Strojno učenje lematizacije neznanih slovenskih besed

Sašo Džeroski, Tomaž Erjavec

Odsek za inteligentne sisteme
Institut "Jožef Stefan"
Jamova 39, 1000 Ljubljana
saso.dzeroski@ijs.si, tomaz.erjavec@ijs.si

Povzetek

Lematizacija je postopek, pri katerem neki besedni obliki v besedilu tvorimo lemo (geslo, iztočnico); je koristna funkcija mnogim programom za obdelavo besedil. Besede slovenskega jezika lahko lematiziramo razmeroma natančno, a v splošnem dvoumno, z uporabo oblikoslovnega leksikona. Za besede, ki niso zajete v leksikonu, je lematizacija lahko edinole hevristična. Pravila za približno lematizacijo neznanih besed v besedilih so sicer lahko napisana ročno, lahko pa so tudi naučena iz vnaprej pripravljenih jezikovnih virov. V članku obravnavamo lematizacijo neznanih slovenskih samostalnikov in pridevnikov v prostih besedilih z uporabo metod strojnega učenja. Problem učenja lematizacije obravnavamo kot sklop dveh problemov: prvi je učenje oblikoslovnega označevanja, drugi pa učenje oblikoslovne analize. Za oblikoslovno označevanje smo uporabili statistični trigramski označevalnik TnT, za oblikoslovno analizo pa odločitvene sezname in induktivno logično programiranje v sistemu CLOG. Kot jezikovni vir za učenje in ovrednotenje sistema smo uporabili oblikoslovno označeni del korpusa MULTEXT-East (slovenski prevod romana '1984' G. Orwella), ki vsebuje približno 100,000 besed. Naučeni model smo nato preizkusili še na besedilu iz odprte domene, ki vsebuje 13.000 pojavitev neznanih samostalnikov in pridevnikov. Predstavljena metoda lematizacije na teh besedah doseže natančnost 81 %.

1. Uvod

Lematizacija je osnovna funkcija mnogih sistemov za obdelavo naravnega jezika. Lematizacija predstavlja korak v normalizaciji besedila, kjer so pregibne besedne oblike v besedilu, npr. *mize* ali *mizi*, poenotene na svojo lemo. Te leme se nato lahko uporabljajo pri leksikalni analizi besedil, npr. kot iskalna funkcija konkordančnikov, avtomatski gradnji leksikonov, pri strojnem prevajanju, itd.

Obstaja več vrst lematizacije: ti. krnenje je enostavno, hitro in fleksibilno, zato pa precej nenatančno in se uporablja predvsem pri zajemanju informacij (Dimec et al., 1998). Pri krnenju skušamo zavreči pregibne in besedotvorne končnice, tako da od besedne oblike ostane samo "krn", ki naj bi bil obenem nespremenjiv nosilec pomena, npr. *miz* ali *ultrazvoč*. Krn lahko, ni pa to nujno, predstavlja tudi oblikoslovno osnovo besede. Metodi lahko povzročajo težave premene v pregibnih besedah (npr. *poslanec/poslanca*) ali pa napačna odločitev o tem, kaj je končnica. Ker so krni tudi uporabniku neprijazni, smo v naših poskusih lematizacijo interpretirali kot iskanje slovarskih iztočnic, npr. *miza*, *lematizirati* ali *škarje*.

Z uporabo velikega leksikona, ki vsebuje informacije o pregibnosti besed, je mogoče natančno, a v splošnem dvoumno (*hotela: hoteti? hotel?*) lematizirati znane besede (Zupan, 1999). Nedvoumna lematizacija je mogoča samo z upoštevanjem konteksta besede, torej z vsaj površinsko skladenjsko analizo besedila (Jakopin and Bizjak, 1997). Za te namene se pogosto uporabljajo oblikoslovni označevalniki (part-of-speech taggers, syntactic wordclass taggers) (van Halteren, 1999), ki se naučijo hevrističnih pravil za označevanje iz pravilno označenega korpusa.

Bolj zahtevna kot lematizacija znanih pa je lematizacija neznanih besed, torej tistih, ki se ne pojavljajo v leksikonu aplikacije. Medtem ko so rezultati pri sistemih z ročno napisanimi pravili precej spodbujajoči vsaj za francoski jezik (Chanod and Tapanainen, 1995), je bilo manj raziskav opravljenih z avtomatsko naučenimi modeli. Poskus, ki obravnava oblikoslovno označevanje neznanih besed, je bil za angleški jezik opisan v (Mikheev, 1997): sistem se s pomočjo leksikona in neoznačenega korpusa nauči pravil za določanje množice oblikoslovnih oznak neke besede glede na njeno predpono in končnico; ta množica je za neko neznano besedo nato posredovana oblikoslovnemu označevalniku, ta pa enoumno določi oblikoslovno oznako in s tem lemo. Ta kombinacija metod doseže natančnost 88 % za neznane besede in 96 % pri znanih besedah.

V članku predstavimo metodo za strojno učenje lematizacije neznanih slovenskih besed. Problema se lotimo v dveh korakih. Najprej se naučimo pravil za oblikoslovno analizo, ki vrnejo lemo besede ob podani besedni obliki in njeni oblikoslovni oznaki, nato pa se naučimo še oblikoslovno označevati besede v slovenskih besedilih. S kombinacijo obeh modelov je nato možno določiti lemo neznanih besed v besedilu.

V poglavju 2 opišemo korpus, ki smo ga uporabili za učenje in preiskušanje oblikoslovne analize in označevalnika. V poglavju 3 opišemo metodo učenja pravil za oblikoslovno analizo, v poglavju 4 pa metodo za učenje oblikoslovnega označevalnika. Pri obeh metodah tudi podamo ovrednotenje postopka na testnem delu korpusa. Poglavje 5 opiše evaluacijo lematizacije s kombiniranjem obeh sistemov, in sicer na testnem delu korpusa ter na korpusu iz odprte domene. Poglavje 6 zaključi in poda nekaj smernic za bodoče delo.

2. Podatki za učenje in ovrednotenje

Projekt MULTEXT-East (Dimitrova et al., 1998) je izdelal korpuse, leksikone in orodja za šest vzhodno- in srednje-evropskih jezikov. Korpus vsebuje roman "1984" G. Orwella v originalu in prevodih, v katerih so vse besede enoumno označene s svojimi oblikoslovnimi oznakami in lemami. Za poskuse, o katerih je govora v tem članku, smo uporabili slovenski prevod romana, ki je označen za povedi

(6,689) in tokeniziran na besede (90,792) in ločila (21,998). Vsaka beseda v tem korpusu je označena s kontekstno odvisnimi oblikoslovnimi podatki, ki sestojijo iz leme besede in njenega oblikoslovnega opisa. Korpus je bil po koncu projekta MULTEXT-East še nadalje očiščen v sklopu projekta ELAN (Erjavec, 1999). Tu je bil tudi zapisan v skladu s priporočili TEI, Text Encoding Initiative (Sperberg-McQueen and Burnard, 1994). V ilustracijo podamo stavek iz romana:

```
<s id="Osl.1.2.3.4">
<w lemma="Winston" msd="Npmsn">Winston</w>
<w lemma="se" msd="Px------y">se</w>
<w lemma="biti" msd="Vcip3s--n">je</w>
<w lemma="napotiti" msd="Vmps-sma">napotil</w>
<w lemma="proti" msd="Spsd">proti</w>
<w lemma="stopnica" msd="Ncfpd">stopnicam</w>
<c>.</c>
</s>
```

Oblikoslovne oznake podajo besedno vrsto ter množico drugih oblikoskladenjskih lastnosti besed; tako npr. oznaka `Ncmpi` pomeni `PoS: Noun, Type: common, Gender: masculine, Number: plural, Case: instrumental`, oziroma `Samostalnik, vrsta: občno_ime, spol: moški, število: množina, sklon: orodnik`.

Podroben opis oznak za slovenščino je podan v (Erjavec, 1998).

V romanu smo vzeli dele I–III "1984" za učno množico, medtem ko nam je IV. del, "Načela Novoreka — Dodatek", ki zajema približno 15 % romana, služil za ovrednotenje sistemov. Dodatek ima precej drugačno strukturo in besedni zaklad kot ostali deli romana, zato predstavlja, kljub temu da je del istega besedila kot učna množica, že razmeroma zahtevno testno množico.

Pri poskusih, ki jih opišemo v tem članku, smo se omejili na samostalnike in pridevnike, saj spada velika večina novih pregibnih besed, tj. takih, ki se bodo pojavile v novih besedilih kljub uporabi velikega leksikona in jih je potrebno lematizirati, v ti dve besedni vrsti. Za lažjo predstavo o obsežnosti in sestavi naših podatkov podamo v Tabeli 1 porazdelitev samostalnikov in pridevnikov (v osnovniku) za celoten roman ter za učno in testno podmnožico le-tega. Zadnji del tabele (IV-neznane) poda števila za neznane besede. Kot neznane obravnavamo tiste samostalnike in pridevnike, ki se pojavijo v testni množici, njihova lema pa se ne pojavi v učni množici.

Prvi stolpec v tabeli poda število vseh besed v besedilu, drugi število različnih besednih oblik in četrti število gesel. Peti stolpec poda število različnih oblikoslovnih oznak; samostalniki v romanu jih imajo 74, neznani samostalniki pa samo 31. Zadnji stolpec pove, koliko besed v besedilu se je pojavilo v obliki leme; pri teh je lematizacija trivialna; tako služijo kot mera najmanjše možne natančnosti sistema. Kot vidimo, bi program, ki vsako besedno obliko neznanega samostalnika ali pridevnika. kar proglasi za lemo, dosegel natančnost 111/279, kar je približno 40 % vseh pojavitev.

3. Oblikoslovna analiza

V tem poglavju opišemo, kako smo uporabili korpus za učenje in ovrednotenje pravil za oblikoslovno analizo slovenskih pridevnikov in samostalnikov. Uporabili smo sistem za induktivno logično programiranje (ILP), ki se nauči odločitvene sezname prvega reda. Ti seznami v programskem jeziku Prolog implementirajo urejene sezname pravil. Takšna, po specifičnosti urejena pravila, se že dolgo uporabljajo v generativnih pristopih h glasoslovju in k oblikoslovju.

Prvi sistem za učenje odločitvenih seznamov prvega reda je bil FOIDL, ki je bil uporabljen na problemu učenja tvorjenja preteklika angleških glagolov (Mooney and Califf, 1995); to je bil obenem eden prvih poskusov uporabiti ILP pri obdelavi naravnega jezika. Naloga je bila, naučiti se logičnega programa, ki bo definiral relacijo `past(PresentVerb,PastVerb)`, kjer sta argumenta ortografska reprezentacija nekega glagola v sedanjiku in pretekliku, pri čemer je `PresentVerb` vhodni, `PastVerb` pa izhodni argument. Program se uči iz primerov, kot so `past([b,a,r,k],[b,a,r,k,e,d])]` in `past([g,o],[w,e,n,t])`.

V našem pristopu za indukcijo pravil za oblikoslovno analizo slovenskih besed smo uporabil sistem za induktivno logično programiranje CLOG (Manandhar et al., 1998), ki se tako kot FOIDL uči odločitvenih seznamov prvega reda. Naši poskusi so pokazali, da je CLOG bistveno bolj učinkovit, zaradi česar ga lahko učimo na večjih podatkovnih množicah in s tem dosežemo večjo natančnost kot s programom FOIDL.

Iz korpusa smo vzeli trojke, ki vsebujejo besedno obliko (`bo`), lemo (`le`) in oblikoslovno oznako (`oo`). Vsaka trojka je tako primer analize v obliki `oo(bo, le)`, pri čemer je `bo` vhodni in `le` izhodni argument. Za vsako oblikoslovno oznako se je tako potrebno naučiti svojo množico pravil. Pri zapisu teh trojk najprej spremenimo vse črke v male ter besedno obliko in lemo zapišemo kot seznam znakov, z neASCII znaki kodiranimi kot entitete SGML. Tako npr. trojka *članki*/*članek*/`Ncmpn` predstavlja relacijo za samostalnik moškega spola v imenovalniku množine in generira sledeči primer:

```
n0mpn([ccaron,l,a,n,k,i],[ccaron,l,a,n,e,k]).
```

Nekateri atributi oblikoslovnih oznak nimajo (skoraj) nobenega vpliva na pregibanje besed, zato zanemarimo njihove vrednosti v predikatih in to označimo z `0`. Primer je viden zgoraj, kjer smo zanemarili razliko med občnimi (`Nc` in lastnimi `Np`) samostalniki; oboji so označeni z `n0`. S tem dobimo posplošene oblikoslovne oznake, kot je npr. `n0mpn` zgoraj. Vsaka posplošena oznaka predstavlja ciljni predikat, ki se ga je potrebno naučiti iz primerov, vsebovanih v učnem delu korpusa.

Kot primer naučene relacije si poglejmo pravila, ki jih inducira CLOG za analizo rodilnika ednine samostalnikov ženskega spola, `n0fsg`. Učna množica za ta koncept vsebuje 608 primerov, iz katerih se je CLOG naučil 13 pravil za analizo. Devet od teh so leksikalne izjeme, ki niso zanimive v kontekstu lematizacije neznanih besed. Ostanejo štiri pravila:

```
n0fsg(A,B):-
  rule(A,B,[t,v,e],[t,e,v]),!.
n0fsg(A,B):-
```

Vir	Vrsta	Pojavitev	Oblik	Gesel	Oznak	=
	Samostalnik	19398	6282	3199	74	7408
I-IV	Pridevnik	7462	3932	1936	121	1207
	Oboje	26860	10214	5135	195	8615
	Samostalnik	18438	6043	3079	74	7049
I-III	Pridevnik	7019	3731	1858	120	1124
	Oboje	25457	9774	4937	194	8173
	Samostalnik	960	533	379	51	359
IV	Pridevnik	443	347	245	62	83
	Oboje	1403	880	624	113	442
	Samostalnik	187	144	127	37	85
IV - neznane	Pridevnik	92	82	72	31	26
	Oboje	279	226	199	68	111

Tabela 1: Porazdelitev samostalnikov in pridevnikov v 1984: vsi, učni, testni in neznani.

```
  rule(A,B,[e,z,n,i],[e,z,e,n]),!.
n0fsg(A,B):-
  rule(A,B,[i],[]),!.
n0fsg(A,B):-
  rule(A,B,[e],[a]),!.
```

Če začnemo spodaj, torej pri najbolj splošnih pravilih, zajame zadnje pravilo samostalnike, ki pripadajo prvi ženski sklanjatvi, kjer končnico leme *-a* nadomesti v rodilniku ednine končnica *-e*. Predzadnje, tretje pravilo zajema samostalnike druge ženske sklanjatve, kjer ima lema ničto končnico, rodilnik ednine pa končnico *-i*. Drugo pravilo poskuša pokriti samostalnike druge sklanjatve, pri katerih prihaja do premene osnove zaradi neobstojnega *-e-*, kot npr. *bolezn-0*, *bolezn-i*. Prvo pravilo poskuša modelirati podobno premeno, ki pa zadeva samo samostalnike, ki se končujejo na *-(t)ev*. Kot vidimo, poskušajo pravila pokriti kanonične sklanjatve in premene, ki pa so podobne tistim, ki so se pojavili v učni množici; tako bo npr. analiza *lestve* vrnila *lestev*, medtem ko bo *bukve* analizirana kot *bukva*.

Za ovrednotenje delovanja naučenih pravil analize smo le-te preizkusili na Dodatku. Za vsako besedo v Dodatku smo vzeli besedno obliko in njeno oblikoslovno oznako ter poskusili s pomočjo pravil priti do pravilne leme; če se dobljena lema ne ujema z zapisano v korpusu, je to napaka. Tabela 2 povzame rezultat tega poskusa, kjer tudi ločimo natančnost na znanih in neznanih besedah.

	Vse	Znane	Neznane
Samostalniki	97.5 %	99.1 %	90.9 %
Pridevniki	97.3 %	96.6 %	100.0 %
Oboje	97.4 %	98.3 %	93.9 %

Tabela 2: Preizkus pravilnosti oblikoslovne analize.

Mogoče je presenetljivo, da je natančnost pri znanih besedah manj kot 100 % in da je pri pridevnikih za znane celo manjša kot za neznane. Razlog se skriva v definiciji '(ne)znane besede': besedna oblika je znana, če se njena lema pojavi v učni množici, kar pa ne pomeni nujno, da je bila videna z vsemi oblikoslovnimi oznakami, npr. s tistimi, v katerih se pojavi v testni množici.

Z naučenim modelom je sedaj mogoče določiti lemo (neznanega) samostalnika oz. pridevnika, če poznamo besedno obliko in njeno oblikoslovno oznako. Besedno obliko seveda najdemo v besedilu, naslednje poglavje pa obravnava določevanje oblikoslovnih oznak neznanih besed v besedilu.

4. Oblikoslovno označevanje

Oblikoslovno označevanje, t. i. *part-of-speech tagging* oz. bolje *syntactic wordclass tagging* (van Halteren, 1999), je v zadnjem desetletju postalo izjemno aktivno področje raziskav. Na kratko povedano, naloga oblikoslovnih označevalnikov je besedam v besedilu pripisati njihovo pravilno, kontekstno odvisno oblikoslovno oznako. V večini primerov se označevalniki naučijo modela jezika iz učnega korpusa, v katerem je bila predhodno vsaka beseda pravilno označena. Ta model jim nato omogoča določiti oblikoslovne oznake besed v novih besedilih z določeno stopnjo natančnosti. Za naše poskuse smo potrebovali natančen, hiter in prilagodljiv označevalnik. Poleg tega je moral biti sposoben delati z veliko množico (preko 1000) oblikoslovnih oznak MULTEXT-East za slovenščino. Pomembno je bilo tudi, da je označevalnik sposoben označevati neznane besede, tj. tiste, ki jih ni predhodno srečal v učni množici oz. niso bile vsebovane v leksikonu označevalnika. V (Džeroski et al., 2000) smo ovrednotili večje število dostopnih označevanikov na slovenskem '1984'; najboljše rezultate smo dosegli s programom TnT (Brants, 2000).

Učni korpus smo konvertirali v format za učenje TnT, kjer je vsakemu 'delcu' besedila (besedi ali ločilu) pripisana oznaka. Oznake besed so bile kar njihove oblikoslovne oznake, medtem ko so ločila označena sama s seboj. S tem dobimo nabor za označevanje, ki vsebuje 1024 oznak: ena za konec stavka, 13 za ločila in 1010 oblikoslovnih oznak.

Model, ki ga označevalnik TnT zgradi ob učenju vsebuje tabelo n-gramov (n=1,2,3) oznak in leksikon besednih oblik skupaj z njihovimi oznakami in številom pojavitev. Tabela n-gramov za naš učni korpus I–III vsebuje 1024 uni-, 12293 bi- in 40802 trigramov; primer sedmih trigramov podamo spodaj:

```
  Vcip3s--n       82
         Afpmsnn 17
         Aopmsn   2
         Ncmsn   12
         Npmsn    1
         Css      2
         Afpnpa   1
         Q        3
...
```

Primer interpretiramo tako, da se je oznaka `Vcps-sma` (*bil*) v učnem korpusu pojavila 544-krat. Od tega ji je 82-krat sledila oznaka `Vcip3s--n` (*je*). Trigram `Vcps-sma`, `Vcip3s--n`, `Afpmsnn` (*mrzel*, *vetroven*) se je pojavil 17-krat, trigram z `Aopmsn` na koncu dvakrat, itd.

Naučeni leksikon ima 15786 vnosov; spodaj podamo nekaj primerov:

```
...
juhe      2      Ncfsg      2
julij     1      Npmsn      1
julija   59      Npfsn     58   Npmsa--y  1
julije    4      Npfsg      4
juliji   10      Npfsd     10
julijin   4      Aspmsa--n  2   Aspmsn    2
...
```

Primer leksikona pove, da se je besedna oblika `juhe` pojavila v korpusu dvakrat; v obeh primerih je bila označena kot `Ncfsg`. Besedna oblika `julijin` pa se je pojavila štirikrat in je bila dvakrat označena kot tožilnik s podsplom neživo `Aspmsa--n` in dvakrat kot imenovalnik `Aspmsn`.

Natančnost označevalnika TnT smo nato ovrednotili na Dodatku; rezultati so podani v tabeli 3. V poskusih nismo uporabili dodatnega (referenčnega) leksikona in nismo spreminjali vnaprej nastavljenih parametrov za glajenje. Takšne izboljšave bi precej izboljšale natančnost označevanja na znanih, posledično pa verjetno tudi na neznanih besedah.

	Natančnost	OK/*
Vse besede	82.5 %	3260/692
Znane	84.3 %	3032/565
Neznane	64.2 %	228/127

Tabela 3: Preizkus TnT na Dodatku.

V tabeli podamo natančnost v odstotkih na pojavitev besed v besedilu (word token), sledi pa število pravilno označenih in število napačno označenih. Natančnost je razmeroma nizka, še posebej pri neznanih besedah.

V tabeli 4 se nato osredotočimo na samostalnike in pridevnike, kjer vidimo, da natančnost pada še naprej z najslabšim rezultatom 58.3 % za neznane samostalnike.

Ob podatku, da je skoraj vsaka druga neznana beseda napačno označena, so upravičeni dvomi, da bo natančnost lematizacije s kombinacijo slabega označevanja in ravno tako nepopolne analize še manjša in s tem praktično neuporabna. Kot bomo videli v naslednjem poglavju, se to na srečo ne zgodi.

	Vrsta	Natančnost	OK/*
Vse	Samostalniki	73.8 %	708/252
	Pridevniki	62.3 %	276/167
	Oboje	70.1 %	984/419
Znane	Samostalniki	77.5 %	599/174
	Pridevniki	60.7 %	213/138
	Oboje	72.2 %	812/312
Neznane	Samostalniki	58.3 %	109/ 78
	Pridevniki	68.4 %	63/ 29
	Oboje	61.6 %	172/107

Tabela 4: Preizkus TnT na samostalnikih in pridevnikih.

5. Lematizacija

V prejšnjih poglavjih smo opisali podatkovno množico, učenje pravil oblikoslovne analize in označevanja ter ločen preizkus natančnosti teh pravil. V tem poglavju podamo rezulate lematizacije neznanih besed, ki jih dosežemo s kombinacijo obeh metod. Opišemo dva preizkusa: prvi je, kot doslej, na Dodatku romana, drugi pa na pravnem besedilu s področja pridruževanja Slovenije EU.

V prvem preizkusu smo najprej oblikoslovno označili Dodatek s TnT, nato pa uporabili dobljene oznake za oblikoslovno analizo samostalnikov in pridevnikov. Kot smo pojasnili v prejšnjih poglavjih, je oblikoslovno označevanje pravilno v 87.5 % primerov za znane in v 61.2 % primerov za neznane samostalnike in pridevnike, medtem ko je analiza pravilna v 98.3 % za znane in 93.9 % za neznane. Natančnost lematizacije s kombinacijo obeh metod je podana v tabeli 5.

	Vrsta	Natančnost	OK/*
Vse	Samostalniki	91.7 %	880/ 80
	Pridevniki	87.6 %	388/ 55
	Oboje	90.4 %	1268/135
Znane	Samostalniki	95.4 %	738/ 35
	Pridevniki	88.0 %	309/ 42
	Oboje	93.1 %	1047/ 77
Neznane	Samostalniki	75.9 %	142/ 45
	Pridevniki	85.9 %	79/ 13
	Oboje	79.2 %	221/ 58

Tabela 5: Rezultati lematizacije na Dodatku 1984.

Kot se vidi iz tabele, bo v povprečju od petih novih neznanih samostalnikov ali pridevnikov en lematiziran napačno. Analiza napak pokaže, da je pri večini TnT označil samostalnik ali pridevnik kot neko drugo besedno vrsto. To se zgodi v 78 primerih, kar je 58 % vseh napak. Očitno je težko ugotoviti že to, ali je neka neznana beseda samostalnik ali pridevnik. Označevalnik je kritična komponenta sistema, čeprav je analiza pogosto neobčutljiva za njegove napake; v precejšnjem številu primerov (245 znanih, 53 neznanih) je napovedana lema besede pravilna, kljub temu da je pripisana oblikoslovna oznaka napačna. To niti ni tako presenetljivo, saj se je TnT pri neznanih besedah največkrat zmotil ravno med sinkretičnimi oblikami (npr. imenovalnik = rodilnik za žive oz. = tožilnik za nežive

N*ms), ki pa ne vplivajo na pravilnost lematizacije.

Čeprav je Dodatek 1984 že precej različen od ostalega romana, ki je bil uporabljen za učenje, smo vseeno hoteli preizkusiti sistem še na povsem drugačnem besedilu in tako priti do boljše ocene o trdoživosti predstavljene metode. V ta namen smo vzeli besedilo s polnim naslovom "Evropski sporazum o pridružitvi med republiko Slovenijo na eni strani in evropskimi skupnostmi in njihovimi državami članicami, ki delujejo v okviru Evropske unije na drugi strani; 10. junij 1996, Luksemburg." To besedilo je bilo zbrano in označeno kot del slovenko-angleškega vzporednega korpusa ELAN (Erjavec, 1999). Besedilo je sestavljeno iz 1,191 prevodnih enot, ki v glavnem ustrezajo stavkom, te pa vsebujejo 12,049 besed in 2,470 ločil.

Vendar pa besedilo v okviru projekta ELAN še ni bilo oblikoslovno označeno ali lematizirano. Da bi lahko besedilo uporabili kot testno množico, je bilo vsako besedo potrebno vsaj enoumno lematizirati. To smo storili v dveh korakih. Najprej nam je podjetje Amebis d. o. o. prijazno lematiziralo besedilo za besede, ki so znane njihovemu oblikoslovnemu analizatorju BesAna, ki vsebuje obširen leksikon slovenskega jezika. S tem je bila vsaka BesAni znana beseda lematizirana, vendar včasih tudi dvoumno. Zato smo v drugem koraku, deloma s programskimi filtri in deloma interaktivno, razdvoumili leme takšnih besed, s čimer smo dobili besedilo, v katerem je veliki večini besed pripisana njihova lema. V tem besedilu smo nato vzeli kot neznane besede tiste samostalnike in pridevnike, ki imajo sicer pripisano lemo, vendar pa se ta lema ni pojavila v učni množici. Nekaj potez te testne množice podamo v Tabeli 6; prvi stolpec poda kot prej število vseh pojavitev besed, druga dva stolpca pa dodajata informacijo o številu različnih besednih oblik in različnih gesel. Kot vidimo, je število neznanih besed približno trikrat večje kot v Dodatku, lema nekega neznanega samostalnika ali pridevnika pa se bo v povprečju pojavilo v besedilu 2.2-krat, v 1.3 besednih oblikah.

	Pojavitve	Besedne oblike	Leme
Znane	12049	3407	1672
Neznane	1458	863	644
- samostalniki in pridevniki	1322	796	595

Tabela 6: Porazdelitev besed v dokumentu EU.

Za preizkus sistema na tem korpusu smo vzeli že naučena modela za označevanje in analizo. Najprej smo oblikoslovno označili celotno besedilo, nato pa analizirali neznane samostalnike in pridevnike. Pri tem poskusu ne moremo več podati natančnosti posameznih komponent, saj korpus ni bil preverjeno označen z oblikoslovnimi oznakami.

Tabela 7 poda rezultate lematizacije neznanih samostalnikov in pridevnikov.

	Pojavitve	Besedne oblike	Leme
Natančnost	81.3 %	79.8 %	75.6 %
Vsi	1322	796	595
Pravilno	1075	635	450
Napaka	247	161	145

Tabela 7: Rezultati lematizacije neznanih samostalnikov in pridevnikov na dokumentu EU.

Prvi del tabele poda natančnost, za katero vidimo, da je z 81.3 % celo rahlo boljša kot 79.2 % na Dodatku. Napačne lematizacije smo računali za besedne oblike in leme tako, da se kot napaka šteje vsaka lematizacija, ki ne vrne pravilnega rezultata ob vsaki pojavitvi. Rezultati pri drugih dveh stolpcih so nekoliko slabši, kot če natančnost merimo za vsako pojavitev posebej; neznana lema v besedilu se v povprečju pojavi večkrat; očitno se sistem bolj moti pri takšnih besednih oblikah oz. lemah.

6. Zaključek

V prispevku smo obravnavali lematizacijo neznanih besed v besedilih na primeru samostalnikov in pridevnikov slovenskega jezika. Problema smo se lotili s kombiniranjem oblikoslovnega analizatorja in oblikoslovnega označevalnika, pri čemer sta bila jezikovna modela obeh komponent naučena induktivno iz vnaprej označenega korpusa MULTEXT-East, ki vsebuje približno 90,000 besed in 1,000 različnih oznak.

Uspešnost naučenega sistema smo ovrednotili na dveh testnih množicah; Dodatku romana '1984' in na besedilu o pridruževanju Slovenije k EU. V obeh primerih je natančnost lematizacije neznanih samostalnikov in pridevnikov okoli 80 %, če jo merimo na posamezno pojavitev. Trivialna natančnost, kjer kot lemo vedno proglasimo kar besedno obliko, je 40 %, tako da je tu izboljšanje očitno. Rezultate bi bilo bolje primerjati s kakšnim drugim sistemom za slovenski jezik, vendar kolikor nam je znano, podobne raziskave še niso bile opravljene.

Že pri trenutni stopnji natančnosti je sistem lahko koristen kot pomoč pri ustvarjanju in izboljševanju jezikovnih virov (leksikonov), ki so dobljeni iz besedilnih korpusov.

Obstaja več načinov, kako v okviru obstoječega sistema zvečati pravilnost lematizacije. Kritična komponenta je trenutno označevalnik, ki bi potreboval vsaj večji leksikon, pa tudi večjo učno množico.

Druga izboljšava bi bila obravnavati dokument kot celoto in najprej narediti leksikon neznanih besed; neznana lema se v povprečju v dokumentu pojavi okoli dvakrat; z izkoriščanjem večjega števila pojavitev iste besedne oblike bi lahko nakopičili več podatkov o neki lemi in se nato tudi bolje odločili, kako jo lematizirti.

Zahvale

Avtorja se zahvaljujeta anonimnim recenzentom za zelo koristne pripombe. Za programsko opremo gre zahvala Sureshu Manadharju za CLOG in Thorstenu Brantsu za TnT. Jezikovni viri, ki smo jih uporabili pri poskusih, so nastali v naslednjih projektih in institucijah: korpus '1984' je bil označen v okviru EU projekta Copernicus 106 MULTEXT-East; besedilo za preizkus sistema je prispeval SVEZ,

Služba Vlade RS za evropske zadeve; označeno je bilo v v okviru EU projekta MLIS ELAN; označevanje z lemami je prispevalo podjetje Amebis, d. o. o.

7. Literatura

Thorsten Brants. 2000. TnT — a statistical part-of-speech tagger. V: *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, WA. http://www.coli.uni-sb.de/~thorsten/tnt/.

Jean-Paul Chanod in Passi Tapanainen. 1995. Creating a tagset, lexicon and guesser for a French tagger. V: *Proceedings of the ACL SIGDAT workshop From Text to Tags: Issues in Multilingual Language Analysis*, Dublin.

Jure Dimec, Sašo Džeroski, Ljupčo Todorovski in Dimitar Hristovski. 1998. Iskalnik za slovenske in angleške dokumente na svetovnem spletu. V: Tomaž Erjavec in Jerneja Gros, ur., *Zbornik konference JEZIKOVNE TEHNOLOGIJE ZA SLOVENSKI JEZIK*, str. 42–48, Ljubjana.

Ludmila Dimitrova, Tomaž Erjavec, Nancy Ide, Heiki-Jan Kaalep, Vladimír Petkevič in Dan Tufiş. 1998. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. V: *COLING-ACL '98*, str. 315–319, Montréal, Québec, Canada. http://nl.ijs.si/ME/.

Sašo Džeroski, Tomaž Erjavec in Jakub Zavrel. 2000. Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. V: *Second International Conference on Language Resources and Evaluation, LREC'00*, str. 1099–1104, Paris. ELRA.

Tomaž Erjavec. 1998. The Multext-East Slovene Lexicon. V: *Proceedings of the 7th Slovene Electrotechnical Conference, ERK '98*, str. 189–192, Portorož, Slovenia. http://nl.ijs.si/et/Bib/ERK98/.

Tomaž Erjavec. 1999. Slovensko-angleški korpus ELAN. *Slavistična revija*, 47(4):515–522. http://nl.ijs.si/elan/.

Primož Jakopin in Aleksandra Bizjak. 1997. O strojno podprtem oblikoslovnem označevanju slovenskega besedila. *Slavistična Revija*, 45(3-4):513–532. http://www.ff.uni-lj.si/sr/index.html.

Suresh Manandhar, Sašo Džeroski in Tomaž Erjavec. 1998. Learning multilingual morphology with CLOG. V: David Page, ur., *Inductive Logic Programming; 8th International Workshop ILP-98, Proceedings*, št. 1446 v Lecture Notes in Artificial Intelligence, str. 135–144. Springer.

Andrei Mikheev. 1997. Automatic rule induction for unknown-word guessing. *Computational Linguistics*, 23(3):405–424.

Raymond J. Mooney in Mary Elaine Califf. 1995. Induction of first-order decision lists: Results on learning the past tense of English verbs. *Journal of Artificial Intelligence Research*, 3(1):1–24.

C. M. Sperberg-McQueen in Lou Burnard, ur. 1994. *Guidelines for Electronic Text Encoding and Interchange*. Chicago and Oxford.

Hans van Halteren, ur. 1999. *Syntactic Wordclass Tagging*. Kluwer.

Jure Zupan. 1999. Problemi in nekaj rešitev računalniških obdelav slovenskih besedil. *Slavistična revija*, 47(3). http://www.ff.uni-lj.si/sr/index.html.