VEČ GLAV VEČ VE: UPORABA MNOŽIČENJA ZA ČIŠČENJE SLOWNETA

Darja Fišer
Filozofska fakulteta, Ljubljana

Aleš Tavčar
Institut »Jožef Stefan«, Ljubljana

UDK 811.163.6'374'371'322

V prispevku predstavljamo projekt čiščenja avtomatsko generiranega semantičnega leksikona sloWNet. Napake, ki se v leksikonu pojavljajo zaradi napačne avtomatske disambiguacije večpomenskih besed, smo odpravili s pomočjo orodja sloWCrowd, ki je zasnovano tako, da odgovore za problematične literale zbira iz široke množice uporabnikov – prostovoljcev. Naloga je oblikovana kot spletna igra, v kateri uporabniki tekmujejo, kdo bo zbral več točk (prispeval več pravilnih odgovorov). Glede na to, da tekmovalci niso izurjeni leksikografi, njihovi odgovori niso nujno zanesljivi, zato orodje omogoča merjenje njihove natančnosti in pri vsakem vprašanju upošteva večinski odgovor, s čimer zagotavlja, da posamezni napačni odgovori sicer zanesljivih uporabnikov ter vsi odgovori nezanesljivih uporabnikov ne vplivajo na dokončno odločitev, ali se določen literal iz leksikona izbriše ali ne.

množičenje, leksikalna semantika, večpomenskost, sloWNet

The paper presents the cleaning of the automatically generated semantic lexicon sloWNet. Errors that occurred due to inappropriate disambiguation of polysemous words were eliminated with a tool called sloWCrowd, which is designed in such a way that it collects multiple answers for problematic literals from a wide number of volunteer users. The task is designed as a web game in which users compete who will collect the highest number of points (contribute the most correct answers). Since the users are not trained lexicographers, the reliability of their answers is questionable, which is why the tool has been designed to measure the users' accuracy and relies on the majority vote for each literal. This means that the individual incorrect answers from otherwise reliable users and all the answers from unreliable users do not affect the final decision whether or not the literal is to be deleted from the lexicon.

crowdsourcing, lexical semantics, polysemy, sloWNet

1 Uvod

V zadnjem desetletju postajajo avtomatski pristopi za izdelavo jezikovnih virov vse popularnejši, saj omogočajo hitrejšo, cenejšo in preprostejšo izdelavo obsežnih virov. Razumljivo pa je, da z njimi ne moremo dosegati človeške kvalitete, zato je treba tako izdelane vire pregledati in v njih odpraviti morebitne napake. Ker je tudi tovrstno delo zamudno in drago, so številni raziskovalci začeli proučevati možnosti, ki bi postopek pospešile in pocenile, pri čemer se kvaliteta zbranih odgovorov ne bi bistveno znižala. Kot so pokazali njihovi poskusi, je nalogo mogoče razdeliti na obvladljive in razumljive kose ter jo ponuditi v reševanje široki množici uporabnikov svetovnega spleta, ki niso nujno strokovnjaki z obravnavanega področja. Kvaliteto je mogoče zagotoviti s preverjanjem zanesljivosti uporabnikov skozi ponavljanje istega vprašanja različnim uporabnikom in filtriranjem njihovih odgovorov (Adda idr. 2011).

Ena izmed najbolj razširjenih platform za množičenje (ang. crowdsourcing) je Mechanical Turk[1] ameriškega podjetja Amazon. Za motivacijo povprečnih uporabnikov svetovnega spleta, da se pridružijo eksperimentu in prispevajo čim več odgovorov, so raziskovalci razvili t. i. igre z razlogom (ang. games with a purpose), ki od uporabnika na zabaven, a strukturiran način pridobivajo želene podatke. Na področju jezikoslovja je najbolj znana igra Word Detectives,[2] s pomočjo katere označujejo anafore v besedilih (Chamberlain idr. 2008), za označevanje besednega pomena pa je bila razvita spletna igra Wordrobe[3] (Venhuizen 2013).

V prispevku predstavljamo odpravljanje napak iz avtomatsko zgrajenega semantičnega leksikona za slovenščino sloWNet (Fišer 2009), ki je bil zasnovan na sorodnem angleškem viru Princeton WordNet (Fellbaum 1998) in je vseboval precej šuma, ki bi ga bilo treba čim prej odpraviti. Najbolj problematične lekseme (literale), ki skoraj zagotovo ne ubesedujejo pojma (sinseta), ki so mu pripisane, smo s pomočjo kontekstualnih informacij iz referenčnega korpusa FidaPLUS[4] in načel distribucijske semantike identificirali avtomatsko (Sagot, Fišer 2012). Te potencialne napake smo s pomočjo orodja sloWCrowd ponudili v glasovanje večjemu številu prostovoljcev in jih glede na zbrane odgovore označili kot validirane oz. jih izbrisali iz leksikona.

V nadaljevanju prispevka predstavimo orodje za množičenje sloWCrowd, v osrednjem razdelku prikažemo rezultate projekta, prispevek pa sklenemo z diskusijo in načrti za prihodnje delo.

2 Orodje sloWCrowd

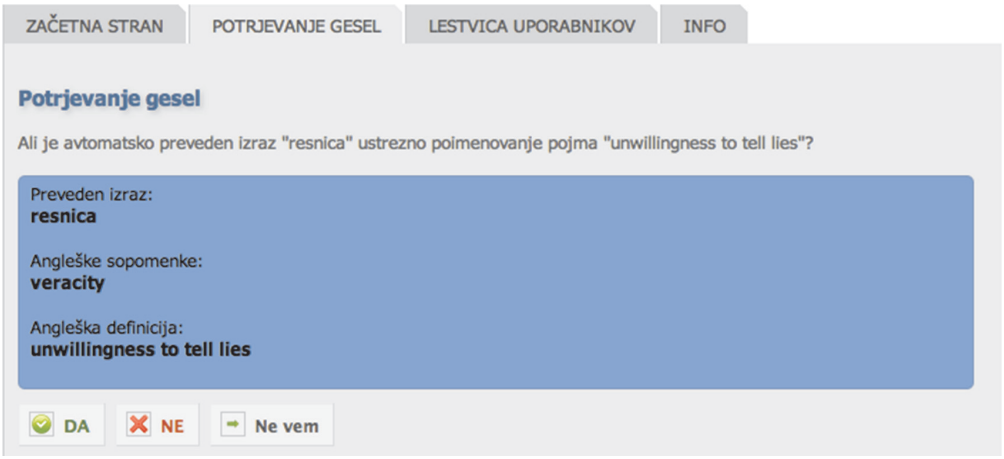
Orodje sloWCrowd[5] smo razvili v sodelovanju z Institutom »Jožef Stefan« (Tavčar idr. 2012). Namenjeno je množičenju različnih tipov nalog na področju gradnje jezikovnih virov in orodij, pri katerih je treba zbrati večje število človeških odgovorov. S prenosom bremena verifikacije na širšo množico se zmanjša čas verifikacije, dobljeni rezultati pa imajo visoko stopnjo zanesljivosti, saj o istem vprašanju glasuje večje število uporabnikov.

Orodje sloWCrowd je sestavljeno iz dveh delov, administratorskega in uporabniškega. V administratorskem delu ustvarimo projekt in vodimo zbiranje odgovorov. Uporabniški del pa omogoča izbiro projekta, pri katerem uporabnik želi sodelovati, predstavitev projekta z navodili za reševanje, reševanje nalog in lestvico najboljših ocenjevalcev glede na število in pravilnost rešenih nalog. Orodje je prosto dostopno[6] ter temelji na popularnem odprtokodnem programskem jeziku za strežniško rabo in razvoj dinamičnih spletnih vsebin PHP in podatkovni bazi MySQL, kar omogoča prenosljivost in enostavno namestitev. Deluje na večini spletnih brskalnikov, saj uporablja splošno razširjene tehnologije.

Osnovna funkcionalnost orodja sloWCrowd je potrjevanje in zavračanje literalov. Po prijavi se uporabniku prikaže glavno okno, kjer rešuje naloge. Primer naloge prikazuje slika 1. Na vrhu zaslona so navodila za reševanje naloge, s katerimi uporabnika prosimo, da presodi, ali je avtomatsko preveden slovenski izraz primeren za izbrani koncept. Nato so za lažje odločanje navedene še dodatne informacije, npr. angleške sopomenke in angleška definicija istega koncepta. Uporabnik lahko s klikom na enega od gumbov na dnu zaslona izbira med tremi možnostmi: DA, NE in NE VEM, s katerim vprašanje preskoči.

Odgovori se nato zapisujejo v bazo MySQL, zbrani pa so prikazani tudi v uporabniškem vmesniku, kot prikazuje slika 2.

1 https://www.mturk.com/mturk/welcome
2 http://anawiki.essex.ac.uk/phrasedetectives
3 http://wordrobe.housing.rug.nl
4 http://fidaplus.net
5 http://nl.ijs.si/slowcrowd
6 http://nl.ijs.si/slowcrowd/sloWCrowd.rar



Slika 1: Primer vprašanja za čiščenje semantičnega leksikona sloWNet. Reševalci se morajo odločiti, ali je preveden izraz (npr. resnica) ustrezen prevod za angleško besedo (npr. veracity) glede na dano angleško definicijo (npr. unwillingness to tell lies). V danem primeru prevod ni točen, zato od reševalcev pričakujemo, da bodo kliknili na gumb NE

Literalu s pripadajočimi angleškimi podatki sledi podatek, ali je literal iz referenčne datoteke s pravilnimi odgovori, ki se uporablja za računanje stopnje natančnosti uporabnika (GS), nato število vseh zbranih odgovorov (Vsi), ter še število vseh pozitivnih (+) in negativnih (–) odgovorov.

3 Čiščenje sloWNeta

3.1 Opis projekta

Ker je bil sloWNet zgrajen avtomatsko, je razumljivo, da vsebuje določeno stopnjo šuma. Glede na to, da za ročno popravljanje celotnega leksikona ni na voljo finančnih sredstev, smo se odločili avtomatsko identificirati najresnejše potencialne napake, ki so posledica napačne disambiguacije večpomenskih besed v fazi prevajanja iz angleščine v

Literal	Synonyms	Definition	GS	All	+	-	?
žreb	caboodle, lot, bunch	any collection in its entirety	-	5	0	5	0
žreb	lot, band, circle, set	an unofficial association of people or groups	-	5	0	5	0
žreb	circumstances, lot, luck, fortune, destiny, portion, fate	your overall circumstances or condition in life (including everything that happens to you)	✓	16	0	15	1
žrtev	sacrifice	a loss entailed by giving up or selling something at less than its value	-	2	1	1	0
župnik	curate, pastor, parson, minister of religion, minister, rector	a person authorized to conduct religious worship	-	5	5	0	0
žur	party	a group of people gathered together for pleasure	-	5	5	0	0

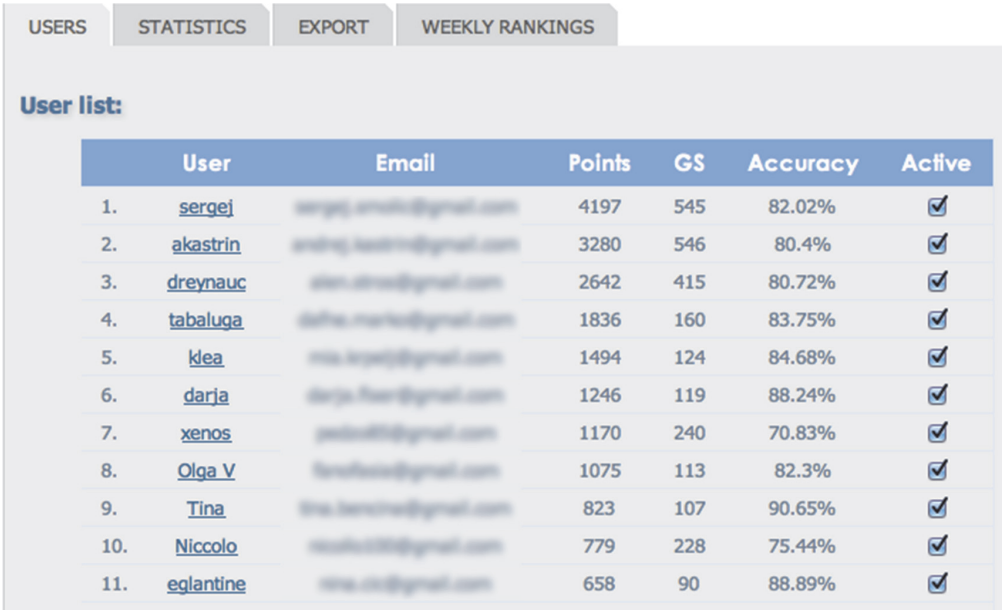
Slika 2: Primer zbranih odgovorov za literale *žreb*, *žrtev*, *župnik* in *žur*, pri katerih bi bilo glede na večinsko mnenje reševalcev vse, razen literala *žrtev*, za katerega še nismo zbrali dovolj odgovorov, treba izbrisati iz leksikona

slovenščino. To smo storili z metodami distribucijske semantike na podlagi primerjave neposredne okolice literalov v semantični mreži in kontekstualnih informacij zanje v referenčnem korpusu Gigafida. Kadar je ujemanje med okolico literala v semantični mreži in njegovim kontekstnim profilom v korpusu izjemno nizko, lahko sklepamo, da skoraj zagotovo ne ubeseduje pojma tistega sinseta, ki mu je pripisan, in je torej najverjetneje napačen (Sagot, Fišer 2012). A ta metoda ni dovolj natančna, da bi jo lahko uporabili za neposredno brisanje iz leksikona, prav tako na podlagi njenih rezultatov ne moremo trditi, da so visoko rangirani literali zagotovo pravilni, temveč nam služi le kot predizbor literalov, ki jih bomo pregledali najprej, ker so najbolj dvomljivi.

Z rangiranega seznama smo za ročni pregled izbrali 8000 literalov z najnižjo stopnjo ujemanja. Če ponazorimo s primerom samostalnika *beseda*, ki je bil napačno pripisan kot prevod angleškemu sinsetu *term, condition: a statement of what is required as part of an agreement* (*pogoj: izjava, kaj je glede na dogovor nujno potrebno upoštevati*), bi bil ta prevod sicer ustrezen za enega od pomenov angleške besede *term* (*izraz*, *termin*), vendar ne za tega, saj *term* v tem pomenu prevajamo kot *pogoj*, zato je treba literal *beseda* iz tega sinseta v sloWNetu izbrisati. To so po eni strani najhujše napake v sloWNetu, ki najbolj negativno vplivajo na uporabno vrednost semantičnega leksikona, po drugi pa predvidevamo, da jih bo prostovoljcem, ki niso izurjeni leksikografi, hkrati tudi najlažje popraviti, saj so tovrstne napake najočitnejše.

V orodju sloWCrowd smo ustvarili projekt, v katerega smo naložili teh 8000 literalov in jih uporabnikom ponudili v reševanje, pri čemer smo uporabnike prosili, da se odločijo, ali je izbrani slovenski literal ustrezno poimenovanje za angleški sinset, ki mu je bil pripisan (glede na prikazane angleške sinonime ter definicijo). Za udobnejše reševanje smo vprašanja oblikovali v sklope, pri čemer se v posameznem sklopu uporabniku prikaže 10 naključno izbranih vprašanj, na katera odgovori, nato pa se odloči, ali želi začeti še en sklop. Orodje uporabniku ponuja določen delež novih in že rešenih nalog iz referenčne datoteke, s katerimi se ugotavlja njegova zanesljivost. Glede na pravilnost uporabnikovih odgovorov se mu prikaže večji ali manjši delež nalog iz referenčne datoteke. Lestvica je progresivna, saj se z višanjem deleža pravilnih odgovorov iz referenčne množice viša delež novih, še neoznačenih vprašanj. Poleg ugotavljanja zanesljivosti uporabnikov glede na referenčne naloge orodje beleži tudi stopnjo ujemanja z drugimi uporabniki. Za motivacijo uporabnikov pri reševanju nalog se njihovi odgovori točkujejo, najboljših pet uporabnikov pa je nato prikazanih na lestvici najboljših ocenjevalcev. Uporabnik dobi točke za vsak pravilni odgovor glede na referenčno datoteko in odgovore ostalih uporabnikov v bazi.

Ker je za reševanje nalog potrebno znanje slovenskega in angleškega jezika, smo k sodelovanju pri projektu povabili predavatelje in študente slovenistike, anglistike in prevajalstva z ljubljanske, mariborske, novogoriške in koprske univerze. Predavatelje smo prosili, da pri pouku na kratko predstavijo projekt, študentje, ki jih je sodelovanje pri projektu zanimalo, pa so na vprašanja odgovarjali doma. Prosili smo jih, da odgovorijo na vsaj 10 sklopov vprašanj (tj. posredujejo odgovore za 100 literalov), po želji pa tudi več. Za reševanje smo jim dali eno študijsko leto časa, da so lahko nalogo izvedli, ko so jim to dopuščale druge obveznosti. Pred reševanjem smo jim posredovali naslednja navodila: »Naloge rešuješ tako, da prebereš slovensko besedo, angleško definicijo in angleške sopomenke ter se odločiš, ali je slovenska beseda ustrezen prevod za to angleško definicijo in sopomenke. Če se s tem strinjaš, klikneš na gumb DA, če se ne strinjaš, klikneš NE, če pa besede ne razumeš ali nisi prepričan, ali je pravilna ali ne, pa klikneš NE VEM.«



	User	Email	Points	GS	Accuracy	Active
1.	sergej	[redacted]	4197	545	82.02%	☑
2.	akastrin	[redacted]	3280	546	80.4%	☑
3.	dreynauc	[redacted]	2642	415	80.72%	☑
4.	tabaluga	[redacted]	1836	160	83.75%	☑
5.	klea	[redacted]	1494	124	84.68%	☑
6.	darja	[redacted]	1246	119	88.24%	☑
7.	xenos	[redacted]	1170	240	70.83%	☑
8.	Olga V	[redacted]	1075	113	82.3%	☑
9.	Tina	[redacted]	823	107	90.65%	☑
10.	Niccolo	[redacted]	779	228	75.44%	☑
11.	eglantine	[redacted]	658	90	88.89%	☑

Slika 3: Seznam uporabnikov, ki so odgovorili na več kot 500 vprašanj. Njihove elektronske naslove smo zaradi varovanja osebnih podatkov zakrili

3.2 Predstavitev in analiza rezultatov

3.2.1 Število uporabnikov in zbranih odgovorov

V orodje sloWCrowd se je v obdobju med 25. septembrom 2012 in 2. septembrom 2013 prijavilo 310 uporabnikov, ki so skupaj odgovorili na 41.587 vprašanj za 7544 različnih literalov. Če ne upoštevamo vprašanj iz referenčne datoteke, ki smo jih uporabnikom zastavili zgolj zato, da smo merili njihovo natančnost, smo doslej zbrali 31.637 novih odgovorov za 7246 od 80.000 želenih literalov, tako da imamo v povprečju zbranih 4,36 od 5 želenih glasov za vsak nov ocenjevan literal. Uporabniki so za posamezno vprašanje porabili v povprečju 10 sekund, kar pomeni, da je bilo v projekt vloženih 115 ur, kar ustreza 14 dnem z 8-urno dnevno delovno obremenitvijo.

Kar nekaj uporabnikov (12 %) je že takoj ugotovilo, da jih naloga ne zanima oz. ji niso kos, saj niso odgovorili niti na eno vprašanje. Vsaj en odgovor je prispevalo 273 reševalcev, na več kot pet vprašanj jih je odgovorilo 254, na več kot 10 pa 210. Čeprav so reševalci, ki so prispevali vsaj en odgovor, v povprečju odgovorili na dobrih 152 vprašanj, je distribucija števila prispevanih odgovorov zelo neenakomerna, saj je 100 uporabnikov rešilo le deset nalog ali manj, medtem ko jih je uporabnik z največjim številom prispevanih odgovorov rešil kar 4197. Kot prikazuje slika 3, je število uporabnikov, ki so prispevali več kot 500 odgovorov, 11. Ti so skupaj prispevali skoraj 58 % vseh zbranih odgovorov za nove literale v projektu.

3.2.2 Zanesljivost uporabnikov

Povprečna zanesljivost uporabnikov je bila izračunana na podlagi odgovorov za 300 literalov iz referenčne datoteke, ki že vsebuje pravilne odgovore; znaša 80,12 %, kar je visoko za naloge v leksikalni semantiki. Pri tem je treba poudariti, da je nihanje med stopnjo zanesljivosti posameznih uporabnikov zelo veliko, saj so bili pri nekaterih uporabnikih vsi prispevani odgovori pravilni, obstajajo pa tudi uporabniki, ki so na vsa zastavljena vprašanja odgovorili narobe. 72,16 % vseh upo-

rabnikov je doseglo vsaj 75-odstotno natančnost, ti pa so prispevali 85 % vseh zbranih odgovorov. Povprečna stopnja natančnosti desetih uporabnikov, ki je prispevala največ odgovorov, pa znaša celo 83,71 %. To pomeni, da so odgovori uporabnikov, ki so v projektu odgovorili na največ vprašanj, hkrati tudi med najbolj zanesljivimi. Vse uporabnike, ki 75-odstotnega praga natančnosti ne dosegajo, je v orodju sloWCrowd mogoče dezaktivirati, kar pomeni, da njihovi odgovori pri izvozu rezultatov ne bodo upoštevani, s čimer bomo izločili najbolj netočne odgovore, pri tem pa bomo izgubili le 15 % zbranih podatkov.

3.2.3 Analiza prispevanih odgovorov

Reševalci so na večino zastavljenih vprašanj (15.861 oz. dobrih 50 %) odgovorili negativno, kar pomeni, da se jim je zdela dobra polovica literalov, ki smo jih z avtomatskimi metodami identificirali kot problematične, dejansko napačnih in jih je treba izbrisati iz sloWNeta. Na 14.984 oz. 47,36 % zastavljenih vprašanj so odgovorili pritrdilno, na zgolj 792 oz. 2,5 % vprašanj pa niso znali oz. želeli odgovoriti in so se jih odločili preskočiti. To pomeni, da je bila naloga razmeroma lahka in napake precej očitne, kot smo tudi predvidevali.

Ker smo zaradi zagotavljanja večje zanesljivosti dobljenih odgovorov isto vprašanje zastavili več različnim reševalcem (pri čemer smo maksimalno število ponovitev istega vprašanja nastavili na pet), smo dokončno odločitev o validaciji oz. izbrisu literala iz semantičnega leksikona sprejeli tako, da smo upoštevali večinsko mnenje vseh, ki so odgovorili na vprašanje. Reševalci so se odločali o ustreznosti pripisanih pomenov za 1476 samostalnikov, ki so pri avtomatski izdelavi sloWNeta razvrščeni v 2901 različnih sinsetov. V največ različnih sinsetov so bili razvrščeni samostalniki *oseba* (35), *sprememba* (35) in *igra* (20), kar že na prvi pogled kaže na to, da so številni med njimi skoraj zagotovo napačni. Z upoštevanjem večinskih odgovorov smo tako iz sloWNeta izbrisali potrjeno napačne literale iz 1264 (44 %) različnih sinsetov, 1446 (50 %) pa smo jih s pomočjo glasovanja uporabnikov označili kot pravilne. 190 literalov (6 %) je prejelo enako število pozitivnih kot negativnih odgovorov, zato še vedno ostajajo vprašljivi in jih bomo morali ponuditi v reševanje še večkrat, preden bomo lahko sprejeli dokončno rešitev zanje.

V manjši meri so med nerazrešenimi literali primeri, za katere je angleška razlaga precej ohlapna in nejasna ali pa so prevodi angleških ustaljenih fraz, ki se v slovenščini uporabljajo v drugem kontekstu ali z drugimi besedami, zaradi česar se uporabniki do njih niso znali enoznačno opredeliti. A v veliki večini primerov literali ostajajo nerazrešeni, ker jih zaradi naključnega izbiranja vprašanj orodje sloWCrowd še ni petkrat ponudilo v reševanje, ne pa zato, ker so bistveno bolj problematični po vsebinski plati, tako da pričakujemo, da bomo za večino od njih v nadaljevanju projekta vendarle dobili enoznačen odgovor. Če bi tudi po petkrat zastavljenem vprašanju zaradi uporabe gumba »NE VEM« kak literal še vedno ostal dvoumen, lahko te predlaga urednik sloWNeta ali pa število iteracij, kolikokrat orodje sloWCrowd zastavi neko vprašanje, prilagodimo tako, da dodamo pogoj, da ga po petih zbranih odgovorih ponavlja še toliko časa, dokler eden od odgovorov ne prevlada.

Temeljitega pregleda večinskih odgovorov še nismo opravili, zato ne moremo z gotovostjo trditi, da je večinsko mnenje uporabnikov vedno tudi jezikovno ustrezna rešitev. A glede na visoko stopnjo ujemanja uporabnikov z referenčno datoteko, ki znaša v povprečju nekaj čez 80 %, nakazuje na to, da so njihovi odgovori vendarle zelo zanesljivi. Prav tako je pregled naključnih 100 večinskih odločitev, dobljenih v tem projektu, pokazal, da je napačnih odgovorov res zelo malo, saj smo med pregledanimi našli le dva napačna literala, ki so ju uporabniki izglasovali kot pravilna:

– *del* (ang. member) – *anything that belongs to a set or class* in

– *člen* (ang. division, part, section) – *one of the portions into which something is regarded as divided and which together constitutes a whole*.

Poleg tega, da je teh primerov res zelo malo, je treba poudariti, da so še ti napake, ki v sloWNetu ostajajo ter jih bomo morda identificirali in odpravili že pri naslednjem pregledovanju. Veliko večjo škodo bi uporabniki povzročili, če bi iz sloWNeta izbrisali sicer pravilne literale, saj je veliko naporneje te kasneje dodati na novo kot brisati napake. A teh primerov v pregledanem vzorcu nismo zaznali.

4 Zaključki

V prispevku smo predstavili orodje sloWCrowd, ki je namenjeno ročni validaciji avtomatsko pridobljenih jezikovnih podatkov. Administrativni del vmesnika omogoča preprosto izdelavo projekta in uvoz podatkov, uporabniški vmesnik pa je zasnovan tako, da lahko uporabnik naloge rešuje čim hitreje in enostavneje.

V projektu, v katerem smo s pomočjo orodja sloWCrowd odpravljali napake v slovenskem semantičnem leksikonu sloWNet, je nekaj čez 300 uporabnikov prostovoljcev pregledalo skoraj 8000 najbolj vprašljivih avtomatsko generiranih literalov iz sloWNeta in jih označilo kot pravilne ali napačne. Njihove odgovore smo zbrali in analizirali, rezultati pa kažejo naslednje:

– da je orodje enostavno za uporabo, saj so se ga uporabniki hitro naučili uporabljati in z njim niso imeli nobenih tehničnih težav,
– da je naloga zastavljena tako, da lahko tudi neizurjeni leksikografi v zelo kratkem času veliko pripomorejo k odpravljanju napak v semantičnem leksikonu, saj za odgovor na eno vprašanje v povprečju potrebujejo le 10 sekund, in
– da so zbrani odgovori zanesljivi, saj je stopnja ujemanja med uporabniki visoka (80 %), število dvoumnih rešitev pa zelo majhno.

V prihodnje nameravamo nadaljevati validacijo preostalih literalov v sloWNetu, in sicer po ključu stopnje večpomenskosti, saj menimo, da so besede, ki se trenutno nahajajo v zelo velikem številu sinsetov, najverjetneje razvrščene tudi v napačne. Projekt nameravamo v prihodnje tudi razširiti na preverjanje drugih besednih vrst, saj so bili za zdaj pregledani le samostalniki. Orodje pa želimo preizkusiti tudi za validacijo wordnetov v drugih jezikih, pri čemer že sodelujemo z univerzo Diderot v Parizu, kjer bi pregledali francoski wordnet, ter s kolegi z zagrebške filozofske fakultete, ki bodo popravljali hrvaški wordnet.

Projekti, ki trenutno tečejo v orodju sloWCrowd, so dostopni na: http://nl.ijs.si/slowcrowd/select_project.php. Ker je orodje prosto dostopno pod licenco Creative Commons, ga je mogoče tudi prenesti, namestiti na lastni strežnik in prilagoditi svojim potrebam. Za namestitev sta potrebna le PHP in MySQL. Namestitvene datoteke so na: http://nl.ijs.si/slowcrowd/sloWCrowd.rar.

Zahvala

Avtorja se zahvaljujeta vsem predavateljem in študentom, ki so sodelovali pri projektu čiščenja sloWNeta.

Literatura

ADDA, Gilles, SAGOT, Benoît, FORT, Karën, MARIANI Joseph, 2011: Crowdsourcing for Language Resource Development: Critical Analysis of Amazon Mechanical Turk Overpowering Uses. *Zbornik konference LTC 2011*. Poznań.

von AHN, Luis, 2006: Games with a Purpose. *Computer* 39/6. 92–94.

CHAMBERLAIN, Jon, POESIO, Massimo, KRUSCHWITZ, Udo, 2008: Phrase Detectives: a Web-based collaborative annotation game. *Zbornik konference iSemantics*. Gradec.

FIŠER, Darja, 2009: Pristopi za avtomatizirano gradnjo semantičnih zbirk. *Terminologija in sodobna terminografija*. Ljubljana: Založba ZRC, ZRC SAZU. 357–370.

FIŠER, Darja, NOVAK, Jernej, 2011: Visualizing sloWNet. *Zbornik konference eLEX2011*. Bled.

SAGOT, Benoît, FIŠER, Darja, 2012: Cleaning noisy wordnets. *Zbornik konference LREC 2012*. Istanbul.

SNOW, Rion, O'CONNOR, Brendan, JURAFSKY, Daniel, NG, Andrew Y., 2008: Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. *Zbornik konference EMNLP 2008*. 254–263.

VENHUIZEN, Noortje J., BASILE, Valerio, EVANG, Kilian, BOS, Johan, 2013: Gamification for word sense labeling. *Zbornik konference IWCS*.