Razpoznavanje imenskih entitet v slovenskem besedilu

Tadej Štajner[*†], **Tomaž Erjavec**[*†], **Simon Krek**[*]

[*]Institut "Jožef Stefan",
Jamova cesta 39 1000 Ljubljana
{tadej.stajner, tomaz.erjavec}@ijs.si, simon.krek@guest.arnes.si

[†]Mednarodna podiplomska šola Jožefa Stefana
Jamova cesta 39, 1000 Ljubljana

Povzetek

Članek predstavi algoritem in implementacijo programa za razpoznavanje imenskih entitet v slovenskem jeziku s pomočjo strojnega učenja. Nadzorovan pristop na osnovi pogojnih naključnih polj je naučen na označenem korpusu ssj500k. V korpusu, ki je prosto dostopen pod licenco Creative Commons, so pri besednih pojavnicah poleg oblikoskladenjskih oznak oz. lastnosti in lem označena tudi osebna, zemljepisna ter stvarna imena. Članek predstavi vpliv na natančnost razpoznavanja ob uporabi oblikoskladenjskih oznak, leksikonov ter konjunkcij sosednjih značilk. Pomembna ugotovitev raziskave je, da oblikoskladenjske oznake koristijo pri razpoznavanju entitet. V kombinaciji z vsemi ostalimi značilkami doseže sistem na testni množici 77% natančnost in 75% priklic, pri čemer so lastna in zemljepisna imena razpoznana bistveno bolje kot stvarna, saj je razred stvarnih imen zelo raznolik in zato težaven za učenje. Programska oprema, razvita in uporabljena v teh poskusih, je prosto dostopna pod licenco Apache 2.0 na naslovu `http://ailab.ijs.si/~tadej/slner.zip`.

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 using the ssj500k annotated Slovene corpus for training data. The corpus, which is available under a Creative Commons licence, is annotated with morphosyntactic tags, as well as named entities of people, locations and real names of other entities. The paper discusses the influence of morphosyntactic tags, lexicons and offset conjunctions of features of neighboring words. An important contribution of this investigation is that morphosyntactic tags benefit named entity extraction. In concert with all other features, it reaches a precision of 77% and a recall of 76%, having stronger performance on personal and geographical named entities than on other entities, since the class of other entities is very diverse and consequently difficult to predict. The software, developed in this research is freely available under the Apache 2.0 licence at `http://ailab.ijs.si/~tadej/slner.zip`.

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, na primer lastno ime osebe, kraja ali organizacije, k vsebini besedila skupaj prispevajo več informacij, kot bi bilo moč razbrati zgolj iz števila posameznih besed.Razpoznavanje entitet obravnava besedilo na drugem nivoju abstrakcije, ker ne govorimo več o posameznih besedah, temveč (največkrat) o dvo- ali večbesednih entitetah. Pri iskanju informacij so lastna imena torej predstavljena kot entiteta, kar nam omogoča, da na besedilni korpus ali podatkovno bazo pogledamo na drugačen način - skozi indeksacijo entitet, ki se v tem korpusu pojavljajo. Na primer, *prikaži mi vse članke o Institutu Jožef Stefan*. V časopisni industriji in založništvu je pogosta praksa, da entitete in ključne besede, ki se pojavijo v člankih, indeksirajo 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 namen povezovanja zgodb v časopisnih člankih (Štajner and Grobelnik, 2009), kjer 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 and Sundheim, 1996), ki se je odvijala v devetdesetih letih in *TREC* (Balog et al., 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 procesiranja naravnega jezika, pri čemer je bilo veliko nalog osredotočenih na razpoznavanje entitet. Najzmogljivejši sistemi trenutno uporabljajo predvsem postopke strojnega učenja, natančneje modele na probablističnih grafih, kot so npr. skriti Markovski modeli (*Hidden Markov Models*) (Rabiner and Juang, 1986) ali pogojna naključna polja (*Conditional Random Fields*) (Lafferty et al., 2001), npr. Mallet (McCallum, 2002) ali Stanford NER (Finkel et al., 2005). V praksi so ti sistemi implementirani z nadzorovanim učenjem na besedilu, kjer so entitete že označene. V procesu učenja se za vsako besedo generirajo posamezne lastnosti, kot na primer 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 stroj-

nem učenju, tako da tvorita hibridni sistem (Cohen and Sarawagi, 2004). Nekateri sistemi uporabljajo tudi nenadzorovano izločanje entitet, saj ta pristop ne zahteva vnaprejšnjega učenja (Etzioni et al., 2005).

V nadaljevanju ima članek naslednjo strukturo: v razdelku 2 predstavimo korpus, na katerem je bil sistem naučen in testiran, v razdelku 3 opišemo razviti razpoznavalnik, v razdelku 4 poskuse, ki smo jih izvedli, razdelek 5 pa vsebuje zaključke.

2. Učni korpus ssj500k

Za nadzorovano učenje je potreben korpus, kjer so pojavitve lastnih imen ustrezno označene. Za slovenski jezik do sedaj še nismo imeli tako označenega korpusa, vendar je bil pred kratkim izdelan ročni označeni korpus ssj500k, ki je bil nastal v okviru projekta Sporazumevanje v slovenskem jeziku (SSJ) in temelji na učnih korpusih jos100k in jos1M, izdelanih v okviru projekta JOS (Erjavec et al., 2010b; Erjavec et al., 2010a). Korpus ssj500k sestavljata dva dela: celotni korpus jos100k ter dodatnih 400.000 besed iz enomilijonskega korpusa jos1M. Vsi jezikoslovni metapodatki (oznake, leme, tokenizacija) so bili v korpusu ssj500k še enkrat ročno pregledani, 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 in razčlenjevalnikov tudi pri teh dveh postopkih. Učni korpus ssj500k je prosto dostopen pod licenco Creative Commons Priznanje avtorstva-Nekomercialno 3.0"na spletnih straneh projekta SSJ `http://www.slovenscina.eu/` oz. `http://nl.ijs.si/ssj/`. 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 segmentirana v tri kategorije: osebna (1.921), zemljepisna (1.284) in stvarna (1.192). 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 potem dolgi rep do stavka s 47 kar lastnimi imeni.

3. Implementacija

V skladu s trenutno prakso obstoječih sistemov za druge jezike implementacija sistema, predstavljenega v tem članku, uporablja nadzorovano učenje s pogojnimi naključnimi polji (*Conditional Random Fields*, ali krajše CRF), ki temelji na sistemu Mallet (McCallum, 2002).

3.1. Model

Pogost pristop pri modeliranju zaznavanja imenskih entitet je verižni model, kjer besede označujemo zaporedno, pri vsaki odločitvi pa med drugim upoštevamo tudi odločitev klasifikacije na prejšnjem koraku. V takšnem modelu, kot je na primer sekvenčni CRF, so stanja določena z željenimi oznakami, ki predstavljajo tipe entitet. Množica stanj modela je torej $osebno, zemljepisno, stvarno, brez$. Ko predstavimo stavek kot zaporedje besed, v postopku označevanja vsaki besedi priredimo oznako najverjetnejšega stanje glede na oznako prejšnje besede ter glede na značilke trenutne besede. V splošnem lahko v modelu CRF predstavimo tudi odvisnosti višjih redov ali odvisnosti v poljubnem acikličnem grafu, vendar je za označevanje besedil najprimernejši verižni model prvega reda. Z drugimi besedami, model, ki ima lastnost, da je trenutno stanje odvisno le od lokalnih značilk ter od predhodnega razreda besede.

Naj bo $G = (V, E)$ graf, kjer je $Y = (Y_v)_{v \in V}$, tako da posamezne dimenzije Y predstavljajo vozlišča G. Iz stališča uporabe je X množica primerov v obliki vektorjev značilk, Y pa naključna spremenljivka, kjer posamezno stanja Y_v predstavljajo tudi ciljne razrede - tipe imenskih entitet. (X, Y) je pogojno naključno polje, če imajo naključne spremenljivke Y_v Markovsko lastnost glede na sosednost, kar pomeni, da je novo stanje odvisno le od prejšnjega stanja: $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim w)$, kjer $w \sim v$ pomeni, da sta w in v soseda. Konkretno, to pomeni, da je oznaka trenutne besede odvisna od značilk trenutne besede ter oznake prejšnje besede.

Pogojna verjetnost med X in Y je tako opisana z množico funkcij značilk oblike $f_k(y, y', x_t)_{k=1}^{K} \in Re^K$. Na primer, $f_{upper-person}$ je lahko tovrstna funkcija, ki vrne 1 v primeru, ko se trenutna beseda začne z veliko začetnico in da je predhodna beseda označena kot osebno ime, sicer pa 0. Linearno verižno pogojno naključno polje (*Linear chain CRF*) je v tem primeru porazdelitev $p(y|x)$, ki jo opišemo z množico parametrov $\Lambda = \lambda_k \in \Re^K$:

$$p(y|x) = \frac{1}{Z(x)} \exp\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\} \quad (1)$$

kjer je $Z(x)$ normalizacijska funkcija. Za uporabo modela je nato potrebno oceniti vrednosti parametrov Λ, ki nam povedo v kolikšni meri je določena značilka povezana z določenim ciljnim razredom. V ta namen se tipično uporablja maksimizacija regulariziranega pogojnega log-verjetja (*conditional log-likelihood*) glede na množico učnih primerov, kar lahko predstavimo s sledečo enačbo:

$$l(\Lambda) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(y_t^{(i)}, y_{t-1}^{(i)}, x_t^{(i)}) -$$

$$-\sum_{i=1}^{N} \log Z(x^{(i)}) - \sum_{k=1}^{K} \frac{\lambda_k^2}{2\sigma^2} \quad (2)$$

Zadnji del enačbe predstavlja regularizacijo, ki se uporablja za preprečevanje prekomernega prileganja modela podatkom. Izraz $\frac{1}{2\sigma^2}$ predstavlja moč regularizacije, ki nam pove, kolikšno kazen dobijo previsoke uteži λ_k. Ker pa $l(\Lambda)$ ni moč maksimizirati v zaprti obliki, se v ta namen uporablja numerična optimizacija s pomočjo delnih odvodov. V ta namen za rešitev optimizacije ocenjevanja parametrov uporabimo optimizacijski algoritem L-BFGS (Byrd et al., 1994). Ko se naučimo parametre modela, jih lahko uporabimo za označevanje neoznačenega besedila. Za to uporabimo inferenčni algoritem *loopy belief propagation* (Sutton and McCallum, 2004).

3.2. Značilke

Pri implementaciji pristopa za razpoznavanje entitet je ključno, da si lahko pomagamo s čim bolj raznolikimi tipi informacij. V ta namen uvajamo štiri kategorije značilk, kjer vsaka kategorija prinaša dodatno informacijo, kar demonstriramo s poskusi v razdelku 4.

3.2.1. Značilke črkovnih vzorcev

Pri obstoječih pristopih za zaznavanje entitet so najbolj tipične značilke grafemskih vzorcev, kjer lahko vsako posamezno besedo opišemo z binarno značilko vzorca. S pomočjo regularnih izrazov smo določili naslednje značilke, ki se že uporabljajo pri zaznavanju imenskih entitet. Značilka dobi vrednost 1 le če beseda ustreza regularnemu izrazu. To množico značilk vzamemo kot osnovo, kateri nato dodajamo ostale razrede.

Uporabljene značilke črkovnih vzorcev so:

- Velika začetnica le na začetku besede (npr. Ljubljana)
- Le velike črke v celotni besedi (npr. IJS)
- Mešane velike črke znotraj besede (npr. iPod)
- Številke v besedi (npr. ZVCP-1)
- Le številke v besedi (npr. 2012)
- Numerični izraz (npr. +3.14)
- Alfanumerični izraz, ki vsebuje le številke in črke (npr. E3)
- Rimska številka (npr. XVIII)
- Vsebuje vezaj ali pomišljaj (npr. Šmarje-Sap)
- Kratica, sestavljena le iz velikih črk, lahko ločenih s piko (npr. I.M.V.)
- Inicialka, posamezna velika črka, ki ji sledi pika. (npr. John *F.* Kennedy)
- Posamezna črka, ne glede na velike ali male črke (npr. odgovor *a*)
- Pozamezna velika črka (npr. plan *B*)
- Ločilo (npr. !)
- Narekovaj (npr. “)
- Le male črke (npr. *pisarna*)

3.2.2. Značilke zunanjih virov znanja

Poleg črkovnih vzorcev lahko uporabljamo tudi zunanje znanje v obliki leksikonov, ki vsebujejo že znana imena entitet. S tem pristopom v model vključimo znanje, ki bi ga bilo le z nadzorovanim učenjem težko nadomestiti. V ta namen definiramo leksikonsko značilko, ki dobi vrednost 1.0 le, če je lema besede vsebovana v določenemu leksikonu, kjer imamo za vsak leksikon po eno binarno značilko. Uporabimo prisotnost *leme*, saj bi bilo pripadnost posameznemu leksikonu pri besedni obliki zaradi bogate slovenske morfologije težko preverjati. Vsak leksikon se prevede v eno značilko. Večino leksikonov smo vzeli iz slovenske različice Wikipedije, ki je prosto dostopni vir in obsega dovolj široko paleto tematskih domen za splošno razpoznavanje entitet. Uporabljajo se sledeči leksikoni:

- kraji v Sloveniji iz slovenske Wikipedije (Wikipedia, 2012f)
- države iz slovenske Wikipedije (Wikipedia, 2012g)
- kraji v tujini iz slovenske Wikipedije (Wikipedia, 2012b)
- občine v Sloveniji iz slovenske Wikipedije (Wikipedia, 2012d)
- 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 pogostih in redkih imen iz Statističnega urada Slovenije (Statistični Urad Slovenije, 2012)
- seznam moških imen iz slovenske Wikipedije (Wikipedia, 2012c)
- seznam ženskih imen iz slovenske Wikipedije (Wikipedia, 2012a)
- seznam priimkov iz slovenske Wikipedije (Wikipedia, 2012e)
- imena dni v tednu
- imena mesecev

V primeru, da besedilo ni lematizirano, lahko uporabimo tudi samodejni lematizator (**?**).

3.2.3. Značilke oblikoskladenjskih lastnosti

Tretji potencialni vir informacij za razpoznavanje entitet so v korpusu že prisotne oblikoskladenjske oznake besed, ki jih prevedemo v značilke po specifikacijah tabele oznak (Erjavec et al., 2010c). Na primer, beseda *narediti* z oznako *Ggdn* dobi značilke *Category=verb, VerbType=main, Aspect=perfective, VForm=infinitive*, beseda *predsednik* z oblikoskladenjsko oznako *Sometd* pa značilke *Category=noun, NounType=common, Gender=masculine, Number=singular, Case=accusative, Animate=yes*. Če besedilo ni označeno, lahko uporabimo ustrezni oblikoskladenjski označevalnik (Rupnik et al., 2008). Uporaba oblikoskladenjskih oznak temelji na predpostavki, da iz vzorcev oznak lahko razbremo prisotnost entitet. Uporaba mestnika v kombinaciji z veliko začetnico lahko denimo nakazuje prisotnost zemljepisnega imena.

3.2.4. Strukturne značilke

Poleg regularnih izrazov, leksikonov in oblikoskladenjskih oznak lahko uporabljamo tudi različne strukturne značilke, ki izvirajo iz same zgradbe stavka kot zaporedja besed. Prva množica strukturnih značilk 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 značilka pa je odvisna od pripadnosti tem razredom (npr. *Length=5*).

Druga množica strukturnih značilk, **konjunkcija sosednjih značilk**, je definirana kot preslikava nad obstoječimi značilkami. To je metoda za generiranje dodatnih značilk, ki za vsako besedo sestavi nove značilke kot kombinacije značilk njenih sosedov znotraj določenega okna. Uporablja se predvsem v tistih verižnih klasifikatorjih, kjer soodvisnosti med značilkami in razredi niso odvisne le od prejšnjega in 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, konjunkcije sosednjih značilk uporabimo kot približek. Na primer, če se trenutna beseda nahaja dve mesti za besedo z veliko začetnico, dobi značilko $kapitalizacija_{-2}$. V nadaljnih 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 značilk v razponu dve mesti naprej ter nazaj, npr. kombinacija $-2,1$ predstavlja konjunkcijo značik besede dve mesti pred trenutno z značilkami naslednje besede. Tovrstno generiranje značilk lahko izredno poveča število možnih značilk in s tem upočasni učenje ter povečuje nevarnost prekomernega prilagajanja.

4. Poskusi

S poskusi smo želili odgovoriti na vprašanja glede smiselnosti uporabe različnih razredov značilk glede na meritve:

- Ali oblikoskladenjske oznake izboljšajo model?
- Ali uporaba leksikonov izboljša model?
- Ali kombinacije parov značilk v soseščini izboljšajo model?

Poskuse smo izvedli z desetkratnim navzkrižnim preverjanjem, kjer naključnih devetdeset odstotkov podatkov uporabimo za učenje, preostalo 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 nadaljni poskus kot osnovo uporablja različico predhodnega poskusa, ki je v tistem krogu imela najboljši izid.

Rezultati v tabeli 2 potrjujejo, 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. Poskusi tudi kažejo, da je sistem razmeroma uspešen pri razpoznavanju osebnih imen, nekaj slabši pri zemljepisnih imenih in neuspešen pri prepoznavanju stvarnih imen. Zemljepisna imena je lažje razpoznati, ker gre za samostalnike z veliko začetnico, ki so tipično v mestniku, stvarna imena pa pogosto sestavljejo daljše zveze s pridevniki (*Evropska komisija*) ali predlogi (*Ministrstvo za obrambo*) in različnimi skloni znotraj besedne zveze, zaradi česar je možnih variacij preveč, da bi jih lahko zajeli v obstoječih učnih podatkih. V nadaljnih poskusih privzemamo, da so oblikoskladenjske oznake vedno prisotne. Z odebeljeno je označena glavna metrika - povprečje F_1 čez vse razrede.

Tip entitete	**Natančnost**	**Priklic**	F_1
Brez oblikoskladenjskih oznak			
Osebna	0.6207	0.6533	0.6342
Zemljepisna	0.4426	0.4868	0.4595
Stvarna	0.3171	0.1970	0.2412
Skupno	0.4932	0.4464	**0.4681**
Z oblikoskladenjskimi oznakami			
Osebna	0.7632	0.8526	0.8046
Zemljepisna	0.7303	0.6770	0.7016
Stvarna	0.5756	0.4283	0.4881
Skupno	0.7011	0.6585	**0.6788**

Tabela 2: Rezultati poskusov glede na uporabljene oblikoskladenjske oznake

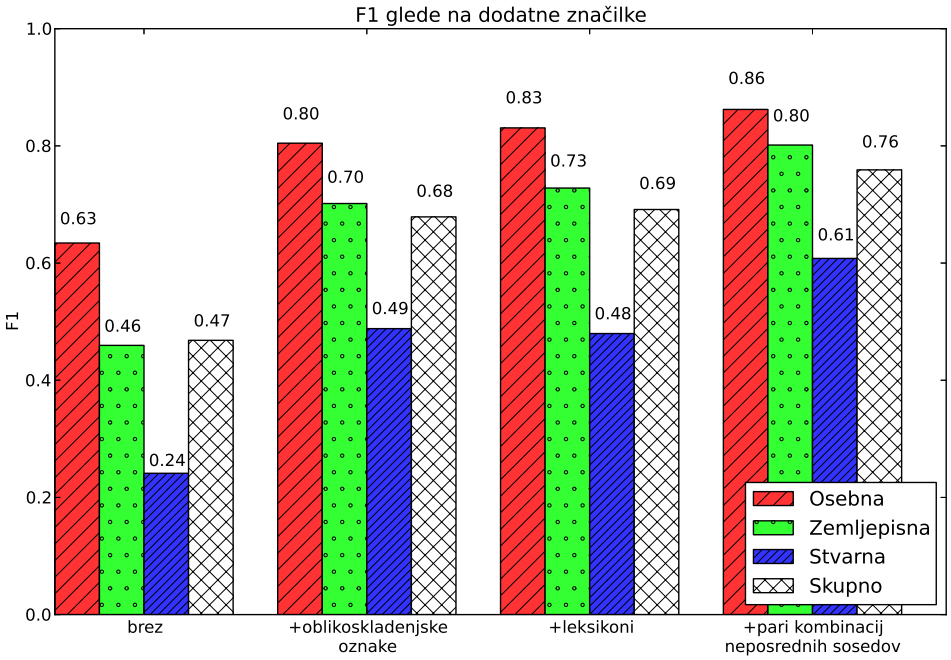
Tip entitete	**Natančnost**	**Priklic**	F_1
Z uporabo leksikonov			
Osebna	0.7851	0.8832	0.8307
Zemljepisna	0.7727	0.6892	0.7280
Stvarna	0.5524	0.4261	0.4797
Skupno	0.7126	0.6718	**0.6914**

Tabela 3: Rezultati poskusov glede na uporabljene leksikone

Tabela 3 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 v drugem delu tabele 2. Skupna F_1 je statistično značilno višja od F_1, ko so uporabljene le oblikoskladenjske oznake. Ker so leksikoni uporabni le v primeru, da je besedilo lematizirano, je uspešna identifikacija imenskih entitet odvisna tudi od obstoja lematizatorja. V nadaljnih hipotezah in poskusih privzemamo uporabo značilk leksikonov in oblikoskladenjskih oznak kot osnovno verzijo.

Iz tabele 4 je razvidno, da je najboljše delovanje modela doseženo takrat, ko uporabimo kombinacije parov značilk neposrednih sosedov, in da je razširjanje soseščine škodljivo, saj se dimenzionalnost prostora značilk s tem močno poveča, kar otežuje proces učenja, saj je število primerov bistveno manjše ne le od števila vseh možnih značilk, temveč tudi ne-ničelnih značilk. Rezultati so podobni pri vseh tipih entitet, kar nakazuje, da je optimalno uporabiti le kombinacije značilk neposrednih sosedov. V primerjavi s tabelo 3 katerakoli uporaba kombinacij parov izboljša F_1, saj se iz 0.69 povzpne od 0.73 do 0.76, odvisno od števila kombinacij soseščine.

Slika 1 kaže rast F_1 metrike glede na dodajanje novih značilk. Meritve skrajno levo uporabljajo le značilke regularnih izrazov ter dolžino besede, naslednje meritve pa



Slika 1: F_1 glede na najboljše kumulativno dodane značilke

Tip entitete	**Natančnost**	**Priklic**	F_1
Značilke neposrednih sosedov			
Osebna	0.8255	0.8788	0.8507
Zemljepisna	0.8093	0.7492	0.7762
Stvarna	0.5851	0.5167	0.5469
Skupno	0.7405	0.7157	**0.7277**
Kombinacije parov značilk neposrednih sosedov			
Osebna	0.8463	0.8816	0.8622
Zemljepisna	0.8279	0.7786	0.8014
Stvarna	0.6472	0.5774	0.6079
Skupno	0.7729	0.7458	**0.7590**
Kombinacije parov značilk soseščine [-2,+2]			
Osebna	0.8310	0.8789	0.8538
Zemljepisna	0.8076	0.7616	0.7819
Stvarna	0.6494	0.5568	0.5966
Skupno	0.7656	0.7334	**0.7489**

Tabela 4: Rezultati poskusov glede na različne kombinacije parov značilk sosedov

kažejo razlike pri dodajanju novih značilk. Tu lahko vidimo, da oblikoskladenjske oznake statistično značilno 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 ravno stvarna imena imela najvišji napredek pri dodajanju parov kombinacij sosednjih značilk, kar lahko pojasnimo s tem, da so stvarna imena pogostokrat daljša in bolj odvisna od širšega konteksta. Podrobnejša analiza napak pokaže, da stvarna imena zajemajo mnogo različnih tipov entitet, kar učnemu modelu otežuje posploševanje. V literaturi (Grishman and Sundheim, 1996) se uporablja ožje definirane tipe, kot na primer *organizacija*, *geopolitična entiteta*, *izdelek* ter *dogodek*, saj je pri ožjih tipih lažje doseči višjo natančnost izločanja.

5. Zaključek

Članek je opisal implementacijo razpoznavanja entitet v slovenskem besedilu s pomočjo nadzorovanega učenja pogojnih naključnih polj z oblikoskladenjskimi in besednimi lastnostmi. Rezultati kažejo na visoko zanesljivost zaznavanja lastnih in zemljepisnih imen in nekoliko manj zanesljivo zaznavanje stvarnih imen, kar je glede na majhen učni korpus zadovoljiv rezultat. Rezultati tudi potrjujejo, da v slovenskem jeziku oblikoskladenjske oznake koristijo pri razpoznavanju entitet, prav tako pa se da kakovost izboljšati z uporabo leksikonov ter kombinacij značilk sosednjih besed. Pri stvarnih imenih bi bilo moč doseči boljši rezultat, če bi jih natančneje delili na organizacije, dogodke, izdelke in ostala stvarna imena, saj je trenutno razred stvarnih imen zelo raznolik in s tem težaven za učenje. Obstoj sistema za razpoznavanje entitet predstavlja tudi pomemben korak za razvoj sistema za razločevanje entitet (Štajner and Mladenić, 2009), ki razpoznavanje nadgradi še z določanjem točne identitete entitete. Razločevanje entitet nam omogoča povezovanje nestrukturiranih besedil s strukturiranimi podatkovnimi bazami, nove metode pa nam omogočajo tudi razločevanje enititet iz slovenskega besedila in povezovanje s podatkovnimi bazami, izraženimi v drugem jeziku (Štajner and Mladenić, 2012). Da bi bil sistem uporaben tudi za razpoznavanje entitet v besedilih brez oblikoskladenjskih oznak, je bila narejena tudi integracija z oblikoskladenjskim označevalnikom (Rupnik et al., 2008), ki je na voljo v slovenski različici spletne storitve Enrycher (Štajner et al., 2010). Programska oprema, razvita in uporabljena v teh poskusih, je prosto dostopna pod licenco Apache 2.0 na naslovu `http://ailab.ijs.si/~tadej/slner.zip`

Zahvale

To delo je podprla Javna agencija za raziskovalno dejavnost Republike Slovenije, 7. okvirni program Evropske Komisije s projektom XLike (ICT-288342-STREP).

6. Literatura

K. Balog, P. Serdyukov, in A.P. de Vries. 2010. Overview of the TREC 2010 Entity Track. *NIST Special Publication: TREC*.

R.H. Byrd, J. Nocedal, in R.B. Schnabel. 1994. Representations of quasi-newton matrices and their use in limited memory methods. *Mathematical Programming*, 63(1):129–156.

W.W. Cohen in S. Sarawagi. 2004. Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods. V: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, str. 89–98. ACM.

T. Erjavec, D. Fišer, S. Krek, in N. Ledinek. 2010a. Jezikovni viri projekta JOS. *Zbornik Sedme konference Jezikovne tehnologije, 14. do 15. oktober 2010 : zbornik 13. mednarodne multikonference Informacijska družba - IS 2010, zvezek C.*, C:42–46.

Tomaž Erjavec, Darja Fišer, Simon Krek, in Nina Ledinek. 2010b. The JOS linguistically tagged corpus of Slovene. V: *Seventh International Conference on Language Resources and Evaluation, LREC'10*, Paris. ELRA.

Tomaž Erjavec, Simon Krek, Špela Arhar, Darja Fišer, Nina Ledinek, Amanda Saksida, Breda Sivec, in Blaž Trebar. 2010c. Oblikoskladenjske specifikacije JOS. `http://nl.ijs.si/jos/msd/`.

O. Etzioni, M. Cafarella, D. Downey, A.M. Popescu, T. Shaked, S. Soderland, D.S. Weld, in A. Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134.

J.R. Finkel, T. Grenager, in C. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. *Ann Arbor*, 100.

R. Grishman in B. Sundheim. 1996. Message understanding conference-6: A brief history. V: *Proceedings of the 16th conference on Computational linguistics-Volume 1*, str. 466–471. Association for Computational Linguistics Morristown, NJ, USA.

J. Lafferty, A. McCallum, in F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. V: *Machine Learning International Workshop*, str. 282–289. Citeseer.

A.K. McCallum. 2002. Mallet: A machine learning for language toolkit.

L. Rabiner in B. Juang. 1986. An introduction to hidden Markov models. *ASSP Magazine, IEEE*, 3(1):4–16.

J. Rupnik, M. Grčar, in T. Erjavec. 2008. Improving morphosyntactic tagging of Slovene language through metatagging. *Informatica Special Issue: Intelligent Systems Guest Editors: Costin Badica*, str. 437–444.

E. Sandhaus. 2008. The New York Times Annotated Corpus. *Linguistic Data Consortium, Philadelphia*.

T. Štajner in M. Grobelnik. 2009. Story link detection with entity resolution. V: *Proceedings of Semantic Search Workshop at WWW2009, Madrid, Spain*.

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

T. Štajner in D. Mladenić. 2012. Cross-lingual named entity extraction and disambiguation. *Proceedings of 4th Jožef Stefan International Postgraduate School Students Conference*, str. 176–181.

Statistični Urad Slovenije. 2012. Seznam pogostih in redkih imen. `http://www.stat.si/imena_top_imena_spol.asp?r=True.`, 5.

C. Sutton in A. McCallum. 2004. Collective segmentation and labeling of distant entities in information extraction. Tehnično poročilo, DTIC Document.

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

Wikipedia. 2012a. Ženska osebna imena. `http://sl.wikipedia.org/wiki/Kategorija:%C5%BDenska_osebna_imena`, 5.

Wikipedia. 2012b. Glavna mesta. `http://sl.wikipedia.org/wiki/Kategorija:Glavna_mesta`, 5.

Wikipedia. 2012c. Moška osebna imena. `http://sl.wikipedia.org/wiki/Kategorija:Mo%C5%A1ka_osebna_imena`, 5.

Wikipedia. 2012d. Občine Slovenije. `http://sl.wikipedia.org/wiki/Kategorija:Ob%C4%8Dine_Slovenije`, 5.

Wikipedia. 2012e. Priimki. `http://sl.wikipedia.org/wiki/Kategorija:Priimki`, 5.

Wikipedia. 2012f. Seznam naselij v Sloveniji. `http://sl.wikipedia.org/wiki/Seznam_naselij_v_Sloveniji`, 5.

Wikipedia. 2012g. Seznam suverenih držav. `http://sl.wikipedia.org/wiki/Seznam_suverenih_dr%C5%BEav`, 5.