sloWCrowd: orodje za popravljanje wordneta z izkoriščanjem moči množic

Aleš Tavčar,* Darja Fišer,† Tomaž Erjavec‡

* Odsek za inteligentne sisteme, Institut Jožef Stefan
Jamova cesta 39, 1000 Ljubljana
ales.tavcar@ijs.si
† Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
darja.fiser@ff.uni-lj.si
‡ Odsek za tehnologije znanja, Institut Jožef Stefan
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

Povzetek

V prispevku predstavimo orodje sloWCrowd, ki smo ga razvili za lažje odpravljanje napak v avtomatsko generiranih semantičnih leksikonih tipa wordnet in je zasnovano tako, da nam odgovore za problematične literale omogoča zbirati iz široke množice uporabnikov. Orodje je prosto dostopno in temelji na razširjenih tehnologijah, kot sta PHP in MySQL. Sestavljata ga administratorski in uporabniški vmesnik. V administratorskem vmesniku izdelamo projekt, sledimo poteku projekta in izvažamo rezultate, v uporabniškem vmesniku pa reševalci glasujejo o (ne)pravilnosti naključno izbranih literalov. Rezultati prvega eksperimenta, ki smo ga izvedli z orodjem sloWCrowd, so spodbudni, saj so bili uporabniki orodja z njim zadovoljni, odločitev o dokončnem izbrisu nekega literala na podlagi ujemanja njihovih odgovorov pa enostavna, hitra in zanesljiva. Dodatna prednost razvitega orodja je, da ga je mogoče prilagoditi za najrazličnejše naloge, pri katerih je koristno sodelovanje večjega števila reševalcev.

sloWCrowd: a Crowdsourcing Tool for Cleaning Wordnets

The paper presents a tool called sloWCrowd, developed to facilitate error correction in automatically generated wordnets by crowdsourcing. The developed tool is open-source and based on popular technologies, such as PHP and MySQL. It consists of an administrator and a user interface. The administrator interface enables the creation of crowdsourcing projects, management of on-going projects and export of the results, while the user interface allows users to vote on the (in)correctness of the randomly displayed literals. The results of the first experiment that was performed to test the sloWCrowd tool are encouraging because the users were satisfied with the tool and the final decision on the deletion of a problematic literal based on the users' inter-annotator agreement was simple, fast and reliable. Another advantage of the tool is that it can be adapted to a broad range of other crowdsourcing tasks.

1. Uvod

Z razmahom avtomatskih pristopov za izdelavo jezikovnih virov, ki glede na zahtevnost problema dajejo rezultate različne kakovosti, so se v jezikovnotehnološki skupnosti povečale tudi potrebe po validaciji oz. čiščenju avtomatsko generiranih vsebin. Ker je tovrstno delo zamudno in drago, so raziskovalci kmalu začeli razmišljati o možnostih, ki bi postopek pospešile in pocenile, pri čemer se kvaliteta zbranih oznak ne bi bistveno znižala. Številni poskusi so pokazali, da je nalogo mogoče razdeliti na obvladljive in razumljive dele 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 vprašanj različnim uporabnikom in filtriranjem njihovih odgovorov (Adda idr. 2011).

Ena najbolj razširjenih platform za uporabo moči množic (ang. *crowdsourcing*) za pridobivanje večje količine ročno potrjenih podatkov je Mechanical Turk[1] ameriškega podjetja Amazon, ki raziskovalcem ponuja administrativno podporo pri izvajanju eksperimentov, po želji pa tudi rekrutacijo reševalcev. S tovrstnimi platformami so zelo zadovoljni predvsem raziskovalci, ki zbirajo večje količine nejezikovnih podatkov (npr. označevanje slik), ter raziskovalci, ki se ukvarjajo z angleščino, saj imajo ti na spletu na voljo največ kompetentnih reševalcev. Z ustrezno zasnovanimi in izvedenimi nalogami pa so rezultati zelo uporabni tudi za zbiranje anotacij za zahtevnejše jezikoslovne in semantične probleme, kot so določanje afekta, presojanje semantične podobnosti besed, prepoznavanje besedilne vsebovanosti, časovno razvrščanje dogodkov, razdvoumljanje ipd. (Snow idr. 2008).

Za motivacijo povprečnih uporabnikov svetovnega spleta, da se pridružijo eksperimentu in prispevajo čim več odgovorov, so raziskovalci z različnih področij razvili t.i. igre z razlogom (ang. *games with a purpose*), ki od uporabnika na zabaven, a strukturiran način pridobivajo željene podatke. Ena prvih takšnih iger je bila ESP Game[2], v kateri sta uporabnika, ki se med seboj nista poznala, morala opisovati slike in zbirala točke vsakič, ko sta pri tem uporabila iste besede (von Ahn 2006). Med projekti, s katerimi so zbirali oznake za jezikovne podatke, pa je najbolj znana igra Word Detectives[3], s pomočjo katere označujejo anafore v besedilih (Chamberlain idr. 2008).

Povod za ta prispevek je bila potreba po odpravljanju napak iz ročno zgrajenega semantičnega leksikona za slovenščino sloWNet (Fišer 2009). Leksikon je zasnovan na sorodnem angleškem viru Princeton WordNet (Fellbaum 1998) in je bil grajen v več korakih s pomočjo

[1] https://www.mturk.com/mturk/welcome [15.5.2012]

[2] Igro je pred nekaj leti kupilo podjetje Google in jo vključilo v svoje produkte, zato je na spletu v prvotni obliki ni več mogoče igrati.

[3] http://anawiki.essex.ac.uk/phrasedetectives/ [15.5.2012]

različnih tipov razpoložljivih dvo- in večjezičnih jezikovnih virov, kot so dvojezični slovarji, vzporedni korpusi in Wikipedija. Analiza vsebine je pokazala, da tako izdelan sloWNet vsebuje precej šuma, ki znižuje uporabno vrednost vira in ga je zato treba čim prej odpraviti. Najbolj problematične lekseme (literale), ki skoraj zagotovo ne ubesedujejo pojma (sinseta), ki so mu pripisane, smo identificirali avtomatsko (Sagot in Fišer, 2012). Za to smo uporabili referenčni korpus FidaPLUS[4], iz katerega smo izluščili kontekstualne informacije za literale iz sloWNeta. V skladu z načeli distribucijske semantike smo nato kontekstualne informacije, pridobljene iz korpusa, primerjali z neposredno okolico literala v sloWNetovi semantični mreži. Kandidate z najslabšim rezultatom smo označili kot potencialne napake, ki jih želimo s pomočjo orodja, ki ga predstavljamo v tem prispevku, ponuditi v glasovanje večjemu številu slovenskih uporabnikov interneta in nato po potrebi izbrisati.

V nadaljevanju prispevka predstavimo orodje sloWCrowd, ki smo ga razvili za zbiranje jezikovnih podatkov iz široke množice uporabnikov. Razdelek 3 predstavi eksperiment, v katerem smo orodje prvič preizkusili. Prispevek zaključimo s sklepnimi mislimi in načrti za prihodnje delo.

2. Orodje sloWCrowd

Z razvojem orodja sloWCrowd smo želeli pridobiti preprosto in prilagodljivo orodje, ki bi omogočalo uporabo množic za verifikacijo avtomatsko generiranih sinsetov in bi bilo uporabno tudi za najrazličnejše naloge na področju gradnje jezikovnih virov in razvoja orodij, pri katerih je potrebno zbrati večje število človeških odgovorov.

Za vire, ki jih gradimo z avtomatskimi pristopi, je značilno, da vsebujejo precej napak, vendar so tudi zelo obsežni, zato bi bilo strokovno in celovito ročno odpravljanje napak preveč zamudno in predrago. S prenosom bremena verifikacije na širšo množico se zmanjša čas verifikacije, dobljeni rezultati pa so lahko celo bolj zanesljivi, saj se o posameznem literalu odloča večje število uporabnikov. Pri tovrstnem načinu zbiranja zanesljivih podatkov je orodje, ki na preprost in zanimiv način izbere ter ponudi naloge uporabnikom, ključno.

Orodje sloWCrowd je sestavljeno iz dveh delov, administratorskega in uporabniškega. V administratorskem delu ustvarimo projekt in vodimo zbiranje odgovorov. Uporabniški del 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 in 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. sloWCrowd deluje na večini spletnih brskalnikov, saj uporablja splošno razširjene tehnologije.

2.1. Zasnova in implementacija

sloWCrowd je spletna aplikacija, napisana v skriptnem jeziku PHP, kar nam omogoča izdelavo dinamičnih spletnih strani, ki so osnova za široko paleto storitev, ki jih danes najdemo na spletu. Prikaz vsebine omogočajo predloge HTML in CSS, zaradi česar je posamezne komponente orodja enostavno prilagajati trenutnim potrebam.

Podatki spletne aplikacije so shranjeni v odprtokodni podatkovni bazi MySQL. Struktura baze je definirana tako, da omogoča enostavno dodajanje novih projektov. V glavni tabeli so informacije o posameznih projektih, vsak projekt posebej pa ima dodeljeni še dve specifični tabeli. Prva vsebuje uporabniške podatke, druga pa uporabnikove odgovore na naloge.

Zaradi zagotavljanja kvalitete zbranih odgovorov smo sloWCrowd zasnovali tako, da se uporabniki pred reševanjem nalog najprej prijavijo v sistem. Prijava nam omogoča beleženje odgovorov uporabnikov, računanje zanesljivosti odgovorov in uporabnikov ter upravljanje z uporabniki. Registracija je zelo preprosta in poteka preko Googlovega računa. Za dostop do Googlove identifikacijske aplikacije smo uporabili odprtokodno PHP knjižnico HybridAuth[5], ki omogoča dostop do večine današnjih socialnih omrežij. Uporabnik s potrditvijo dostopa do prijavne aplikacije sistemu dovoli identifikacijo s podatki iz Googlovega računa, preko katerega orodje dobi uporabnikovo ime in elektronski naslov. Tak način registracije uporabniku prihrani vpisovanje osebnih podatkov in olajša prijavo v sistem, saj mu ni potrebno vsakič vpisovati uporabniškega imena in gesla.

sloWCrowd ima implementiran mehanizem za ugotavljanje zanesljivosti uporabnikov. Slednje je ključno pri obliki reševanja problemov, kjer sodeluje veliko različnih uporabnikov (od ekspertov do običajnih uporabnikov). Pri vsakem projektu je datoteki z nalogami, ki jih rešujejo uporabniki, mogoče dodati še datoteko, ki vsebuje referenčne naloge, t.j. naloge s predhodno označenimi pravilnimi odgovori. Uporabnik med reševanjem dobiva naloge iz obeh datotek, pri čemer referenčna datoteka služi za ugotavljanje zanesljivosti uporabnika:

$$\text{zanesljivost} = \frac{\text{število pravilnih odgovorov}}{\text{število odgovorov}}$$

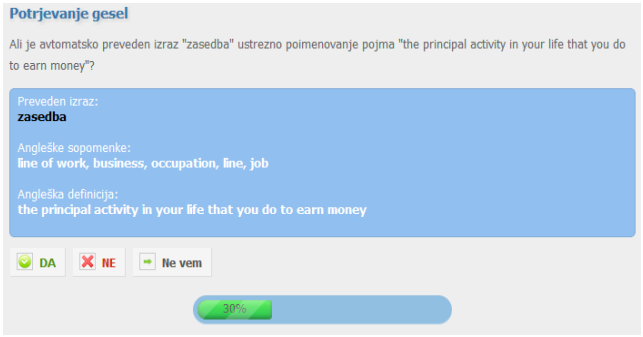
Pravilen odgovor je tisti, ki je enak odgovoru v referenčni datoteki, število odgovorov pa vključuje vse odgovore na naloge iz referenčne datoteke. Mero zanesljivosti uporabljamo za določanje razmerja med nerešenimi in referenčnimi nalogami, ki jih orodje ponudi uporabniku, omogoča pa tudi naknadno izločanje odgovorov zelo nezanesljivih uporabnikov.

[4] http://fidaplus.net/ [15.5.2012]

[5] http://hybridauth.sourceforge.net/index.html [15.5.2012]

2.2. Uporabniški vmesnik

Osnovna funkcionalnost orodja sloWCrowd je potrjevanje in zavračanje literalov. Po prijavi uporabnika se prikaže glavno okno, kjer uporabnik rešuje naloge. Na Sliki 1 je prikazan primer take naloge. Na vrhu zaslona so navodila za reševanje naloge, v konkretnem projektu preverjanje ustreznosti avtomatsko prevedenega slovenskega izraza za izbran koncept. Nato so za lažje odločanje navedene še dodatne informacije, kot na primer angleške sopomenke in angleška defiinicija za isti koncept. Na dnu zaslona so trije gumbi, med katerimi izbira uporabnik, in sicer DA, NE in NE VEM, s katerim vprašanje preskoči.



Slika 1. Potrjevanje in zavračanje literalov

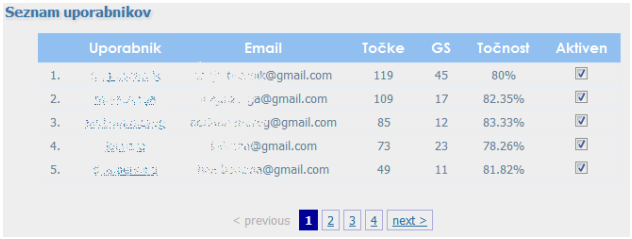
V vsakem sklopu se uporabniku prikaže 10 naključno izbranih vprašanj. Delež rešenih vprašanj v posameznem sklopu je prikazan na dnu zaslona. Orodje uporabniku ponudi določen delež novih in že rešenih nalog, s katerimi se ugotavlja zanesljivost uporabnika. Glede na pravilnost uporabnikovih odgovorov se uporabniku 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 točke dobi za vsak pravilen odgovor glede na referenčno datoteko in odgovore ostalih uporabnikov v bazi.

2.3. Administratorski vmesnik

V orodje je vključen tudi administratorski vmesnik, ki je namenjen skrbnikom orodja. V njem lahko urejajo aktivne projekte in definirajo nove. Skrbnik doda nov projekt tako, da vnese ime, opis projekta, v sistem naloži referenčno datoteko odgovorov, datoteko še nerešenih odgovorov in izbere eno od besedilnih datotek, v kateri je definirana celotna tekstovna vsebina projekta. V primeru, da skrbnik želi definirati projekt v drugem jeziku ali vzpostaviti projekt z drugimi funkcionalnostmi, le izbere ustrezno besedilno datoteko. Ob potrditvi se v bazi avtomatsko kreirajo tabele, ki jih projekt potrebuje in uporabniki lahko začnejo z reševanjem nalog projekta. Preostale funkcionalnosti so skupne vsem definiranim projektom in služijo pregledu poteka reševanja.

Prva funkcionalnost vmesnika je pregled vseh registriranih uporabnikov, razvrščenih po številu točk, ki so jih dosegli, njihova zanesljivost, na zahtevo pa tudi prikaz posameznih odgovorov. Uporabnike, ki ne dosežejo zadovoljive zanesljivosti, je mogoče onemogočiti z izklopom kljukice v polju Aktiven, s čimer se njegovi odgovori pri prikazu ne upoštevajo. Primer seznama uporabnikov je prikazan na Sliki 2.



Slika 2. Pregled vseh uporabnikov (imena uporabnikov so prekrita zaradi varovanja osebnih podatkov)

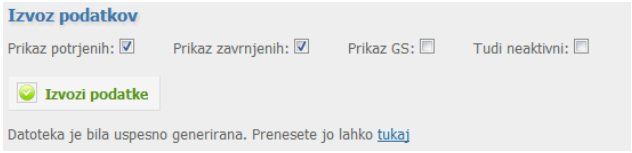
Naslednja funkcionalnost je prikaz seznama vseh rešenih nalog, število odgovorov na posamezno nalogo ter število potrditev in zavrnitev s strani uporabnikov. Na Sliki 3 je prikazan del odgovorov uporabnikov, ki smo jih zbrali med validacijo avtomatsko generiranih prevodov v sloWNetu. Literali, ki imajo v stolpcu GS (gold standard) zeleno kljukico, so iz datoteke z referenčnimi odgovori. Iz števila potrditev in zavrnitev lahko vidimo, da že z razmeroma majhnim številom zbranih odgovorov na posamezno vprašanje skupni rezultat hitro konvergira k pravilni rešitvi. Opazimo lahko namreč, da so v večini primerov uporabniki izbrali enak odgovor. Naj omenimo, da je v izvedenem eksperimentu množica referenčnih odgovorov precej večja od množice neoznačenih nalog, zato je število zbranih odgovorov za posamezni literal iz referenčne množice precej manjše od števila odgovorov na nove literale. Na Sliki 3 tako lahko vidimo, da so iz referenčne datoteke samo trije primeri: en odgovor za literal »simbol«, dva odgovora za literal »slovo« in eden za literal »soba«.

Slika 3. Seznam odgovorov

Orodje omogoča tudi enostavno filtriranje odgovorov, saj lahko skrbnik izbere, ali naj se literali iz referenčne datoteke prikažejo med rezultati ali ne. Pri izbiri posameznega literala se prikaže seznam, ki vsebuje vse odgovore uporabnikov, ki so se odločali o tem literalu. Za vsakega uporabnika se prikaže datum odgovora, ali je literal iz referenčne datoteke in pravilna rešitev. Vsakemu reševalcu je pripisana tudi njegova zanesljivost (glej Slika 4). Poleg tega lahko skrbnik izbira med prikazom uporabnikov, ki so literal zavrnili in uporabnikov, ki so literal potrdili.

Slika 4. Seznam odgovorov.

Zadnja funkcionalnost je izvoz vseh odgovorov v besedilno datoteko (Slika 5). Uporabnik lahko izbira, kateri podatki bodo vključeni v izvoz: samo potrditve ali zavrnitve literalov, ali naj bodo vključeni tudi odgovori na literale iz referenčne datoteke in odgovori uporabnikov, ki zaradi nezanesljivosti ne bodo upoštevani.



Slika 5. Izvoz podatkov

Izvoženi podatki so grupirani po literalih, kar pomeni, da so odgovori vseh uporabnikov za isti literal izpisani skupaj. Izvoženim podatkom je poleg vseh informacij o literalu dodano še uporabniško ime ocenjevalca, datum odgovora in njegova odločitev.

3. Preizkus orodja sloWCrowd

3.1. Opis eksperimenta

Razvito orodje smo preizkusili v manjšem eksperimentu, v katerem smo 10 uporabnikov, večinoma študentov in diplomantov prevajalstva ter tujih jezikov, prosili, da se prijavijo v orodje in rešijo 5 sklopov nalog, od katerih vsak vsebuje po 10 vprašanj. Za reševanje smo jim dali 10 dni časa. Pri tem smo jim dali naslednja navodila za reševanje:

»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.«

Vprašanja so bila sestavljena iz 100 samostalniških literalov, ki so na podlagi prejšnje raziskave (Sagot in Fišer 2012) najbolj vprašljivi in najverjetneje napačni. Enak delež vprašanj, ki so se uporabnikom naključno prikazala, pa je bil iz referenčne datoteke, ki je vsebovala vnaprej rešene naloge in nam služi za izračun stopnje zanesljivosti posameznih uporabnikov.

3.2. Predstavitev in analiza rezultatov

Pregled in analizo rezultatov začnemo s predstavitvijo sodelujočih pri eksperimentu. V Tabeli 1 so prikazani uporabniki, ki so sodelovali pri eksperimentu. Drugi stolpec v tabeli vsebuje število vseh literalov, o katerih so se uporabniki odločali. Od teh je v tretjem stolpcu prikazano število neoznačenih in v četrtem število označenih literalov. V četrtem stolpcu je prikazana zanesljivost uporabnikov, izračunana na podlagi odgovorov za literale iz referenčne datoteke, v zadnjem stolpcu pa je prikazana točnost odgovorov uporabnika za še neoznačene literale. Slednja je izračunana na podlagi odgovorov vseh uporabnikov (ang. *inter-annotator agreement*), pri čemer privzamemo, da je večinsko mnenje pravilno, odstopanja od njega pa nepravilna (čeprav bi v teoriji lahko tudi večina prispevala napačne odgovore, tovrstnih težav pri tej nalogi ne pričakujemo v omembe vrednem številu primerov). Iz primerjave stolpcev Zanesljivost in Točnost lahko ugotovimo, da so vrednosti obeh podobne in relativno visoke.

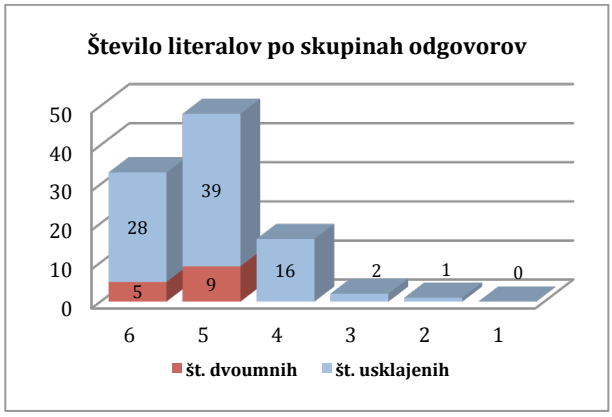
Upor.	Lit.	Neozn.	Ozn.	Zanesljivost	Točnost
U1	122	77	45	80,0 %	90,1 %
U2	110	94	16	82,3 %	91,5 %
U3	90	79	11	83,3 %	89,9 %
U4	78	56	22	78,3 %	87,5 %
U5	48	38	10	81,8 %	97,4 %
U6	49	35	14	92,8 %	88,6 %
U7	48	36	12	91,7 %	86,1 %
U8	48	36	12	91,7 %	91,7 %
U9	42	30	12	92,3 %	96,7 %
U10	40	29	11	81,8 %	75,9 %

Tabela 1. Pregled rezultatov, zbranih v eksperimentu

Iz tabele lahko opazimo, da so štirje uporabniki rešili veliko več nalog, kot je bilo od njih zahtevano, ostali pa nekaj nalog manj od zahtevanih 50. Ti uporabniki so verjetno nekajkrat izbrali možnost »Ne vem«, teh odgovorov pa v sistemu zaenkrat ne beležimo. Analiza odgovorov kaže, da so uporabniki so na splošno dosegli visoko zanesljivost v primerjavi z referenčno datoteko. Devet uporabnikov je doseglo zanesljivost, večjo od 80 %, štirje uporabniki pa celo zanesljivost nad 90 %. Povprečna dosežena zanesljivost pa znaša 85,6 %, kar je za naloge s področja leksikalne semantike zelo visok rezultat. Opazimo lahko tudi, da so uporabniki dosegli visoko stopnjo medsebojnega ujemanja. Tudi v tem primeru je le en uporabnik dosegel točnost, manjšo od 80 %, medtem ko povprečna točnost znaša 89,5 %. Na splošno pa so uporabniki dosegli večjo stopnjo točnosti kot zanesljivosti, kar pomeni, da so se bolj strinjali med seboj kot z referenčno množico. Opazimo pa trend, da se pri večini

uporabnikov z večanjem stopnje zanesljivosti veča tudi stopnja točnosti in obratno.

Pri analizi odgovorov nas je zanimalo tudi, kolikokrat so se literali med eksperimentom ponavljali. Na Sliki 6 je prikazana razporeditev literalov po pogostosti pojavitve. Opazimo lahko, da se je največ literalov (49) pojavilo petkrat, 33 literalov se je ponovilo šestkrat, 16 literalov se je ponovilo štirikrat, dva literala trikrat in en literal dvakrat. Na istem grafu je prikazano tudi število literalov, pri katerih so se uporabniki večinsko strinjali o pravilnosti pomena (zgornji del) in število dvoumnih literalov, za katere smo zbrali enako število potrditev in zavrnitev (spodnji del). V primeru petih odgovorov se prišteje literal med dvoumne tudi, če so trije uporabniki glasovali na en način, dva pa na drug.

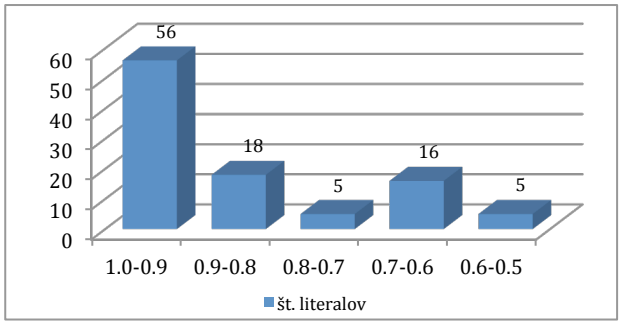


Slika 6. Število literalov po pogostosti pojavitve

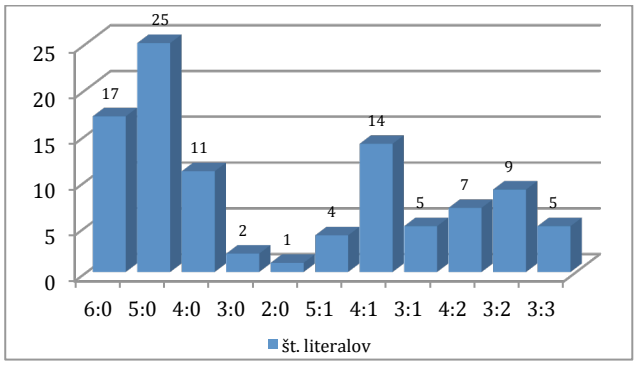
Opazimo lahko, da so se uporabniki v večini primerov strinjali o (ne)pravilnosti prevoda literala. Iz dobljenih rezultatov lahko sklepamo, da že z manjšim številom odgovorov dobimo zanesljivo oceno o nekem literalu, saj so že pri štirih odgovorih uporabniki dosegli konsenz glede pomena literala. Tega sicer zgolj na podlagi izvedenega eksperimenta ne moremo z gotovostjo potrditi, saj je število uporabljenih literalov premajhno, zato nameravamo v prihodnje minimalno število potrebnih odgovorov preveriti v večjem eksperimentu.

Pri nadaljnji analizi smo opazovali število literalov, ki so jih uporabniki potrdili ali zavrnili z določeno gotovostjo. Na Sliki 7 vidimo razporeditev literalov v skupine verjetnosti. V prvi so vsi literali, ki so bili s strani uporabnikov soglasno potrjeni ali zavrnjeni. V ostalih stolpcih pa so razporejeni literali s padajočim razmerjem med potrditvijo in zavrnitvijo. Opazimo lahko, da so bili uporabniki v večini primerov soglasni pri odločitvah o literalih, iz česar lahko sklepamo, da lahko z uporabo orodja sloWCrowd v večini primerov konvergiramo k splošno sprejemljivi pravilni rešitvi.

Na Sliki 8 je podrobneje prikazana razporeditev literalov po razmerjih odgovorov. Vsebinsko je podobna prejšnji sliki, le da natančneje prikaže razporeditev literalov po skupinah razmerij. Razmerje *x:y* predstavlja število potrditev (zavrnitev) *x* in število zavrnitev (potrditev) *y* za nek literal. Zadnja dva stolpca v grafu vsebujeta dvoumne literale, kjer je število potrditev in zavrnitev približno enako.



Slika 7. Število literalov po združenih deležih odgovorov



Slika 8. Število literalov po razmerju odgovorov

V nadaljevanju smo pregledali vse dvoumne literale. V Tabeli 2 so izpisani vsi literali, kjer je enako število potrditev in zavrnitev (3:3). V prvem stolpcu je prevod, nato sopomenke v angleščini in še angleška definicija.

Literal	Sopomenke	Definicija
hip	piece, while, spell, patch	a period of indeterminate length (usually short) marked by some action or condition
položaj	place, position	the particular portion of space occupied by something
prostor	room	opportunity for
razglas	rescript, fiat, order, edict, decree	a legally binding command or decision entered on the court record (as if issued by a court or judge)
vprašanje	topic, matter, subject, issue	some situation or event that is thought about

Tabela 2. Literali, za katere se uporabniki niso strinjali, ali so pravilni ali napačni.

Največ problemov predstavljajo literali, za katere je angleška razlaga precej ohlapna in zato nejasna. Naslednjo skupino problematičnih literalov predstavljajo angleške ustaljene fraze, ki se v slovenščini uporabljajo v drugem kontekstu ali z drugimi besedami. Med odgovori smo naleteli tudi na napačne odločitve. Dva primera, ki sta v resnici napačna, uporabniki pa so izglasovali, da sta pravilna, sta literala »del« (ang. *member*) v pomenu »*anything that belongs to a set or class*« in »člen« (ang. *division, part, section*) v pomenu »*one of the portions into which something is regarded as divided and which together constitutes a whole*«. Vendar je teh primerov zanemarljivo malo, pa še te glasovi drugih uporabnikov slej ko prej preglasujejo in s tem izničijo negativni vpliv napake na končne rezultate.

3.3. Diskusija in možnosti za izboljšave

Z eksperimentom, v katerem smo preizkusili razvito orodje sloWCrowd, smo ugotovili, da je njegova glavna prednost, da uporabnikom na preprost in zanimiv način ponudi v reševanje različne naloge. Kot je znano za večino iger z razlogom, se je tudi tu pokazalo, da vpeljava točkovanja uporabnike pritegne k pravilnemu reševanju večjega števila nalog, saj je kar nekaj tekmovalcev rešilo precej več nalog, kot je bilo od njih zahtevano. Čeprav na podlagi opravljenega eksperimenta težko zanesljivo določimo spodnjo mejo zahtevanih odgovorov, ki so potrebni za zanesljivo oceno literala, menimo, da 10 uporabnikov ob nizki obremenitvi in v krajšem času lahko validira 200–300 literalov, kar pomeni, da bi za ocenjevanje celotnega sloWNeta potrebovali 300–400 uporabnikov.

Da bi lahko orodje v prihodnje še izboljšali in nadgradili, smo prostovoljce, ki so v eksperimentu reševali naloge, dodatno prosili še, da nam posredujejo odgovore na naslednja vprašanja (možni odgovori NITI NAJMANJ, NE PREVEČ, PRECEJ, ZELO):

- Ali se ti je zdelo delo z orodjem sloWCrowd enostavno?
- Ali je zdelo delo z orodjem sloWCrowd zanimivo?
- Ali so se ti zdele naloge razumljive?
- Ali je bila dolžina posamezne naloge primerna?
- Ali imaš v zvezi z orodjem oz. eksperimentom kakšno pripombo, ki nam bo pomagala orodje še izboljšati?

Večina uporabnikov je menila, da je delo z orodjem sloWcrowd PRECEJ enostavno in zanimivo, nekaj uporabnikov je celo ocenilo enostavnost in zanimivost orodja z ZELO. Podobne ocene so uporabniki dodelili razumljivosti podanih nalog, s tem, da je en uporabnik podal oceno NE PREVEČ, z obrazložitvijo, da je pri veliko nalogah odgovor odvisen od konteksta, v katerem se literal uporablja, predvsem v primeru nejasnih definicij. Po njegovem mnenju zgolj prikaz angleških sopomenk in definicij v teh primerih ne zadošča, zato predlaga, da omogočimo še prikaz drugih semantičnih relacij, ki izhajajo iz ocenjevanega pojma, kar bomo v naslednji različici orodja tudi upoštevali. Podobne pomisleke so imeli tudi nekateri drugi uporabniki, saj je na primer eden od njih predlagal, da uvedemo še dodaten gumb »odvisno od konteksta«. Taka rešitev bi verjetno zmanjšala število zaželenih odločitev (DA, NE), saj bi se uporabniki v primeru dvoma odločali zanjo. Vendar bi s tem po našem mnenju po nepotrebnem dodatno zapletli nalogo in s tem zmanjšali uporabnost rezultatov. Menimo, da bomo tovrstne težave ustrezno odpravili že z zgornjim ukrepom. Bi pa bilo možno, da bi literale, za katere večje število uporabnikov izbere odgovor »NE VEM«, pregledal ekspert. Naslednja želja, ki jo je izrazil eden od uporabnikov, je povratna informacija, ali je bil izbrani odgovor pravilen. To bi v naslednji različici orodja lahko upoštevali tako, da bi uporabniku ob vsakem odgovoru sporočili, ali si je z odgovorom prislužil dodatno točko ali ne ter prikazali trenutno razmerje med zbranimi odgovori. Zadnja pripomba se nanaša na možnost predlaganja boljših prevodov, kar je pravzaprav funkcionalnost že razvitega orodja sloWTool (Fišer in Novak 2011), ki je namenjeno celovitemu urejanju wordneta. Glavni cilj orodja sloWCrowd je predvsem minimalistično in enostavno okolje za čim hitrejše pridobivanje odgovorov na ozko specializirana vprašanja.

4. Zaključek

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 eksperimentu, s katerim smo orodje testirali, smo uporabnike prosili, da pregledajo 100 najbolj vprašljivih avtomatsko generiranih literalov iz sloWNeta in označijo, ali so pravilni ali napačni. Po zaključenem preizkusu in analizi njihovih odgovorov ugotavljamo, da je orodje tako z administrativnega kot z uporabniškega vidika enostavno za uporabo, zbrani odgovori pa zanesljivi, saj je stopnja ujemanja med uporabniki visoka in število dvoumnih rešitev majhno. S tem tako nismo dobili samo orodja, s katerim je mogoče popravljati wordnet, temveč tudi platformo za evalvacijo uspešnosti različnih pristopov in avtomatskih metod za avtomatizirano luščenje leksikalno-semantičnih informacij iz strukturiranih in nestrukturiranih jezikovnih virov.

Glede na komentarje prostovoljcev, ki so pri preizkusu sodelovali, bomo v prihodnje še nekoliko izboljšali točkovanje uporabnikov, na večjem eksperimentu pa skušali ugotoviti optimalno število potrebnih odgovorov na vsako zastavljeno vprašanje za zagotavljanje zanesljivih rezultatov po eni stran in čim večje količine validiranih primerov po drugi. Projekti, ki trenutno tečejo v orodju sloWCrowd, so dostopni na naslovu http://nl.ijs.si/slowcrowd/select_project.php. Ker je orodje prosto dostopno pod licenco Creative Commons, pa 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.

Literatura

G. Adda, B. Sagot, K. Fort, J. Mariani. 2011. Crowdsourcing for Language Resource Development: Critical Analysis of Amazon Mechanical Turk Overpowering Uses. *Zbornik konference LTC*, Poznań, Poljska.

L. von Ahn. Games with a Purpose. 2006. *Computer*, 39/6, str. 92-94.

J. Chamberlain, M. Poesio, U. Kruschwitz. 2008. Phrase Detectives: A Web-based collaborative annotation game. *Zbornik konference iSemantics*, Gradec, Avstrija.

D. Fišer, J. Novak. 2011. Visualizing sloWNet. *Zbornik konference eLEX*. Bled, Slovenija.

D. Fišer. 2009. Pristopi za avtomatizirano gradnjo semantičnih zbirk. *Terminologija in sodobna terminografija*. Ljubljana: Založba ZRC, ZRC SAZU, str. 357-370.

B. Sagot, D. Fišer. 2012. Cleaning noisy wordnets. *Zbornik konference LREC*, Istanbul, Turčija.

R. Snow, B. O'Connor, D. Jurafsky, A. Y. Ng. 2008. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. *Zbornik konference EMNLP*, str. 254-263.