Oblikoslovno označevanje slovenskega jezika: primer korpusa SVEZ-IJS

Tomaž Erjavec†, Bence Sárossy‡

†Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova 39, 1000 Ljubljana
tomaz.erjavec@ijs.si
‡Budimpešta, steksz@freemail.hu

Povzetek

Avtomatsko oblikoslovno označevanje (*part-of-speech tagging* oz. *word-class syntactic tagging*) je postopek, pri katerem se vsaki besedi, ki se v besedilu pojavi, pripiše oblikoslovna oznaka. Za slovenski jezik so raziskave te, za jezikovne tehnologije izredno pomembne, komponente še v povojih. V prispevku predstavljamo evalvacijo označevalnika TnT, ki je vgrajen v program totale in naučen modela označevanja na jezikovnih virih MULTEXT-East. Testna domena je prosto dostopen vzporedni angleško-slovenski korpus SVEZ-IJS, ki vsebuje pravni red Evropske unije. V prispevku opisujemo ročno popravljen vzorec iz korpusa in podrobno analiziramo napake pri označevanju. Predstavljamo tudi enostaven pretvorbeni program, ki popravi nekaj najbolj pogostih napak, in podajamo zaključke in smernice za nadaljnje delo.

Word-class Syntactic Tagging of Slovene: the case of the SVEZ-IJS corpus

Part-of-speech tagging or, more accurately, word-class syntactic tagging is a procedure that assigns to each word token appearing in a text its morphosyntactic description. Research on this important component of many language technology applications is, for the Slovene language, still at a preliminary stage. In the paper we evaluate the accuracy of the TnT tagger, as a part of the totale annotation tool, which had been trained on the MULTEXT-East language resources for Slovene. The test data come from the freely available parallel English-Slovene corpus SVEZ-IJS, which contains legal acts of the EU. Presented are the details of the manually corrected sample from the corpus and an analysis of the tagging errors. The paper also discusses a simple transformation-based program that fixes some of the more common errors, and concludes with some directions for future work.

1. Uvod

Oblikoslovno označevanje (*part-of-speech tagging* oz. *word-class syntactic tagging*) (van Halteren, 1999) je postopek, s katerim vsaki besedi, ki se v besedilu pojavi, pripišemo enolično oblikoslovno oznako. Programi za oblikoslovno označevanje morajo poljubnim besednim oblikam določiti možne oznake, nato pa izmed teh oznak izbrati pravo glede na kontekst, v katerem se besedna oblika pojavi. Tako ima npr. besedna oblika *hotel* tri možne oznake: dve samostalniški (ednina imenovalnika in tožilnika) in eno glagolsko (deležnik na -l moškega spola), v stavku *Šel je v hotel* pa mora biti ta pojavnica označena kot samostalnik v tožilniku.

Oblikoslovno označevanje je bilo najprej razvito za angleški jezik, v katerem so nabori oblikoslovnih oznak razmeroma majhni (~50, odvisno od specifike posameznih naborov oznak), problemi pa predvsem v dvoumnosti glede besedne vrste. Dosti kasneje so bili označevalniki razviti tudi za oblikoslovno bistveno bogatejše jezike, predvsem češčino (Hajič in Hladka, 1998). Pri pregibno bogatih jezikih, kot sta češčina ali slovenščina, je različnih oznak tudi po več tisoč, največji problem pa je, vsaj na prvi pogled, predvsem v razdvoumljanju sinkretičnih pregibnih oblik, torej v okviru besedne vrste.

Večina sodobnih oblikoslovnih označevalnikov se modela določenega jezika nauči, in to iz ročno označenih korpusov, po možnosti podprtih z oblikoslovnimi leksikoni. Takšni programi so sicer robustni (analizirajo tudi njim neznane besede oz. besedne zveze in so odporni na napake v besedilu), vendar pa delajo napake. Natančnost označevanja je odvisna od velikosti učne množice, konkretnega nabora oznak, besedilnozvrstne podobnosti testne in učne množice in seveda od izbranega algoritma za učenje in označevanje.

Tekst v slovenskem jeziku so verjetno prvi skušali oblikoslovno označiti na ZRC SAZU (Jakopin in Bizjak, 1997), in sicer z v urejevalnik EVA vgrajenim označevalnikom, ki deluje na osnovi pravil in velikega podpornega leksikona. Označevalnik je polavtomatski, saj se pri neznanih besedah in besedah, ki mu jih ni uspelo razdvoumiti, zanaša na intervencijo človeka. Zato je koristen predvsem za izdelavo ročno označenih korpusov, za kar je bil do sedaj tudi uporabljan.

Na tem mestu velja omeniti še avtomatsko označevanje korpusa FidaPLUS (http://www.fidaplus.net/), ki ga je razvilo podjetje Amebis, d.o.o. Vendar pa je ta označeni korpus še v nastajanju, o označevanju pa, kolikor nam je znano, še ni bilo nobene publikacije

Na IJS so poskusi z avtomatskim označevanjem vezani na nabor oblikoslovnih oznak, rabljenih pri projektu MULTEXT-East (Erjavec, 2004, http://nl.ijs.si/ME/), in na ročno označen korpus tega projekta, tj. Orwellov roman »1984«. Prvi poskusi (Erjavec et al., 2000) so pokazali, da ima od štirih preizkušenih javno dostopnih označevalnikov najboljše lastnosti statistični označevalnik TnT (Brants, 2000), ki je dosegel stopnjo natančnosti 89,2%, pri čemer je bila testna množica ravno tako iz romana »1984«. V članku smo analizirali tudi natančnost označevanja po besednih vrstah, ki je bila za TnT 96,6%. V nadaljevanju dela z avtomatskimi označevalniki na IJS (Erjavec in Džeroski, 2004) smo označevanje rahlo izboljšali in implementirali ter evalvirali tudi lematizacijo neznanih besed, tj. pripisa osnovne oblike vsaki besedi v besedilu, npr. *hotela* → *hotel* oz. *hoteti*, odvisno od konteksta. Problema sta povezana, saj je naša

implementacija lematizacije izrazito odvisna od oblikoslovnega označevanja.

Kolikor nam je znano, je edina druga raziskava, ki je imela namen evalvirati avtomatsko oblikoslovno označevanje slovenskega jezika, Lönneker (2005), ki opisuje uporabo označevalnika TreeTagger (Schmid, 1994) na ročno označenem korpusu ZRC SAZU, podaja pa tudi primerjavo z rezultati iz raziskave Erjavec et al. (2000). K primerjavi rezultatov raziskave Lönneker (2005) z našimi rezultati se bomo vrnili v razdelku 5.

V pričujočem članku ponavljamo evalvacijo označevanja iz raziskave Erjavec et al. (2000) v razširjeni obliki na novem korpusu, in sicer na vzorcu iz slovenskega dela korpusa SVEZ-IJS (Erjavec, 2006). Zanimalo nas je, kakšna je natančnost označevanja na korpusu, ki je po zvrstnosti zelo drugačen od učnega, katere napake so najbolj pogoste in škodljive in ali jih je mogoče, in do katere mere, na enostaven način odpraviti.

V nadaljevanju naprej predstavljamo program totale (Erjavec et al., 2005), ki ga uporabljamo za avtomatsko označevanje, nato korpus SVEZ-IJS in testni vzorec iz tega korpusa, ki smo ga ročno popravili. Temu sledi analiza napak pri avtomatskem označevanju, opis pretvorbenega programa, ki popravi najbolj pogoste napake, primerjava stopenj natančnosti, doseženih pri posameznih eksperimentih, ter nekaj zaključkov.

2. Testni podatki

Podatki, ki smo jih uporabili za evalvacijo, so sestavljeni iz vzorca korpusa SVEZ-IJS. Vzorec je bil avtomatsko označen s programom totale in nato ročno pregledan in popravljen.

2.1. Označevanje s totale

Za jezikoslovno označevanje uporabljamo program totale (tokenisation, tagging, and lemmatisation) (Erjavec et al., 2005), ki:

1. Besedilo tokenizira, torej razdeli na besede, ločila in povedi (z v totale vgrajenim programom mlToken).
2. Besede oblikoslovno označi (z v totale vgrajenim programom TnT) (Brants, 2000).
3. Besedilo lematizira (z v totale vgrajenim programom CLOG, upoštevajoč pripisane oblikoslovne oznake).

Takó oblikoslovno označevanje kot lematizacija se izvajata s programi, ki se modela jezika naučijo iz vnaprej pripravljenih podatkov, torej iz ročno označenega korpusa in oblikoslovnega leksikona. Naš model oblikoslovnega označevanja za slovenski jezik je naučen iz korpusa MULTEXT-East, tj. romana »1984« G. Orwella (100.000 pojavnic), ter majhnega vzorca iz korpusa IJS-ELAN (5.000 pojavnic), lematizator pa je naučen na oblikoslovnem slovarju MULTEXT-East (polne paradigme 15.000 lem), ki smo ga uporabili tudi za boljše delovanje označevalnika (Erjavec in Džeroski, 2004).

Korpus in leksikon vsebujeta oblikoslovne oznake MULTEXT-East za slovenski jezik, in sicer se v korpusu pojavi 1.023 različnih oznak, v leksikonu pa je predvidenih kar 2.083 oznak. Oznake so izvorno angleške[1] (tj. sestavljajo jih prve črke angleških besed) in so sestavljene iz niza črk, pri čemer prva črka označuje besedno vrsto, ostale pa, glede na besedno vrsto, vrednosti njenih atributov. Tako npr. oznaka *Ncmsn* pomeni: *PoS = noun, type = common, gender = masculine, number = singular, case = nominative*, ekvivalentna slovenska oznaka *Somei* pa: *besedna vrsta = samostalnik, vrsta = občno ime, spol = moški, število = ednina, sklon = imenovalnik*.

2.2. Korpus SVEZ-IJS

Vzporedni angleško-slovenski korpus SVEZ-IJS, http://nl.ijs.si/svez/ (Erjavec, 2006), zajema pravni red EU, t. i. Acquis Communautaire. Inačica 1.0 tega korpusa vsebuje 2 x 5 milijonov besed, nastala pa je leta 2004 na osnovi tedanjega pomnilnika prevodov prevajalske skupine pri SVEZ, Službi vlade RS za evropske zadeve (Erbič et al., 2005). Korpus je sestavljen iz poravnanih segmentov v angleškem in slovenskem jeziku, ki tipično vsebujejo eno poved ali del povedi. Korpus je zanimiv iz več razlogov:

- Je velik in vsebuje kvalitetno poravnane povedi v za nas najbolj aktualnem jezikovnem paru.
- Vsebuje (skoraj) identična besedila kot JRC-Acquis (Steinberger, 2006), v katerega pa je dodatno vključenih še 19 drugih jezikov, vendar pa ima korpus slabšo poravnavo.
- Je prosto dostopen za raziskovalne namene.

Ker je SVEZ-IJS torej izredno koristna podatkovna množica za raziskave in razvoj jezikovnih tehnologij za slovenski jezik, se nam je zdelo koristno preučiti napake v njegovih oblikoslovnih oznakah in poskusiti zmanjšati njihovo število.

2.3. Vzorec

Za evalvacijo označevanja smo iz avtomatsko označenega korpusa najprej izdvojili vzorec, v katerega smo vključili po 3 zaporedne slovenske segmente na vsakih 1.000 segmentov, s čimer smo zajeli približno 3 ‰ slovenskega besedila v korpusu. Ta vzorec je bil potem izvožen v Excelovo tabelo in tam ročno popravljen, pri čemer smo obdržali prvotne, avtomatsko pripisane oznake. S tem smo dobili testno množico, iz katere izvirajo vse statistike v pričujočem članku.

Enota	n	Razmerje	
Znakov	513.650		A
Segmentov	821	625 A/B	B
Vseh pojavnic	15.765	19 C/B	C
Ločil (pojavnic)	2.346	15% C	Č
Besed (pojavnic)	13.419	85% C	D
Besed (različnic)	5.189	2,59 D/E	E
Lem (različnic)	3.062	4,38 E/F	F
Oblik. oznak (različnic)	452	29,69 D/G	G

Tabela 1: Testni podatki, osnovna statistika.

Podrobna analiza velikosti vzorca je podana v Tabeli 1, kjer v stolpcu *n* npr. vidimo, da vsebuje vzorec približno pol milijona znakov in nekaj več kot 15.000 pojavnic, od katerih je skoraj šestina ločil, ter da je

[1] Oznake je možno avtomatsko preslikati oz. prevesti v slovenski jezik, in take se uporabljajo v korpusu FIDA.

različnih besednih oblik okoli 5.200, lem pa 3.000. Pri tem se za različne besedne oblike štejejo ortografsko različne oblike, tako da sta npr. *tudi* in *Tudi* dve različnici, vse leme pa so zapisane v malih črkah, zato imata besedni obliki *koren* in *Koren* enako lemo. Končno nas vrstica G opozarja na (presenetljivo majhno) število oblikoslovnih oznak v vzorcu, okoli 450.

Stolpec *Razmerja* podaja ulomke različnih (v stolpcu *n* prikazanih) števil in nas s tem seznanja npr. s povprečno dolžino segmenta v znakih (625) oz. pojavnicah (19) ali pa prikazuje, koliko različnih besednih oblik v povprečju pokriva ena lema (4.4).

	n	% vseh	% besed
Besed	13.419	85,1%	100%
(Znanih)	10.996	69,7%	81,9%
(Neznanih)	2.423	15,4%	18,1%
Samostalnik (N)	4.928	31,3%	36,7%
Glagol (V)	1.287	8,2%	9,6%
Pridevnik (A)	1.694	10,7%	12,6%
Prislov (R)	373	2,4%	2,8%
Števnik (M)	795	5,0%	5,9%
Zaimek (P)	743	4,7%	5,5%
Veznik (C)	1.102	7,0%	8,2%
Predlog (S)	1.787	11,3%	13,3%
Členek (Q)	107	0,7%	0,8%
Okrajšava (Y)	474	3,0%	3,5%
Neuvrščene (X)	128	0,8%	1,0%

Tabela 2: Testni podatki, znane/neznane besede in statistika po besednih vrstah.

V Tabeli 2 je bolj podrobno predstavljeno razmerje med besedami. Stopnja napake pri avtomatskem označevanju je seveda zelo odvisna od tega, ali se je neka besedna oblika v učni množici oz. leksikonu pojavila ali pa je sistemu povsem neznana. V tabeli vidimo, da neznane besede obsegajo okoli 15% vseh pojavnic in več kot 18% vseh besed, kar lepo kaže na leksikalno različnost korpusa SVEZ-IJS od korpusa MULTEXT-East oz. na pomanjkljivo in premajhno učno množico. Tabela podaja še statistiko po (ročno določenih, torej pravilnih) besednih vrstah, ki kaže, da je v besedilu največ samostalnikov, pridevnikov in predlogov, ki skupaj tvorijo kar polovico pojavnic. To pomeni, da je za visoko stopnjo natančnosti pomembna predvsem pravilna interpretacija teh treh besednih vrst, predvsem samostalnikov.

Posebno pozornost si zaslužita zadnji dve kategoriji v razpredelnici, in to iz dveh razlogov. Neuvrščenih besed (X) in okrajšav (Y) namreč ne kategorizirata po besednih vrstah in sta zato nekakšni metakategoriji. Tako so okrajšave z oblikoskladenjskega vidika tipično samostalniki, poleg tega pa kategorija Y zajema tudi kratice, ki lahko skladenjsko delujejo kot celotne fraze, kot npr. *itd.* Z X (tj. kot neuvrščene besede) označujemo tujejezične besede, ki jih je običajno po več skupaj, s skladenjskega vidika pa delujejo kot samostalniške fraze, npr. *Carte de séjour de résident privilégié de Monaco*. Druga posebnost kategorij, označenih z X in Y, pa je število pojavnic, ki jih zajemata, saj take pojavnice v korpusu SVEZ-IJS tvorijo razmeroma velik del besedila, 4,5%. Kot bomo videli, sta ti dve kategoriji odgovorni za velik delež napak pri avtomatskem označevanju.

3. Analiza avtomatskega označevanja

Na osnovi ročno popravljenega vzorca smo nato evalvirali natančnost avtomatskega označevanja s totale/TnT. Ob tem je pomembno ločiti dve vrsti napak: pri prvi vrsti je oblikoslovna oznaka sicer napačna, vendar je pravilno določena vsaj besedna vrsta, pri drugi vrsti pa je napačna tudi oznaka besedne vrste. Program pri prvi vrsti napak npr. napačno določi sklon ali število (ali oboje) ali pa (pod)vrsto besedne vrste, npr. lastna imena proti občnim imenom, kakovostni pridevniki proti vrstnim pridevnikom, pri drugi vrsti napak pa npr. proglasi samostalnik za glagol ali pa neuvrščeno besedo za samostalnik. Razlikovanje med temi napakami je koristno zato, ker je za marsikatero potencialno uporabo oz. uporabnika oblikoslovnih oznak pomembna samo besedna vrsta, ne pa npr. pregibne lastnosti. Zanimivo je ločiti tudi natančnost označevanja za znane in neznane besede, torej tiste, ki jih sistem v učni množici ni srečal. V Tabeli 3 predstavljamo absolutno število napak v oznaki in besedni vrsti, podajamo pa tudi natančnost lematizacije. Za vsako kategorijo (oz. število) prikazujemo ustrezno natančnost glede na vse pojavnice in glede na besedne pojavnice v vzorcu.

	n	Točnost pojavnice	Točnost besede
Napačna oznaka	1.799	88,6%	86,6%
Za znane besede	950	92,9%	91,4%
Za neznane besede	849	65,0%	65,0%
Napačna bes. vrsta	748	95,3%	94,4%
Za znane besede	155	98,8%	98,6%
Za neznane besede	593	75,5%	75,5%
Napačna lema	220	98,6%	98,4%
Za znane besede	88	99,3%	99,2%
Za neznane besede	132	94,6%	94,6%
Za napačno oznako	217	87,9%	87,9%
Za pravilno oznako	3	99,8%	99,8%

Tabela 3: Točnost avtomatskega označevanja.

Po najbolj strogi oceni ima sistem natančnost 86,6%. Tako nizka je predvsem zaradi precejšnjega števila neznanih besed. Natančnost glede na besedno vrsto je precej večja, 94,4%, je pa tudi tu napaka pri neznanih besedah bistveno večja kot pri znanih. Lematizacija ima celo večjo natančnost, 98,4%, kot označevanje besedne vrste, saj veliko napak označevalnika na pravilno lematizacijo ne vpliva, je pa natančnost za neznane besede zopet precej manjša. V zadnjih dveh vrsticah vidimo, da so za napačno delovanje lematizatorja skoraj v celoti krive napačne oblikoslovne oznake: samo v treh primerih se zgodi, da je oblikoslovna oznaka pravilna, lema pa ne.

3.1. Napake med besednimi vrstami

Zaradi pomembnosti napak pri označevanju besedne vrste bomo analizo le-teh ločili od tistih napak, ki nastanejo pri označevanju v okviru besedne vrste.

V Tabeli 4 podamo matriko števila napak glede na dejansko besedno vrsto (vodoravno) in besedno vrsto, ki jo je pojavnici pripisal označevalnik (navpično). Števila v diagonali tako predstavljajo napake, ki se zgodijo pri označevanju v okviru besedne vrste, ostale pa kažejo, katere zamenjave so najbolj pogoste, na primer, da so bili samostalniki petindevetdesetkrat označeni kot glagoli.

Tabela kaže, da je označevanje (glede na nabor) odprtih besednih vrst (zapisane so v krepkem tisku) bistveno manj uspešno kot označevanje funkcijskih besed, kar je razumljivo, saj je večina slednjih označevalniku poznana. Do neke mere je izjema zaimek, pa še ta samo glede stopnje napak pri označevanju v okviru besedne vrste. Razlog za majhno absolutno natančnost oznak za zaimek je, da imajo oblikoslovne oznake te besedne vrste izrazito razvejano strukturo, tako da zajemajo zaimki skoraj polovico (čez tisoč) vseh oblikoslovnih oznak, in da so besedne oblike vseh zaimkov sicer vsebovane v leksikonu, ne pa tudi v učnem korpusu, zato so napake pri sinkretičnih pregibanjih pogoste.

	N	***V***	***A***	***R***	*M*	*P*	*C*	*S*	*Q*	***I***	***X***	***Y***	*
N	**609**	6	9	4	47	0	1	1	0	0	69	241	987
V	95	**18**	2	1	28	2	2	0	0	0	35	17	200
A	28	1	**275**	12	8	3	0	0	0	0	14	9	350
R	14	1	4	**15**	0	1	1	1	0	0	6	11	54
M	0	0	1	0	**11**	6	0	0	0	0	0	18	36
P	1	0	1	0	2	**105**	0	0	0	0	0	1	110
C	0	1	0	3	0	0	**0**	0	10	0	0	11	25
S	1	0	0	0	1	0	0	**18**	0	0	1	6	27
Q	0	0	0	3	0	0	2	0	**0**	0	0	0	5
I	1	0	0	0	0	0	0	0	0	**0**	1	0	2
X	0	0	0	0	0	0	0	0	0	0	**0**	0	0
Y	0	0	0	0	1	0	0	0	0	0	2	**0**	3
*	749	27	292	38	98	117	6	20	10	0	128	314	1799

Tabela 4: Pregled napak pri določitvi besednih vrst.
N = samostalnik, V = glagol, A = pridevnik, R = prislov, P = zaimek, S = predlog, C = veznik, Q = členek, I = medmet, M = številnik, Y = okrajšava, X = neuvrščeno

Največkrat je napačna besedna vrsta pripisana samostalnikom, številnikom, neuvrščenim besedam in okrajšavam. Pri samostalnikih je relativna napaka sicer majhna, vendar je zaradi velikega števila samostalnikov za celotno natančnost vseeno pomembna. Posebej je opazna zamenjava samostalnika z glagolom, ki deloma verjetno izhaja tudi iz narave učne množice.

Pri ostalih treh besednih vrstah, namreč številnikih, neuvrščenih besedah in okrajšavah, so razlogi za veliko število napak podobni. Besede so po eni strani skoraj vedno neznane (še posebej zato, ker jih niti učni korpus niti leksikon tako rekoč ne vsebujeta, z izjemo omejenega števila številnikov), po drugi strani pa so oblikoskladenjsko slabo definirane, kar še posebej velja za neuvrščene besede in okrajšave. K temu problemu se bomo vrnili v razdelkih 4 in 5.

3.2. Napake v okviru besednih vrst

V tem razdelku bomo bolj podrobno pogledali najbolj pogoste napake pri označevanju v okviru besedne vrste. Kot kaže Tabela 4, je največ napak pri samostalnikih, ki imajo v oznakah MULTEXT-East pet atributov: vrsta, spol, število, sklon in živost. Kar 85% napak je vezanih na sklon, čeprav tudi v kombinaciji s številom, podobno pa je tudi pri pridevnikih in zaimkih, čeprav je pri teh večkrat napačen tudi spol. Bolj podrobna analiza teh napak je pokazala, da je v velikem številu primerov pravilno oznako pravzaprav nemogoče ugotoviti samo na osnovi lokalnega konteksta, torej takega, ki ga uporablja TnT. Zato bi bil za odpravljanje takšnih napak potreben drugačen pristop.

Ostale besedne vrste so za označevanje manj težavne. Pri glagolih sta najbolj problematični lastnosti spol in število.

4. Pretvorbeno označevanje s pravili

Ključno vprašanje seveda je, kako izboljšati natančnost označevanja. Kot preizkus v to smer smo implementirali program, ki popravi nekatere napake označevalca. V tem razdelku opisujemo delovanje programa in evalviramo rezultate.

Program, napisan v programskem jeziku Perl, vzame kot vhodne podatke že označeno besedilo, s čimer ima dostop do podatkov o obliki in tipu (ločilo, beseda) vsake pojavnice, za besede pa lahko preveri še pripisano (mogoče napačno) oblikoslovno oznako in frekvenco besedne oblike v učnih podatkih (0 = neznana beseda). Za označevanje vsake besedne pojavnice program nato sproži kaskado ročno napisanih pravil. Vsako pravilo je oblike »če *pogoj*, potem *pripiši oblikoslovno oznako,* drugače *naslednje pravilo*«. V pogojih uporabljamo funkcijo *feature*, ki vzame kot prvi argument lastnost, kot drugi pa pojavnico ter vrne vrednost te lastnosti za pojavnico. Kot primer podajamo prvi dve pravili:

```
        ...
①       elsif ($freq == 0  and feature("idwrd",$sent[$focus])=~/^[IVX]+$/)
          {$outmsd="Mc---r" }
②       elsif ($freq == 0 and
                feature("case",$sent[$focus]) eq 'uc' and
                not (feature("case",$sent[$focus-1])  eq  'uc' or
                     feature("case",$sent[$focus+1]) eq  'uc'))
          {$outmsd="Y"}
        ...
```

Prvo pravilo oblikoslovno oznako popravi v oznako `Mc---r`, ki označuje rimsko številko, saj je bilo napačno prepoznavanje le-teh eden večjih problemov pri označevanju številnikov. Pogoj določa, da mora biti beseda neznana (`$freq == 0`), obenem pa mora biti besedna oblika (lastnost `idwrd`) pojavnice, katere oznako popravljamo (`$sent[$focus]`), sestavljena samo iz znakov `I`, `V` in `X` (regularni izraz `/^[IVX]+$/`). Drugo pravilo popravi oznako besede tako, da jo uvrsti v kategorijo `Y`, tj. označi jo kot okrajšavo, in sicer če je beseda neznana, sestavljena iz samih velikih črk, pred in za njo pa nista kapitalizirani besedi. Na ta način npr. pravilno označimo zglede tipa *Čist dobiček ECB se prenese ...*, ne pride pa do napačne

označitve neznane besede v zgledih tipa *profesor dr. Walter HALLSTEIN , državni sekretar ...*

Zaenkrat smo implementirali pet pravil, ki izhajajo iz analize nekaj najbolj pogostih, pa obenem najbolj »popravljivih« napak. Prvi dve pravili sta že bili opisani, tretje spremeni oznako v okrajšavo, če je neznana beseda sestavljena iz številk in največ treh črk (npr. *2002/917/ES*), četrto spremeni oznake vseh namenilnikov v oznake za samostalnike moškega spola v osnovni obliki, peto pa spremeni oznako *a*, če mu sledi ločilo, tako, da mu namesto vezniške vloge podeli vlogo okrajšave.

Tabela 5 podaja rezultat popravkov pri označevanju s temi petimi pravili. V prvem stolpcu so števila za popravek besedne vrste, v drugem pa za popravek celotne oblikoslovne oznake. Prva vrstica prikazuje število pravilno popravljenih pojavnic, druga vrstica pa število pojavnic, ki so bile po prvotnem označevanju pravilne, vendar so bile popravljene v napačno besedno vrsto oz. oblikoslovno oznako. V tretji vrstici je število oznak, ki so bile napačne po prvotnem označevanju in popravljene v ravno tako napačno oznako, zadnja vrstica pa opozarja na število oznak, ki jih je neko pravilo sicer popravilo, vendar so popravljene oznake identične izvornim. Vrednosti, prikazane v zadnjih dveh vrsticah, na natančnost označevanja ne vplivajo, je pa vseeno zaželeno, da so vrednosti v predzadnji vrstici čim manjše, saj so nove napake bolj kompleksne vrste kot pa izvorne napake. Absolutno natančnost dobimo tako, da odštejemo drugo vrstico od prve. Rezultat je podan v zadnji vrstici.

	Besedna vrsta	Oblikoslovna oznaka
Popravljeno	291	289
Pokvarjeno	4	4
Zamešano	14	16
Enako	2	2
Izboljšanje	287	285

Tabela 5: Rezultat pri avtomatskem popravljanju napak.

Napake v oblikoslovnih oznakah se z uporabo programa za popravke torej zmanjšajo za 287 pojavnic. S tem napako glede na besedne pojavnice zmanjšamo za 16% oz. dvignemo natančnost s 86,6% na 88,9%. Razlika ni ogromna, vendar pa maksimiziranje te mere natančnosti niti ni bil naš cilj, saj vsa pravila popravljajo oznako besedne vrste. Izboljšanje natančnosti za besedne vrste pa je dosti bolj opazno: z uporabo petih pretvorbenih pravil se ta izboljša za 38,4% oz. se z 94,4% poveča na 96,6% absolutne natančnosti.

5. Primerjave natančnosti označevanja

V Tabeli 6 podajamo pregled in primerjavo natančnosti označevanja po oblikoslovnih oznakah in besednih vrstah. Prva vrstica predstavlja rezultate iz raziskave Erjavec et al. (2000), v kateri je bil tako za učno kot za testno množico uporabljen korpus MULTEXT-East, tj. roman »1984«. Druga vrstica prikazuje glavno evalvacijo označevanja s programom TnT/totale na vzorcu iz korpusa SVEZ-IJS. Natančnost je sicer nižja, kar glede na precejšnje razlike med obema korpusoma ni presenetljivo, vendar ne v taki meri, kot bi mogoče pričakovali. Naslednja vrstica podaja rezultate, dobljene po izvajanju pretvorbenega programa, ki vsebuje 5 ročno napisanih pravil, dobljenih s pomočjo analize najbolj pogostih napak. Treba je opozoriti, da so tri pravila od petih (in to tista, ki pokrijejo največ primerov) popravljala oznake okrajšav oz. rimskih številk.

Zdi se, da so te napake in napake pri nerazporejenih besedah v večji meri problem tokenizacije kot pa samega oblikoslovnega označevanja. Robustna rešitev problema označevanja okrajšav in tujejezičnih citatov bi bila zato prej v izdelavi dodatnega modula, ki označevalniku na osnovi tipografskih značilnosti pojavnic dopolni leksikon za konkreten dokument. Zato je zanimivo pogledati še, kakšna bi bila natančnost označevanja, če bi pojavnice kategorij X in Y iz evalvacije izpustili. Natančnost se v tem primeru znatno poveča in doseže že 894% za oblikoslovne oznake oziroma 97,6% za besedne vrste.

V tabeli prikazujemo za primerjavo z rezultati našega sistema označevanja tudi natančnost najboljšega označevanja, predstavljenega v raziskavi Lönneker (2005), ki je glede na oblikoslovno oznako 83,6%, glede na besedno vrsto pa ni podana. Poskusa se v marsičem razlikujeta (v naboru oblikoslovnih oznak, velikosti učnega korpusa in sestavi testnega korpusa), zato je natančnosti težko neposredno primerjati, vseeno pa razlika v rezultatih preseneča, posebej glede na to, da je učni korpus pri poskusu ZRC/TreeTagger vseboval preko milijon besed in je bil torej desetkrat večji kot naš. Lönneker (2005) postavi nekaj hipotez, zakaj je natančnost pri njenih poskusih manjša, mdr. izpostavi bolj podrobne oznake, ki npr. ločijo različne vrste imen (osebna, krajevna, mitološka), in manjšo konsistentnost pri samem označevanju učnega in testnega korpusa. Zdi se nam verjetno, da je dodaten razlog tudi to, da je označevalnik TnT boljši kot TreeTagger, predvsem pri označevanju neznanih besed. Avtorica žal ne podaja posebej natančnosti za znane in neznane besede, tako da te hipoteze ne moremo preveriti.

	Oblikoslovna oznaka	Besedna vrsta
1984: TnT	89,2%	96,6%
SVEZ-IJS: TnT	86,6%	94,4%
SVEZ-IJS: TnT + Trans	**88,9%**	**96,6%**
SVEZ-IJS – X,Y: TnT	89,4%	97,6%
ZRC ISJ: TreeTagger	83,6%	?

Tabela 6: Ocene napak.

6. Zaključki

V članku smo analizirali natančnost avtomatskega oblikoslovnega označevanja z označevalnikom TnT, ki je vgrajen v program totale in izšolan na oblikoslovnih virih MULTEXT-East za slovenski jezik. Evalvacija je potekala na ročno popravljenem vzorcu iz slovenskega dela korpusa SVEZ-IJS, velikem približno 15.000 pojavnic, ki vsebuje okoli 15% sistemu neznanih besed. Evalvacija je pokazala, da je absolutna natančnost glede na besedne pojavnice v vzorcu 86,6% za polno označevanje oz. 94,4%, če opazujemo samo napake v besedni vrsti. V

primeru, da označevanje izboljšamo s pretvorbenim programom, ki odpravi nekaj najbolj pogostih in obenem enostavnih napak, se natančnost poveča na 88,9% za oznake oz. na 96,6% za besedne vrste.

Dodaten način, kako izboljšati natančnost, smo že omenili, in sicer s predprocesiranjem, ki bi identificiralo okrajšave in nerazporejene besede (tujejezične citate). Očiten korak k doseganju večje natančnosti bi bil tudi povečanje učnega korpusa in raznovrstnosti besedil v njem, pri čemer je glavna težava zamudnost (cena) takšnega označevanja. Žal pa bo to delo treba podvajati, saj veliki ročno označeni korpus ZRC (za razliko od npr. virov MULTEXT-East in korpusa SVEZ-IJS) izven matične institucije ni na voljo.

V literaturi najdemo tudi številne druge napotke, kako je možno označevanje izboljšati. Zanimiv pristop, in javno dostopen program, je opisan v raziskavi Brill (1992) in je bil tudi izhodišče za naš pretvorbeni program. Bistvena razlika pa je, da smo mi pravila napisali ročno, program, opisan v omenjeni raziskavi, pa se pretvorbenih pravil nauči na osnovi učne množice, pri čemer lahko sami definiramo lastnosti, glede na katere naj se program uči. Drugačen pristop, ki se je tudi že uporabljal za jezike z bogatim naborom oznak in majhno učno množico, je opisan v Tufiş (2006). Pristop predvideva redukcijo števila oblikoslovnih oznak, vendar tako, da je polne oznake še vedno možno rekonstruirati iz leksikona.

Zahvala

Avtorja se zahvaljujeta anonimnima recenzentoma za koristne pripombe. Raziskavo, opisano v pričujočem prispevku je podprl raziskovalni program ARRS »Tehnologije znanja« (prvi avtor), in štipendija CMEPIUS RS (drugi avtor).

Literatura

Brill, E. (1992). A Simple Rule-Based Part of Speech Tagger. V: Proceedings of the Third Conference on Applied Natural Language Processing, ACL. Trento, Italija.

Brants, T. (2000). TnT - A Statistical Part-of-Speech Tagger. V: Proceedings of the Sixth Applied Natural Language Processing Conference, ANLP-2000. Seattle, WA. 224--231.

Erjavec, T., Džeroski, S., Zavrel, J. (2000). Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. V: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000). ELRA, Pariz.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. V: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'2004). ELRA, Pariz.

Erjavec, T., Džeroski, S. (2004). Machine Learning of Language Structure: Lemmatising Unknown Slovene Words. Applied Artificial Intelligence, 18, 17--41.

Erjavec, T., Ignat, C., Pouliquen, B., Steinberger, R. (2005). Massive multilingual corpus compilation: ACQUIS Communautaire and totale. V: Proceedings of the Second Language Technology Conference. April 2004, Poznan.

Erjavec, T. (2006). The English-Slovene ACQUIS corpus. V: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'2006). ELRA, Pariz.

Erbič, D., Krstič Sedej, A., Belc J., Zaviršek-Žorž, N., Gajšek, N., Željko, M. (2005). Slovenščina na spletu v dokumentih slovenske različice pravnega reda Evropske unije, terminološki zbirki in korpusu. V: Zbornik Simpozija Obdobja 24: Razvoj slovenskega strokovnega jezika. Ljubljana.

Hajič, J., Hladka, B. (1998). Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. COLING-ACL'98. ACL.

van Halteren, H. (ur.) (1999). Syntactic Wordclass Tagging. Kluwer.

Jakopin, P., Bizjak, A. (1997). O strojno podprtem oblikoslovnem označevanju slovenskega besedila. Slavistična Revija. 45/3-4. 513--532.

Lönneker, B. (2005). Strojno oblikoslovno označevanje slovenskih besedil: Kako daleč smo. Slavistična revija 53/2. 193--210.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. V: Proceedings of International Conference on New Methods in Language Processing. Manchester. 44--49.

Tufiş, D. (2006). Tagset Design for High Accuracy POS Tagging and Automatically Building Mapping between Arbitrary Tagsets. Workshop on Annotation Science: State of the Art in Enhancing Automatic Linguistic Annotation (LREC'06). ELRA, Paris.