tudi verjetnostno porazdelitev razredov *Y*, označeno s *P(Y)*. Z drugimi besedami, učenje modela je osnovano na tem, da imajo različni razredi entitet različne porazdelitve lastnosti, npr. samostalniki v sklonu mestnika so z višjo verjetnostjo v razredu

zemljepisnih imen kot pa v razredu osebnih imen.

Iz teh definicij sledi, da je *(X,Y)* pogojno naključno polje, če ima *Y* markovsko lastnost glede na sosednost besed. Naj bosta w in v besedi. Potem mora veljati $p(Y_v|X,Y_w,\ w \neq v) = p(Y_v|X,Y_w,\ w \approx v)$, kjer $w \approx v$ pomeni, da sta w in v soseda. Ta omejitev ima za posledico, da lastnosti besede, ki se nahaja dve mesti zadaj, na trenutno označevanje nimajo neposrednega vpliva.

Pogojna verjetnost med *X* in *Y* je tako opisana z množico funkcij lastnosti oblike $f_k(y,y',x_t)$. Npr. $f_{upper\text{-}person}$ je lahko tovrstna funkcija, ki vrne 1 v primeru, ko se trenutna beseda začne z veliko začetnico in ko je predhodna beseda označena kot osebno ime, sicer pa 0. Linearno verižno pogojno naključno polje (*Linear chain CRF*) je verjetnostna porazdelitev $p(y|x)$, ki jo opišemo z množico numeričnih parametrov Λ, ki predstavljajo uteži posameznih lastnosti glede na prejšnjo in trenutno oznako besede.

Za uporabo modela je nato potrebno oceniti vrednosti parametrov Λ, ki nam povedo, v kolikšni meri je določena lastnost povezana z določenim ciljnim razredom. Algoritem za učenje s pomočjo optimizacijskega algoritma na vrednostih Λ maksimira verjetnost $P(Y|X;\Lambda)$, tako da je končna Λ takšna, da je pogojna verjetnost največja. Da se prepreči prekomerno prilagajanje učnim podatkom zaradi prevelikih uteži posameznih lastnosti, se v optimizacijski funkciji uporabi tudi regularizacija parametrov Λ. V ta namen se tipično uporablja maksimizacija regularizirane pogojne logaritemske verjetnosti (*conditional log-likelihood*), izračunane iz množice učnih primerov. Ker pa $l(\Lambda)$ ni moč maksimirati v zaprti obliki, se v ta namen uporablja numerična optimizacija s pomočjo delnih odvodov. Za rešitev optimizacije ocenjevanja parametrov uporabimo optimizacijski algoritem L-BFGS (Byrd in dr. 1994). Ko se naučimo parametrov modela, jih lahko uporabimo za označevanje neoznačenega besedila. Za to uporabimo inferenčni algoritem *loopy belief propagation* (Sutton in dr. 2004).

3.2 Lastnosti besed

Pri implementaciji pristopa za razpoznavanje entitet je ključno, da si lahko pomagamo s čim bolj raznolikimi tipi informacij. V ta namen uvedemo štiri kategorije lastnosti, pri katerih vsaka kategorija prinaša dodatno informacijo, kar prikazujemo s poskusi v naslednjem razdelku.

3.2.1 LASTNOSTI ČRKOVNIH VZORCEV

Lastnosti črkovnih vzorcev	**Primer**
Velika začetnica	*Ljubljana*
Le velike črke v celotni besedi	*IJS*
Mešane velike črke znotraj besede	*iPod*
Številke v besedi	*ZVCP-1*
Le številke v besedi	*2012*
Numerični izraz	*+3.14*
Alfanumerični izraz, ki vsebuje le številke in črke	*E3*
Rimska številka	*XVIII*
Vsebuje vezaj ali pomišljaj	*Šmarje-Sap*
Le velike črke, lahko ločene s piko	*I.M.V.*
Inicialka, tj. posamezna velika črka, ki ji sledi pika	John *F.* Kennedy
Posamezna črka ne glede na velike ali male črke	odgovor *a.*
Posamezna velika črka	Plan *B*
Ločilo	*!*
Narekovaj	``
Le male črke	*pisarna*

Tabela 2: Črkovni vzorci in primeri pojavnic, ki jim ustrezajo.

Obstoječi pristopi pri razpoznavanju entitet največkrat vključujejo tipične lastnosti grafemskih vzorcev, pri katerih lahko vsako posamezno besedo opišemo z binarno vrednostjo prisotnosti te lastnosti v vzorcu. S pomočjo regularnih izrazov smo določili lastnosti, ki se že uporabljajo pri zaznavanju imenskih entitet. Vsaka lastnost dobi vrednost 1 le, če beseda ustreza regularnemu izrazu. To množico lastnosti vzamemo kot osnovo, ko ji nato dodajamo ostale razrede.

3.2.2 LASTNOSTI IZ ZUNANJIH VIROV ZNANJA

Poleg črkovnih vzorcev lahko uporabljamo tudi zunanje znanje v obliki leksikonov, ki vsebujejo že znana lastna imena. S tem pristopom v model vključimo znanje, ki bi ga bilo le z nadzorovanim učenjem težko nadomestiti. V ta namen definiramo leksikonsko lastnost, ki dobi vrednost 1 le, če je lema besede vsebovana v določenem leksikonu, pri čemer imamo za vsak leksikon po eno binarno lastnost. Uporabimo prisotnost leme, saj bi bilo pripadnost posameznemu leksikonu pri besedni obliki zaradi bogate slovenske pregibnosti težko preverjati. Večino leksikonov smo vzeli iz slovenske različice Wikipedije, ki je prosto dostopen vir in obsega dovolj široko paleto tematskih domen za splošno razpoznavanje entitet. Uporabljajo se naslednji leksikoni:

- kraji v Sloveniji iz slovenske Wikipedije (Wikipedia 2012e)
- države iz slovenske Wikipedije (Wikipedia 2012f)
- kraji v tujini iz slovenske Wikipedije (Wikipedia 2012b)
- občine v Sloveniji iz slovenske Wikipedije (Wikipedia 2012č)
- tipične besede v lokacijah (npr. vas, mesto, trg, gora)
- tipične besede v organizacijah (npr. institut, ministrstvo)
- tipične predpone in pripone osebnih imen (npr. dr, mag, ml)
- seznam imen iz Statističnega urada Republike Slovenije (Statistični urad Republike Slovenije 2012)
- seznam moških imen iz slovenske Wikipedije (Wikipedia 2012c)
- seznam ženskih imen iz slovenske Wikipedije (Wikipedia 2012a)
- seznam priimkov iz slovenske Wikipedije (Wikipedia 2012d)

- imena dni v tednu
- imena mesecev

Ker leksikoni besed vsebujejo le besede v lematizirani obliki, je treba vsako učno in testno množico lematizirati, da lahko zanesljivo zaznavamo lastnosti na podlagi leksikonov. V tem primeru uporabimo lematizator LemmaGen (Juršič in dr. 2007), ki je že vključen v oblikoskladenjski označevalnik. Ta korak za korpus ssj500k ni potreben, saj že vsebuje leme in oblikoskladenjske oznake.

3.2.3 OBLIKOSKLADENJSKE LASTNOSTI

Tretji potencialni vir informacij za razpoznavanje entitet so oblikoskladenjske oznake besed, ki jih prevedemo v lastnosti po tabeli, ki je del oblikoskladenjskih specifikacij za slovenski jezik JOS (Erjavec in dr. 2010a). Npr. beseda *narediti* s slovensko oznako *Ggdn* oz. angleško *Vmen* dobi lastnosti *Category=verb, Type=main, Aspect=perfective, VForm=infinitive*, beseda *predsednik* z oblikoskladenjsko oznako *Sometd* oz. *Ncmsay* pa lastnosti *Category=noun, Type=common, Gender=masculine, Number=singular, Case=accusative, Animate=yes*. Če besedilo ni označeno, lahko uporabimo ustrezni oblikoskladenjski označevalnik (Rupnik in dr. 2008). Uporaba oblikoskladenjskih oznak temelji na predpostavki, da iz vzorcev oznak lahko razberemo prisotnost entitet. Uporaba mestnika v kombinaciji z veliko začetnico lahko denimo nakazuje prisotnost zemljepisnega imena.

3.2.4 STRUKTURNE LASTNOSTI

Poleg regularnih izrazov, leksikonov in oblikoskladenjskih oznak lahko uporabimo tudi različne strukturne lastnosti, ki izvirajo iz zgradbe stavka kot zaporedja besed. Prva množica strukturnih lastnosti izvira iz dolžine besede, ki jo razbijemo v razrede dolžin 1, 2, 3 ali 4, od 5 do 9 ali več kot 10 znakov, sama lastnost pa je odvisna od pripadnosti tem razredom (npr. *Length=5_9*

pri besedi z dolžino 7 znakov).

Druga množica strukturnih lastnosti, konjunkcija sosednjih lastnosti, je definirana kot preslikava nad obstoječimi lastnostmi. To je metoda za generiranje dodatnih lastnosti, ki za vsako besedo sestavi nove lastnosti kot kombinacije lastnosti njenih sosedov znotraj določenega okna. Uporablja se predvsem v tistih verižnih klasifikatorjih, pri katerih soodvisnosti med lastnostmi in razredi niso odvisne le od prejšnjega ter trenutnega stanja, ampak tudi od širše okolice, kar je lahko še posebej poudarjeno pri jezikih s prostim besednim redom. Ker je eksplicitno modeliranje soodvisnosti višjega reda računsko zelo zahtevno, uporabimo konjunkcije sosednjih lastnosti kot približek.

Npr., če se trenutna beseda nahaja dve mesti za besedo z veliko začetnico, dobi lastnost, ki jo opišemo z "velika začetnica na mestu -2" oz. krajše $f_{upper@-2}$. V nadaljnjih poskusih obravnavamo tri možne razpone vzorcev: le predhodna in naslednja *((-1),(1))*, vse možne kombinacije parov predhodnika, trenutnega in naslednika *((-1,0),(-1,1),(0,1)),* tretji razpon pa predstavlja vse možne kombinacije parov lastnosti v razponu dve mesti naprej ter nazaj. Tako npr. kombinacija *(-2,1)* predstavlja konjunkcijo značilk besede dve mesti pred trenutno z lastnostmi naslednje besede. Tovrstno generiranje lastnosti izredno poveča število možnih lastnosti in s tem upočasni učenje ter povečuje nevarnost prekomernega prilagajanja.

4 POSKUSI

S poskusi smo želeli odgovoriti na vprašanja o smiselnosti uporabe različnih razredov lastnosti glede na meritve:

- Ali oblikoskladenjske oznake izboljšajo model?
- Ali uporaba leksikonov izboljša model?
- Ali kombinacije parov lastnosti v soseščini izboljšajo model?
- Ali delitev razreda stvarnih imen na organizacije ter druga stvarna imena izboljša model?

- Ali se model uspešno posploši na druge testne podatke?

4.1 Poskusi na ssj500k

Poskuse na ssj500k smo izvedli z desetkratnim navzkrižnim preverjanjem, pri katerem naključnih 90 % podatkov uporabimo za učenje, preostale pa za testiranje. Kakovost rezultata merimo z več metrikami: natančnostjo, ki nam pove, koliko od dobljenih entitet je pravilnih, priklicem, ki nam pove, koliko znanih entitet smo identificirali, ter F_1, ki je geometrijsko povprečje natančnosti in priklica.

Zaradi preglednosti obravnavamo vsako hipotezo posebej. Vsak nadaljnji poskus kot osnovo uporablja različico predhodnega poskusa, ki je imela v tistem krogu najboljši izid.

Tip entitete	**Natančnost**	**Priklic**	**F_1**
Brez oblikoskladenjskih oznak			
Osebno	0,73	0,80	0,76
Zemljepisno	0,66	0,64	0,65
Organizacija	0,57	0,44	0,49
Stvarno	0,27	0,20	0,23
Skupaj	0,63	0,59	**0,61**
Z oblikoskladenjskimi oznakami			
Osebno	0,80	0,88	0,84
Zemljepisno	0,77	0,75	0,76
Organizacija	0,61	0,53	0,56
Stvarno	0,41	0,25	0,30
Skupaj	0,72	0,68	**0,70**

Tabela 3: Rezultati poskusov glede na uporabljene oblikoskladenjske oznake.

Tabela 3: Rezultati poskusov glede na uporabljene oblikoskladenjske oznake. kaže, da so oblikoskladenjske oznake pri razpoznavanju entitet izjemno koristne, saj sta tako priklic kot tudi natančnost pri vseh meritvah statistično značilno višja kot brez uporabe oznak. Statistična značilnost je ugotovljena s pomočjo uporabe T-testa pri p-vrednosti, manjši od 0.05. Poskusi tudi kažejo, da je sistem razmeroma uspešen pri razpoznavanju osebnih imen, nekaj slabši pri zemljepisnih imenih, čemur sledijo imena organizacij. Pri prepoznavanju stvarnih imen je razmeroma neuspešen, vendar oblikoskladenjske oznake stanje opazno izboljšajo. Zemljepisna imena je lažje razpoznati, ker gre za samostalnike z veliko začetnico, ki so tipično v mestniku, organizacijska imena pa pogosto sestavljajo zveze s pridevniki (*Evropska komisija*) ali predložnimi zvezami (*Ministrstvo za obrambo*) in z različnimi skloni znotraj besedne zveze. Pri stvarnih imenih je možnih variacij preveč, da bi bile zajete v obstoječih učnih podatkih. V primerjavi s predhodnim modelom za prepoznavanje entitet, ki še ni ločeval organizacij in stvarnih imen (Štajner in dr. 2012), so rezultati za vse razrede boljši že na osnovni množici lastnosti, kar nakazuje boljšo informativnost ob delitvi na štiri razrede. V nadaljnjih poskusih privzemamo, da so oblikoskladenjske oznake vedno prisotne.

Tip entitete	**Natančnost**	**Priklic**	**F_1**
Z uporabo leksikonov			
Osebno	0,83	0,89	0,86
Zemljepisno	0,79	0,77	0,78
Organizacija	0,62	0,55	0,58
Stvarno	0,41	0,26	0,31
Skupaj	0,73	0,69	**0,71**

Tabela 4: Rezultati poskusov glede na uporabljene leksikone.

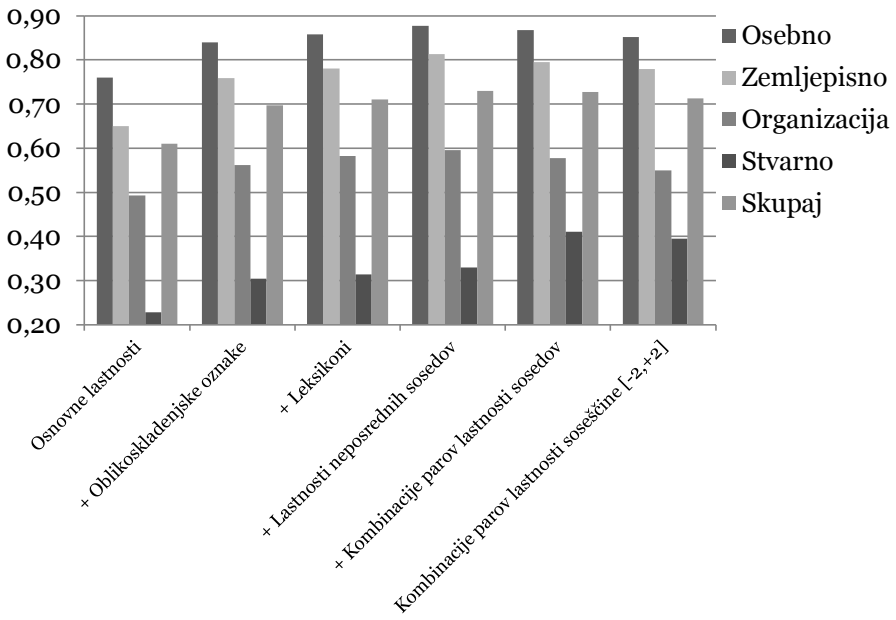
Tabela 4 kaže, da uporaba leksikonov opazno dvigne priklic in natančnost pri osebnih ter zemljepisnih imenih, medtem ko stvarna imena nimajo statistično značilne spremembe v primerjavi z uporabo le oblikoskladenjskih oznak brez leksikonov, kar Tabela 3 prikazuje v drugem delu. Skupna F_1 je statistično značilno višja od F_1, pri kateri so uporabljene le oblikoskladenjske oznake. Ker so leksikoni uporabni le, če je besedilo lematizirano, je uspešna identifikacija imenskih entitet odvisna tudi od obstoja lematizatorja. V nadaljnjih hipotezah in poskusih privzemamo uporabo lastnosti leksikonov ter oblikoskladenjskih oznak kot osnovno različico. V primerjavi s predhodno raziskavo (Štajner in dr. 2012) so tudi v tem poskusu rezultati F_1 opazno višji kot pri modelu s tremi razredi, celo pri obstoječih razredih: osebna imena se izboljšajo z 0.83 na 0.86, zemljepisna pa z 0.70 na 0.78. Medtem ko je stari razred stvarnih imen dosegel 0.49, v tem modelu organizacije doseže 0.58, vendar za ceno slabšega rezultata pri stvarnih imenih.

Tabela 5 prikazuje, da je najboljše delovanje modela doseženo takrat, ko uporabimo kombinacije parov lastnosti neposrednih sosedov, in da je razširjanje soseščine lahko celo škodljivo, saj se dimenzionalnost prostora lastnosti s tem močno poveča, kar otežuje učenje, saj je število primerov bistveno manjše ne le od števila vseh možnih lastnosti, temveč tudi neničelnih lastnosti. Rezultati so podobni pri vseh tipih entitet, kar nakazuje, da je optimalno uporabiti le lastnosti neposrednih sosedov. To se razlikuje od rezultatov iz Štajner in dr. (2012), saj je bil tam najboljši rezultat dosežen s pomočjo kombinacij lastnosti neposrednih sosedov, tu pa so zadostovale le lastnosti neposrednih sosedov. Možna razlaga je, da je ravno ta izboljšava najbolj pomagala razredu stvarnih imen, ki so bila zelo raznolika, tako da je bila korist od dodatnih lastnosti bolj opazna. V scenariju, v katerem je razred organizacij ločen od razreda preostalih stvarnih imen, pa je možno doseči boljši rezultat tudi z manj kompleksnim modelom.

Tip entitete	**Natančnost**	**Priklic**	**F_1**
Lastnosti neposrednih sosedov			
Osebno	0,85	0,91	0,88
Zemljepisno	0,82	0,80	0,81
Organizacija	0,63	0,58	0,60
Stvarno	0,39	0,30	0,33
Skupaj	0,74	0,72	**0,73**
Kombinacije parov lastnosti neposrednih sosedov			
Osebno	0,85	0,89	0,87
Zemljepisno	0,80	0,79	0,79
Organizacija	0,59	0,57	0,58
Stvarno	0,50	0,35	0,41
Skupaj	0,74	0,71	**0,73**
Kombinacije parov lastnosti soseščine [-2,+2]			
Osebno	0,83	0,88	0,85
Zemljepisno	0,78	0,79	0,78
Organizacija	0,59	0,52	0,55
Stvarno	0,50	0,33	0,40
Skupaj	0,73	0,69	**0,71**

Tabela 5: Rezultati poskusov glede na različne kombinacije parov lastnosti sosedov.

V primerjavi z leksikoni tu katerakoli uporaba kombinacij parov izboljša F_1, saj se z 0.69 povzpne na višino od 0.73 do 0.76, odvisno od števila kombinacij soseščine.



Slika 1: F_1 glede na najboljše kumulativno dodane lastnosti.

Slika 1 kaže rast F_1 glede na kumulativno dodajanje novih lastnosti – vsak stolpec ima poleg spodaj navedenih lastnosti tudi vse lastnosti predhodnika. Meritve skrajno levo uporabljajo le lastnosti regularnih izrazov ter dolžino besede, naslednje meritve pa kažejo razlike pri dodajanju novih lastnosti. Tu vidimo, da oblikoskladenjske oznake izboljšajo kakovost na vseh tipih entitet, medtem ko leksikoni izboljšajo le osebna in zemljepisna imena, saj za stvarna imena še nismo uporabili primernega leksikona. Kljub temu pa so imela ravno stvarna imena najvišji napredek pri dodajanju parov kombinacij sosednjih lastnosti, kar lahko pojasnimo s tem, da so stvarna imena pogosto daljša in bolj odvisna od širšega konteksta.

Podrobnejša analiza napak je pokazala, da kljub izločitvi organizacijskih entitet v lasten razred stvarna imena zajemajo več različnih tipov entitet, kar učnemu modelu otežuje posploševanje. V literaturi (Grishman, Sundheim 1996) se v nekaterih domenah uporablja še ožje definirane tipe, kot npr.

geopolitična entiteta, izdelek in dogodek, saj je pri ožje definiranih tipih lažje doseči višjo natančnost izločanja. To potrjujejo tudi rezultati naših poskusov: ko razred stvarnih imen razdelimo na organizacije ter preostala stvarna imena, so rezultati pri prepoznavanju organizacij boljši, kot so bili v predhodni analizi, kjer so bile organizacije del stvarnih imen. Ker pa imajo zaradi te delitve preostala stvarna imena manjšo učno množico, je na tem razredu kakovost samodejnega označevanja slabša.

4.2 Poskusi na slWaC

Da bi preizkusili splošnost pridobljenega modela, smo izvedli tudi merjenje s pomočjo neodvisno označenega testnega korpusa. V ta namen smo uporabili testni del korpusa slWaC (Ljubešić, Erjavec 2011), ki obsega 361 stavkov in se ujema v razredih entitet s ssj500k. Preizkus je potekal tako, da smo s celotnim korpusom ssj500k naučili model, ki je v prejšnjem podpoglavju deloval najbolje, nato pa s tem modelom označili testni korpus slWaC.

	Predlagani pristop, CRF-SlNER, učenje na ssk500k, testiranje na slWaC			**sl-MSD-DISTSIM, učenje in testiranje na slWaC, rezultati iz Ljubešić in dr. (2012)**		
Tip	**Natančnost**	**Priklic**	**F_1**	**Natančnost**	**Priklic**	**F_1**
Osebno	0,76	0,74	0,75	0.86	0.84	0.85
Zemlj.	0,79	0,77	0,78	0.80	0.75	0.77
Org.	0,24	0,41	0,30	0.89	0.36	0.52
Stvarno	0,24	0,14	0,18	0.47	0.24	0.32
Skupaj	0,62	0,61	**0,62**	0.81	0.70	**0.75**

Tabela 6: Rezultati poskusov pri uporabi testnega korpusa slWaC v primerjavi z modelom sl-MSD-DISTSIM, naučenem na korpusu slWaC.

Tabela 6 prikazuje rezultate testiranja našega pristopa CRF-SlNER v primerjavi s pristopom sl-MSD-DISTSIM (Ljubešić in dr. 2012) na testnem

korpusu slWaC, s čimer lahko ocenimo stopnjo posplošitve modela, saj je pristop CRF-SlNER naučen na ssj500k, testiran pa na slWaC. Zaradi vsebinsko različnega učnega in testnega korpusa lahko pričakujemo padec kakovosti označevanja, ki nam pove, kako dobro je model posplošen: nižji kot je padec natančnosti, boljša je posplošitev označevalnika.

Rezultati kažejo, da je pri osebnih imenih F_1 nižji, pri zemljepisnih imenih je primerljiv, organizacije in stvarna imena pa so opazno manj natančno označeni. Iz tega lahko razberemo, da se označevalni model zelo dobro posploši za označevanje zemljepisnih imen, nekoliko manj na osebnih imenih ter relativno slabo za imena organizacij in stvarna imena.

Razlike med korpusi so pogosto tudi tematske: različne tematike imajo različne porazdelitve tipov entitet in nasploh različne entitete. Medtem ko je ssk500k sestavljen iz mešanice različnih žanrov, je slWaC sestavljen iz besedil s spletnih strani.

5 ZAKLJUČEK

V članku smo opisali implementacijo razpoznavanja entitet v slovenskem besedilu s pomočjo nadzorovanega učenja pogojnih naključnih polj z oblikoskladenjskimi in besednimi lastnostmi. Rezultati kažejo visoko zanesljivost razpoznavanja lastnih, zemljepisnih in organizacijskih imen ter slabo razpoznavanje preostalih stvarnih imen. Ugotovili smo tudi, da so v slovenskem jeziku za doseganje dobrega rezultata pri razpoznavanju entitet potrebne oblikoskladenjske oznake, prav tako pa se da kakovost izboljšati z uporabo leksikonov ter kombinacij lastnosti sosednjih besed, kar je znano tudi iz podobnih pristopov pri tujih jezikih.

Članek potrjuje hipotezo, da je moč izboljšati rezultate, če namesto razredov osebnih, zemljepisnih in stvarnih imen uporabljamo razrede osebnih, zemljepisnih, organizacijskih ter preostalih stvarnih imen. Rezultati so se izboljšali, ker tak model organizacije prepoznava veliko natančneje, če sodijo v

samostojen razred, kot če so podmnožica stvarnih imen. Kljub temu, da je zaradi slabega prepoznavanja preostalih stvarnih imen skupni končni rezultat slabši, je pri ostalih razredih vidno izboljšanje F_1: z 0.86 na 0.88 pri osebnih ter z 0.80 na 0.81 pri zemljepisnih imenih.

Preizkus modela na neodvisno označenem korpusu slWaC je prikazal odlično posplošenost pri zemljepisnih imenih, relativno dobro posplošenost pri osebnih ter slabšo posplošenost pri organizacijskih in stvarnih imenih.

Obstoj sistema za razpoznavanje entitet je pomemben korak pri razvoju sistema za razločevanje entitet (Štajner, Mladenić 2009), ki razpoznavanje nadgradi še z določanjem točne identitete. Razločevanje entitet omogoča povezovanje nestrukturiranih besedil s strukturiranimi podatkovnimi bazami, nove metode pa omogočajo tudi razločevanje entitet iz slovenskega besedila in povezovanje s podatkovnimi bazami, izraženimi v drugem jeziku (Štajner, Mladenić 2012).

Da bi bil sistem uporaben tudi za razpoznavanje entitet v besedilih brez oblikoskladenjskih oznak, je bila narejena integracija z oblikoskladenjskim označevalnikom (Rupnik in dr. 2008), ki je na voljo v slovenski različici spletne storitve Enrycher (Štajner in dr. 2010).

Programska oprema, razvita in uporabljena v prikazanih poskusih, je prosto dostopna[2] pod licenco Apache 2.0.

ZAHVALA

Raziskavo je podprla Javna agencija za raziskovalno dejavnost Republike Slovenije s programom Mladi raziskovalec in 7. okvirni program Evropske komisije s projekti XLike (ICT-288342-STREP), MetaNet (ICT-249119-NoE) ter LT-Web (ICT-287815-CSA).

[2] http://ailab.ijs.si/~tadej/slner.zip

LITERATURA

Balog, K., Serdyukov, P., in de Vries, A. P. (2010): Overview of the TREC 2010 Entity Track. Dostopno prek: http://trec.nist.gov/pubs/trec19/papers/ENTITY.OVERVIEW.pdf (11. maj 2013).

Byrd, R. H., Nocedal, J., in Schnabel, R. B. (1994): Representations of Quasi-Newton Matrices and their Use in Limited Memory Methods. *Mathematical Programming*, 63 (1): 129–156.

Che, W., Wang, M., Manning, C. D., in Liu, T. (2013): Named Entity Recognition with Bilingual Constraints. *Proceedings of NAACL-HLT*: 52–62. Atlanta.

Cohen , W. W., in Sarawagi, S. (2004): Exploiting Dictionaries in Named Entity Extraction: Combining Semi-Markov Extraction Processes and Data Integration Methods. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 89–98. Seattle.

Erjavec, T., Fišer, D., Krek, S., in Ledinek, N. (2010a): Jezikovni viri projekta JOS. V T. Erjavec, J. Žganec Gros (ur.): *Zbornik Sedme konference Jezikovne tehnologije*: 42–46. Ljubljana: Institut Jožef Stefan.

Erjavec, T., Fišer, D., Krek, S., in Ledinek, N. (2010b): The JOS Linguistically Tagged Corpus of Slovene. *Seventh International Conference on Language Resources and Evaluation (LREC'10)*: 1806–1809. Valetta.

Erjavec, T., Krek, S., Arhar, Š., Fišer, D., Ledinek, N., Saksida, A., Sivec, B., in Trebar, B. (2010c): *Oblikoskladenjske specifikacije JOS*. Dostopno prek: http://nl.ijs.si/jos/msd/html-sl/ (22. avgust 2013).

Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., Weld, D. S., in Yates, A. (2005): Unsupervised Named-Entity Extraction from the web: An Experimental Study. *Artificial Intelligence,* 165 (1): 91–134.

Finkel, J. R., Grenager, T., in Manning, C. (2005): Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*: 363–370. Stroudsburg.

Grishman, R., in Sundheim, B. (1996): Message Understanding Conference-6: A Brief History. *Proceedings of the 16th Conference on Computational Linguistics, Volume 1*: 466–471. Stroudsburg.

Jung, J. J. (2012): Online Named Entity Recognition Method for Microtexts in Social Networking Services: A Case Study of Twitter. *Expert Systems with Applications*, 39 (9): 8066–8070.

Juršič, M., Mozetič, I., in Lavrač, N. (2007): Learning Ripple Down Rules for Efficient Lemmatization. *Proceedings of the 10th International Multiconference Information Society*: 206–209. Ljubljana.

Lafferty, J., McCallum, A., in Pereira, F. (2001): Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data. *Machine Learning International Workshop*: 282–289. San Francisco.

Ljubešić, N., in Erjavec, T. (2011): hrWac in slWaC: Compiling Web Corpora for Croatian and Slovene. V I. Habernal, V. Matoušek (ur.): *Text, Speech and Dialog: Proceedings of the 14th International Conference, TSD*: 395–402. Pilsen: Springer Berlin Heidelberg.

Ljubešić, N., Stupar, M., in Jurić, T. (2012): Building Named Entity Recognition Models For Croatian and Slovene. V T. Erjavec, J. Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*: 129–134. Ljubljana: Institut Jožef Stefan. Ljubljana.

McCallum, A. K. (2002): MALLET: *A Machine Learning for Language Toolkit*. Dostopno prek: http://mallet.cs.umass.edu (4. november 2013).

Munro, R., in Manning, C. D. (2012): Accurate Unsupervised Joint Named-

Entity Extraction from Unaligned Parallel Text. *Proceedings of the 4th Named Entity Workshop*: 21–29. Stroudsburg.

Rabiner, L., in Juang, B. (1986): An Introduction to Hidden Markov Models. *ASSP Magazine: IEEE*, 3 (1): 4–16.

Rupnik, J., Grčar, M., in Erjavec, T. (2008): Improving Morphosyntactic Tagging of Slovene Language through Metatagging. *Informatica: Intelligent Systems*, Special Issue: 437–444.

Sandhaus, E. (2008): *The New York Times Annotated Corpus*. Linguistic Data Consortium.

Štajner, T., Erjavec, T., in Krek, S. (2012): Razpoznavanje imenskih entitet v slovenskem besedilu. V T. Erjavec, J. Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*: 191–196. Ljubljana: Institut Jožef Stefan. Ljubljana.

Štajner, T., in Grobelnik, M. (2009): Story Link Detection with Entity Resolution. *Proceedings of Semantic Search Workshop at WWW2009*. Madrid.

Štajner, T., in Mladenić, D. (2009): Entity Resolution in Texts Using Statistical Learning and Ontologies. *3rd Asian Semantic Web Conference*: 91–104. Shanghai.

Štajner, T., in Mladenić, D. (2012): Cross-Lingual Named Entity Extraction and Disambiguation. *Proceedings of 4th Jožef Stefan International Postgraduate School Students Conference*: 176–181. Ljubljana.

Štajner, T., Rusu, D., Dali, L., Fortuna, B., Mladenić, D., in Grobelnik, M. (2010): A Service Oriented Framework for Natural Language Text Enrichment. *Informatica*, 34: 307–313.

Statistični urad Republike Slovenije (2012): *Seznam pogostih in redkih imen*. Dostopno prek:

http://www.stat.si/imena_top_imena_spol.asp?r=True (26. september 2013).

Sutton, C., in McCallum, A. (2004): Collective Segmentation and Labeling of Distant Entities in Information Extraction. Dostopno prek: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.141.584 (23. september 2013).

Wikipedia (2012a): *Ženska osebna imena*. Dostopno prek: http: //sl.wikipedia.org/wiki/Kategorija:%C5%BDenska_osebna_imena (23. maj 2013).

Wikipedia (2012b): *Glavna mesta*. Dostopno prek: http: //sl.wikipedia.org/wiki/Kategorija:Glavna_mesta (23. maj 2013).

Wikipedia (2012c): *Moška osebna imena*. Dostopno prek: http: //sl.wikipedia.org/wiki/Kategorija:Mo%C5%A1ka_osebna_imena (23. maj 2013).

Wikipedia (2012č): *Občine Slovenije*. Dostopno prek: http: //sl.wikipedia.org/wiki/Kategorija:Ob%C4%8Dine_Slovenije (23. maj 2013).

Wikipedia (2012d): *Priimki*. Dostopno prek: http://sl.wikipedia.org/wiki/Kategorija:Priimki (23. maj 2013).

Wikipedia (2012e): *Seznam naselij v Sloveniji*. Dostopno prek: http://sl.wikipedia.org/wiki/Seznam_naselij_v_Sloveniji (23. maj 2013).

Wikipedia (2012f): *Seznam suverenih držav*. Dostopno prek: http://sl.wikipedia.org/wiki/Seznam_suverenih_dr%C5%BEav (23. maj 2013).

NAMED ENTITY RECOGNITION IN SLOVENE TEXT

This paper presents an approach and an implementation of a named entity extractor for Slovene language, based on a machine learning approach. It is designed as a supervised algorithm based on Conditional Random Fields and is trained on the ssj500k annotated corpus of Slovene. The corpus, which is available under a Creative Commons CC-BY-NC-SA licence, is annotated with morphosyntactic tags, as well as named entities for people, locations, organisations, and miscellaneous names. The paper discusses the influence of morphosyntactic tags, lexicons and conjunctions of features of neighbouring words. An important contribution of this investigation is that morphosyntactic tags benefit named entity extraction. Using all the best-performing features the recognizer reaches a precision of 74% and a recall of 72%, having stronger performance on personal and geographical named entities, followed by organizations, but performs poorly on the miscellaneous entities, since this class is very diverse and consequently difficult to predict. A major contribution of the paper is also showing the benefits of splitting the class of miscellaneous entities into organizations and other entities, which in turn improves performance even on personal and organizational names. The software, developed in this research is freely available under the Apache 2.0 licence at http://ailab.ijs.si/~tadej/slner.zip, while development versions are available at https://github.com/tadejs/slner.

Keywords: named entity extraction, natural language processing, Slovene language tools