Avtomatska razširitev in čiščenje sloWNeta

Darja Fišer, Benoît Sagot

Avtomatska razširitev in čiščenje sloWNeta

Darja Fišer,* Benoît Sagot†

* Oddelek za prevajalstvo, Filozofska fakulteta
Aškerčeva 2, 1000 Ljubljana, Slovenija
† UMR-I ALPAGE, UFR de Linguistique de l'Université Paris Diderot
Case 7003, 75205 Paris Cedex 13, France

Povzetek

V prispevku predstavljamo jezikovno neodvisno in avtomatsko razširitev wordneta z uporabo heterogenih že obstoječih jezikovnih virov, kot so strojno berljivi slovarji, vzporedni korpusi in Wikipedija. Pristop, ki ga preizkusimo na slovenščini, upošteva tako eno- kot večpomenske besede, splošno in specializirano besedišče, pa tudi eno- in večbesedne lekseme. Izluščenim besedam enega ali več pomenov pripišemo s pomočjo klasifikatorja, ki temelji na naboru različnih značilk, predvsem pa na distribucijski podobnosti. V naslednjem koraku s pomočjo distribucijskih informacij, izluščenih iz velikega korpusa, identificiramo in odstranimo zelo dvomljive kandidate. Avtomatska in ročna evalvacija rezultatov pokaže, da uporabljeni pristop daje zelo spodbudne rezultate.

Automatic extension and cleaning of sloWNet

In this paper we present a language-independent and automatic approach to extend a wordnet by recycling different types of already existing language resources, such as machine-readable dictionaries, parallel corpora and Wikipedia. The approach, applied to Slovene, takes into account monosemous and polysemous words, general and specialized vocabulary as well as simple and multi-word lexemes. The extracted words are assigned one or several synset ids based on a classifier that relies on several features including distributional similarity. In the next step we also identify and remove highly dubious (literal, synset) pairs, based on simple distributional information extracted from a large corpus in an unsupervised way. Automatic and manual evaluation show that the proposed approach yields very promising results.

1. Uvod

Avtomatski pristopi k izdelavi wordnetov so se uveljavili, ker je ročna izdelava časovno preveč potratna, da bi bila uresničljiva v večini raziskovalnih scenarijev, prav tako pa so avtomatske in polavtomatske metode že dale zadovoljive rezultate v številnih projektih, kot so npr. EuroWordNet, BalkaNet in Asian WordNet Vossen 1999, Tufis 2000, Sornlertlamvanich 2012). Pristop, predstavljen v tem prispevku, poleg avtomatske izdelave odlikuje predvsem dejstvo, da zanjo niso potrebni nobeni specializirani ali kompleksni algoritmi ali orodja, ki obstajajo samo za jezikovno-tehnološko najbolje podprte jezike, jezikovno odvisna pravila ali dragi leksikalni viri.

Z izkoriščanjem že obstoječih heterogenih jezikovnih virov, kot so strojno berljivi slovarji, vzporedni korpusi in Wikipedija, s predstavljenim pristopom maksimiziramo količino izluščenih leksikalnih informacij iz vsakega uporabljenega vira. Za razliko od večine sorodnih raziskav, ki temeljijo zgolj na Wikipediji, lahko s tem pristopom zajamemo vse besedne vrste, ne zgolj samostalnikov. Pristop je celovit tudi v smislu obravnave tako eno- kot večpomenskih, pa tudi eno- in večbesednih leksemov. Predstavljeni pristop je nadgradnja razvoja prvih različic slovenskega wordneta (Fišer in Sagot 2008), v okviru katere smo izboljšali tako obseg kot tudi natančnost razširjenega vira.

V 2. razdelku povzamemo sorodne raziskave, v 3. opišemo luščenje kandidatov za razširitev sloWNeta s pomočjo klasifikatorja, v 4. pa predstavimo filtriranje nezanesljivih kandidatov z uporabo načel distribucijske semantike. V 5. razdelku rezultate ročno in avtomatsko ovrednotimo, prispevek pa sklenemo z diskusijo in načrti za prihodnost.

2. Sorodne raziskave

Avtomatski pristopi izgradnje wordneta se med seboj razlikujejo predvsem glede na vrsto vira, ki ga za gradnjo uporabljajo. Najstarejši pristopi so temeljili na strojno berljivih dvojezičnih slovarjih, ki so služili za prevajanje sinsetov v Princeton WordNetu pod predpostavko, da slovarski prevodi predstavljajo isti pojem v ciljnem jeziku (Knight in Luk 1994, Yokoi 1995). Glavna ovira tega pristopa je, da dvojezični slovarji tipično niso pojmovno zasnovani, temveč slonijo na tradicionalnih leksikografskih načelih, kar otežuje razdvoumljanje slovarskih iztočnic. Pogosto je problematičen tudi njihov obseg oz. slovar za določen jezikovni par sploh ni na voljo.

Te slabosti so presegli različni pristopi, ki za prevajanje sinsetov uporabljajo dvo- in večjezične leksikone, izluščene iz vzporednih korpusov (Resnik in Yarowsky 1997, Fung 1995). Osrednja predpostavka tovrstnih pristopov je, da se pomeni večpomenskih besed v izvornem jeziku pogosto prevajajo v različne besede v ciljnem jeziku. Po drugi strani pa velja, da če se dve ali več besed v izvirniku prevajajo v isto besedo v ciljnem jeziku, imajo le-te pogosto skupne pomenske elemente. Posledično je mogoče identificirati pomenske razlike polisemnih besed oz. besede z enakim pomenom združiti v sopomenske nize, kar so dokazali Dyvik (2002), Ide idr. (2002) in Diab (2004).

Tretja skupina pristopov, ki so postali popularni v zadnjih nekaj letih, pa za iskanje prevodnih ustreznic uporablja Wikipedio, obsežno spletno enciklopedijo, ki v številnih jezikih nastaja s sodelovanjem zainteresiranih uporabnikov svetovnega spleta. Z njihovo pomočjo so raziskovalci zgradili nove wordnete s povezovanjem Wikipedijinih strani z najpogostejšim pomenom v Princeton WordNetu (Suchanek 2008), z izkoriščanjem

Wikipedijinih kategorij in drugih strukturnih informacij (Ponzetto in Navigli 2009) ter z luščenjem ključnih besed iz Wikipedijinih člankov (Reiter idr. 2008). Ruiz-Casado idr. (200) in Declerck idr. (2006) so s pomočjo modela vektorskega prostora mapirali Wikipedijine strani na WordNet. Najnaprednejši pristopi pa Wikipedijo in sorodne projekte, kot je npr. Wiktionary, uporabljajo za indukcijo wordnetov za številne jezike hkrati (de Melo 2009, Navigli in Ponzetto 2010, Navigli in Ponzetto 2012).

S pristopom, predstavljenim v pričujočem prispevku, smo za razširitev slovenskega wordneta uporabili vse vire, ki jih imamo na voljo: splošne in specializirane dvojezične slovarje, vzporedne korpuse in Wiki vire. Predstavljeni pristop je nadgradnja osnovne različice algoritma, za katere so bili uporabljeni isti viri (Erjavec in Fišer 2006, Fišer in Sagot 2008), v kateri smo osnovni pristop izboljšali tako, da deluje tudi za luščenje prevodnih ustreznic večpomenskih besed, kar omogoči poln izkoristek virov, ki so na voljo, pri čemer visoko stopnjo natančnosti zagotavlja ponderiranje kandidatov za sinsete glede na izbrane značilke.

3. Razširitev sloWNeta

Motivacija za razširitev sloWNeta izhaja iz dejstva, da smo pri izdelavi prve različice besede iz vzporednega korpusa razdvoumili s pomočjo ostalih jezikov v korpusu, medtem ko smo iz slovarjev in Wikipedije zaradi pomanjkanja ustreznih večjezičnih ali strukturnih informacij uporabili le enopomenske lekseme (Fišer in Sagot 2008), s čimer je precejšen delež dragocenih leksikosemantičnih informacij ostal neizkoriščen. Pristop, ki smo ga uporabili za razširitev sloWNeta tudi s temi informacijami, in tako bistveno izboljšali njegovo pokritost, opisujemo v tem razdelku, v naslednjem pa predstavimo varnostni mehanizem, s katerim smo kljub razširitvi v sloWNetu zagotovili visoko stopnjo natančnosti.

3.1. Verjetnostni klasifikator

Predstavljeni pristop temelji na verjetnostnem klasifikatorju, ki za odločanje o razvrščanju neke besede iz dvojezičnega slovarja oz. Wikipedije v obstoječi sloWNet uporablja niz značilk. Učno množico za klasifikator smo izdelali tako, da smo za vse izluščene pare (literal, sinset) – se pravi besedo v določenem pomenu –, ki že obstajajo v prejšnji različici sloWNeta, privzeli, da so pravilni, za vse ostale pa, da so nepravilni. Tako izdelana učna množica seveda ni popolna, saj po eni strani kot ustrezne pare (literal, sinset) obravnava napake, podedovane iz avtomatično generirane prve različice sloWNeta, po drugi pa tudi povsem ustrezne pare (literal, sinset) obravnava kot napačne, samo zato, ker se v prejšnji različici niso pojavili. Naš cilj v predstavljeni raziskavi je identificirati ravno te in z njimi razširiti wordnet.

Uporabili smo klasifikator največje entropije megam (Daume 2004) in s pomočjo značilk, opisanih v naslednjem razdelku, v razširjen wordnet vključili vse pare (literal, sinset), ki presegajo eksperimentalno določen prag verjetnosti 0,1.

3.2. Izbor značilk

Najpomembnejša značilka modelira semantično bližino med nekim literalom in potencialnimi sinseti. Naj jo ponazorimo na primeru angleško-slovenskega para *organ-organ*:

- Angleški leksem *organ* se v PWN 3.0. pojavi v 6 različnih sinsetih, zato smo za ta dvojezični par tudi generirali 6 različnih kandidatov (literal, sinset).
- Naša naloga je, da ugotovimo, kateri od teh 6 kandidatov so ustrezni, torej v katere sinsete v sloWNetu je potrebno dodati slovenski literal *organ*.
- Semantično bližino slovenskega literala z vsakim od 6 možnih sinsetov smo izračunali tako, da smo za vsak sinset izdelali vektor, v katerega smo vključili vse literale iz tega slovenskega sinseta in iz vseh sinsetov, ki so od osrednjega oddaljeni največ dve koleni.
- Tako na primer potencialen sinset {*organ*, *pipe organ*} iz PWN predstavlja naslednji slovenski vektor: {*glasbilo*, *Anton Bruckner*, *glasbenik*, *Johann Sebastian Bach*, *pisalni*, *klavirska*, *harmonika*,...}.
- Podoben vektor smo za potencialno slovensko prevodno ustreznico *organ* zgradili s pomočjo programskega pakega SementicVectors (Widdows in Ferraro 2008) iz korpusa FidaPLUS in ju nato primerjali v skladu z načeli distribucijske semantike.
- Semantična podobnost potencialne slovenske ustreznice *organ* s sinsetom {*organ*, *pipe organ*} znaša le 0.021, medtem ko primerjava s sinsetom {*organ, a fully differentiated structural and functional unit in an animal that is specialized for some particular function*}, znaša 0.668, kar je tudi jezikovno ustrezna rešitev.

Poleg mere semantične podobnosti smo uporabili tudi nekatere druge značilke, kot so število vseh angleških iztočnic, ki imajo pripisano isto slovensko ustreznico, najnižja stopnja večpomenskosti med vsemi angleškimi iztočnicami s pripisano isto slovensko ustreznic, število virov, iz katerih smo dvojezični par izluščili, in dolžina kandidata za prevodno ustreznico. Kot je razvidno iz tabele 1, se je v skladu s pričakovanji kot najbolj relevantna značilka izkazala semantična podobnost, saj ima največjo utež. H končnemu rezultatu pozitivno prispevata tudi indeks najnižje stopnje večpomenskosti za angleške literale in število različnih angleških literalov, ki imajo pripisano isto prevodno ustreznico. Po drugi strani pa na verjetnost ustreznosti kandidata za določen sinset negativno vpliva število pojavnic v slovenskem literalu.

Značilka	Utež
Semantična podobnost	6,24
Št. virov	0,55
Št. ang. literalov z isto ustreznico	0,33
Min. večpomenskost za ang. literal	2,69
Št. besed v slo. literalu	-1,87
Vir: Wikipedija	0,92
Vir: ang. Wiktionary	0,27
Vir: slo. Wiktionary	-0,07
Vir: SpeciesWiki	0,10
Vir: ang-slo slovar	0,15
Vir: slo-ang slovar	0,79

Tabela 1. Modeli za razvrščanje novih kandidatov (literal, sinset), naučeni na osnovnem wordnetu.

3.3. Rezultati klasifikacije

Naučen model smo uporabili za klasifikacijo 685.633 kandidatov, ki smo jih izluščili iz dvojezičnih slovarjev, vzporednega korpusa in Wikipedije. Mejni prag 0,1 je preseglo 68.070 kandidatov. Med njimi je 5.056 (7 %) takšnih, ki so že obstajali v prejšnji različici sloWNeta, novih pa je 63.010 (93 %) kandidatov. To pomeni, da je 25.102 sinsetov, ki so bili doslej prazni, dobilo vsaj en literal. Razširjena različica ima tako 141 % več neprazniih sinsetov kot pred razširitvijo. Število parov (literal, sinset) pa je še večje, saj smo jih z razširitvijo dobili 82.721, kar pomeni povečanje za 244 %.

4. Filtriranje sloWNeta

Kljub spodbudnim rezultatom postopek ni popoln, zato novi sinseti vsebujejo tudi precej šuma. Tega smo želeli odpraviti z jezikovno neodvisnim korpusnim pristopom za detekcijo in filtriranje potencialnih napak v avtomatsko generiranih sinsetih, s čimer bi dobili čistejši in uporabnejši semantični leksikon.

Čiščenje temelji na metodah distribucijske semantike za merjenje semantične podobnosti med besedami (Lin idr. 2003), vendar naš cilj ni prepoznavanje najbolj sorodnih besed glede na sobesedilo, v katerem se pojavljajo, temveč izhajamo iz (nepopolnega) seznama sinonimov, na katerem iščemo tiste, ki nanj ne sodijo. To pomeni, da je naloga podobna tistim na področju leksikalne substitucije (Mihalcea idr. 2010), pri čemer nas najbolj zanimajo kandidati, ki so na seznamu rangirani najnižje. Poleg tega je tudi naše razumevanje sinonimije strožje, saj je naš cilj očistiti vse sinsete v avtomatsko generiranem wordnetu, za katerega je znano, da ima pomene zelo nadrobno razdelane. Za to nalogo je ključno, da je naše dojemanje polisemije prevodno motivirano. To pomeni, da ne glede na število sinsetov, v katerih se beseda pojavi, so za naše potrebe pomenske razlike relevantne le, če se v ciljnem jeziku leksikalizirajo različno. Pri tem naj poudarimo še, da smo čiščenje wordneta zaenkrat sicer izvedli za vse besedne vrste, a zgolj za enobesedne literale (saj je stopnja večpomenskosti za večbesedne literale zelo nizka, medtem ko je procesiranje večbesednih literalov zahtevnejše, zato jih v nadaljevanju ne omenjamo, čeprav smo jih izluščili, predvsem iz Wikipedije, in vključili v razširjen sloWNet).

Naš cilj je identificirati in izločiti najočitnejše napake v sinsetih, ki so se v njem pojavile zaradi napačne besedne poravnave vzporednega korpusa ali napačnega razdvoumljanja homonimov, saj ravno tovrstne napake najbolj znižujejo uporabno vrednost wordneta. Zato predstavljeni pristop temelji na preprosti tezi: leksemi (oz. pari (literal,sinset)) se v korpusi sopojavljajo s semantično povezanimi leksemi, ki so eksplicitno kodificirani s semantičnimi relacijami v wordnetu. Pristop je sestavljen iz dveh korakov:

- primerjava kontekstualne podobnosti leksemov v referenčnem korpusu s sinseti v wordnetu,
- globalna razvrstitev vseh parov (literal, sinset) glede na dobljene rezultate.

Za vsak par (literal, sinset) smo najprej zgradili vektor, v katerega smo vključili vse literale iz vseh sinsetov, ki so s ciljnim povezani z ročno izbranimi semantičnimi relacijami (hipernimija, hiponimija, holonimija, meronimija, derivacija) v razdaji 0-2. Nato smo za vsakega od teh literalov zgradili še kontekstni vektor iz referenčnega korpusa, v katerega smo vključili vse polnopomenske besede, ki se pojavijo v istem odstavku. V zgrajenih vektorjih smo nato primerjali stopnjo prekrivanja kontekstov z algoritmom, zelo podobnim Leskovemu, ki je klasična mera za razdvoumljanje večpomenskih besed s pomočjo slovarjev (Lesk 1986). Pare (literal, sinset) smo nato rangirali tako, da smo za vsak iteral, ki je povezan s sinsetom in se pojavi v istem odstavku, rezultat povečali za število pojavitev v korpusu, deljeno s številom sinsetov, v katerih se v wordnetu pojavi. S tem smo dali manjšo težo zelo polisemnim literalom. Rezultat smo nato normalizirali še s številom polnopomenskih besed v odstavku.

Če pristop ponazorimo na primeru literala *ikona*, opazimo, da se v sloWNetu pojavi v 4 sinsetih, med drugim tudi v teh dveh:

- *eng-30-07269916-n {icon}*; ikona je v tem primeru ustrezen prevod, v vektorju sinseta pa so naslednji povezanimi literali: *znak, točka, simbol, računalništvo...*,
- *eng-30-03931044-n {icon, ikon, image, picture}*; ikona v tem primeru ni ustrezen prevod, v vektorju sinseta pa so naslednji povezani literali: *fotografija, podoba, predstaviti, prikaz...*

V korpusu FidaPLUS se samostalnik *ikona* pojavi 3.488-krat. Seštevek vseh rezultatov za pojavitev besede ikona v korpusu za pravilen par *(ikona, eng-30-07269916-n)* glede na zgoraj omenjen vektor znaša le 1,02, medtem ko globalni rezultat za nepravilen par *(ikona, eng-30-03931044-n)* znaša 5,99, kar pomeni, da tak globalni rezultat ni učinkovit indikator napak v razširjenem wordnetu. Zato smo ga nadgradili tako, da smo ga normalizirali z vsoto vseh globalnih rezultatov za par (literal, sinset) za celotni sinset, kar meri prispevek določenega literala med vsemi literali v sinsetu. Prispevek literala smo nato normalizirali še s številom pojavitev tega literala v korpusu, dobljen rezultat pa je hkrati tudi končni rezultat.

V našem prejšnjem primeru globalni rezultat za celotni sinset za pojem *eng- 30-07269916-n* znaša 1,02, za pojem *eng-30-03931044-n* pa 234, medtem ko prispevka literala za ta sinseta znašata 1 in 0,026. Normaliziran prispevek literala oz. končni rezultat za par *(ikona, eng-30-07269916-n)* znaša 0,287 za par *(ikona, eng-30-03931044-n)* pa le 0,007, s čimer je napačen kandidat v razširjenem wordnetu ustrezno identificiran.

Glede na izmerjeno 18-odstotno stopnjo napake razširjenega sloWNeta, smo za mejni prag pri filtriranju sloWNeta določili takšno vrednost, ki iz njega izloči primerljiv delež parov (literal, sinset). Ta znaša $4 \cdot 10{-}6$ in iz sloWNeta izloči 12,578 parov oz. tretjino vseh identificiranih potencialnih napak.

5. Vrednotenje rezultatov

Vrednotenje rezultatov smo opravili ročno in avtomatsko, pri čemer smo ročno evalvirali razširjen sloWNet in detekcijo napak, za avtomatsko vrednotenje pa smo izdelan vir primerjali z manjšim zlatim standardom ter z avtomatsko generiranima večjezičnima viroma Universal WordNet (de Melo 2009) in BabelNet 2.0 (Navigli in Ponzetto 2010, Navigli in Ponzetto 2012), ki vsebujeta tudi slovenščino.

5.1. Ročno vrednotenje razširjenega sloWNeta in avtomatske detekcije napak

Ročno vrednotenje razširjenega sloWNeta smo opravili na vzorcu 100 naključnih parov (literal, sinset) iz razširjenega sloWNeta za vsako besedno vrsto. Na podlagi teh rezultatov lahko podamo skupno oceno celotnega wordneta, in sicer tako, da rezultate za posamezno besedno vrsto ponderiramo z relativnim številom parov (literal, sinset). Kot je razvidno iz tabele 2, smo najvišjo stopnjo natančnosti izmerili za prislove (96%), najnižjo za glagole (59%), medtem ko skupni rezultat za razširjen sloWNet 3.0 znaša 82%.

Bes. vrsta	Σ parov	% parov	% pravilnih	% nepravilnih
sam.	55.383	67 %	87 %	13 %
prid.	12.438	15 %	85 %	15 %
gl.	14.053	17 %	59 %	41 %
prislov	847	1 %	96 %	4 %
Skupaj	82.721	100 %	82 %	18 %

Tabela 2. Rezultati ročnega vrednotenja razširjenega sloWNeta.

Ročno vrednotenje detekcije napak v razširjenem sloWNetu smo opravili na 100 naključnih parih (literal, sinset), ki jih je algoritem prepoznal kot napačne. Stopnja natančnosti tega avtomatskega postopka je 64 %.

Glede na to, da smo za razširitev sloWNeta uporabili prag, s katerim smo iskali optimalno razmerje med priklicem in natančnostjo, smo v leksikon vnesli tudi nekaj šuma. Zato je spodbudno, da z detekcijo potencialnih napak identificiramo kandidate, med katerimi je 64 % dejanskih napak. Ne samo, da z algoritmom najdemo več napak, kot bi jih s pregledovanjem naključnih kandidatov, temveč algoritem uspešno izkoristi razširjeno mrežo, ki doslej ni bila na voljo.

5.2. Avtomatsko vrednotenje razširjenega sloWNeta

Avtomatska primerjava z ročno izdelanim zlatim standardom, izdelanim z ročno validacijo prve različice sloWNeta, ki je temeljila na srbskem wordnetu (Erjavec in Fišer 2006) in vsebuje osnovni nabor sinsetov, pokaže 70 % natančnost, kar je sicer manj, kot je pokazala ročna evalvacija v prejšnjem razdelku, a je treba poudariti, da zlati standard vsebuje predvsem osnovni nabor sinsetov, ki vsebujejo zelo splošno besedišče, za katerega je značilna visoka stopnja večpomenskosti, tako da je za ta segment besedišča avtomatska naloga bistveno težja.

Nekoliko drugačna evalvacija, ki ovrednoti predvsem pristop, s katerim smo sloWNet izdelali, pa je primerjava s sorodnimi leksikosemantičnimi viri, in sicer z BabelNetom (Navigli in Ponzetto 2010, Navigli in Ponzetto 2012) in Universal WordNetom (de Melo 2009). Čeprav vsi trije viri temeljijo na primerljivih virih, je bil za razliko od nas osnovni cilj UWN in BabelNeta izdelati večjezično semantično mrežo.

Zato je razumljivo, da je slovenski del UWN z 9.924 pari (literal, sinset) precej manjši od sloWNeta, ki jih vsebuje 82.721. Kot je razvidno iz tabele 3, 5.590 (56 %) od 9.924 slovenskih parov v UWN vsebuje tudi sloWNet 3.0.

Podrobni rezultati					
		so v BabelNetu 2.0		niso v BabelNetu 2.0	
		so v UWN	niso v UWN	so v UWN	niso v UWN
so v sloWNetu 3.0	št. parov	2.239	12.468	3.351	64.663
	natančnost	98 %	98 %	92 %	86 %
niso v sloWNetu 3.0	št. parov	901	116.367	3.433	-
	natančnost	100 %	70 %	72 %	-
Primerjava z BabelNetom 2.0					
		so v Babelnetu 2.0		niso v Babelnetu 2.0	
so v sloWNetu 3.0	št. parov	14.707		68.014	
	natančnost	98 %		86 %	
niso v sloWNetu 3.0	št. parov	117.257		-	
	natančnost	70 %		-	
Primerjava z UWN					
		so v UWN		niso v UWN	
so v sloWNetu 3.0	št. parov	5.590		77.131	
	natančnost	94 %		88 %	
niso v sloWNetu 3.0	št. parov	4.334		-	
	natančnost	78 %		-	
Pregled posameznih virov					
	sloWNet 3.0	BabelNet 2.0		UWN	
št. parov	82.721	131.964		9.924	
natančnost	88 %	73 %		87 %	

Tabela 3. Primerjava razširjenega sloWNeta z BabelNetom in UWN.

Po drugi strani pa sloWNet vsebuje še 77.131 (93 %) parov, ki jih v UWN ni. Za razliko od UWN BabelNet 2.0 vsebuje kar 131.964 parov. Med njimi jih je 14.707 (11 %) tudi v sloWNetu 3.0. 69.014 (82 %) parov iz sloWNeta ni v BabelNetu. Če primerjamo vse tri vire, kar 64.663 parov najdemo samo v sloWNetu, po drugi strani pa je samo 901 parov tako v BabelNetu in UWN, v sloWNetu pa manjkajo.

Za boljšo predstavo o kvaliteti primerjanih virov smo ročno ovrednotili po 50 naključnih parov (literal, sinset) za vsakega od zgoraj naštetih scenarijev. Skupna natančnost za sloWNet znaša 88 %, kar je primerljivo z UWN, venar je UWN precej manjši. Natančnost 64.663 parov, ki jih najdemo samo v sloWNetu, je 86 %, kar je bistveno več kot natančnost parov, ki so samo v UWN (72 %), in tistih, ki so samo v BabelNetu (72 %).

Pristopi, ki so bili uporabljeni za gradnjo teh treh virov, so komplementarni, saj so praktično vsi pari, ki jih vsebujejo vsi trije viri, pravilni (2,239), zelo kvalitetni pa so tudi pari, ki si jih delita po dva vira (92 %).

BabelNet, sicer zelo obširen vir, je manj zanesljiv, saj je pravilnih samo 70 % parov, ki jih ni v sloWNetu, v primerjavi z 78 % pari v UWN, ki jih ni v sloWNetu.

Napake v sloWNetu poleg napačne disambiguacije so povezane z zastarelim dvojezičnim slovarjem, ki vsebuje precej arhaičnih izrazov, napake v parih, ki jih najdemo samo v UWN in BabelNetu, pa večinoma tičijo v napačni normalizaciji, kot so neprevedeni ang. izrazi, naslovi strani v Wikipediji, ki niso literali (*Seznam Arheoloških Dob*), semantično ustrezne slovenske besede, a napačne besedne vrsto, v ženski obliki, se začnejo s številko, končajo s piko ali disambiguatorjem (npr. *Mars (bog)*). Glede na opravljeno primerjalno analizo ugotavljamo, da je sloWNet 3.0 najkvalitetnejši leksikosemantični vir za slovenščino, ki je trenutno na voljo.

6. Zaključek

V prispevku smo predstavili avtomatsko širitev in čiščenje slovenskega semantičnega leksikona sloWNet s pomočjo različnih že obstoječih več- in dvojezičnih virov, kot so dvojezični slovarji, vzporedni korpusi in Wiki viri. S širitvijo smo dragocene leksikosemantične informacije v njih izkoristili v največji možni meri, ne samo večpomenskih iz korpusa in enopomenskih iz slovarja in Wikipedije, kot je to bilo storjeno v gradnji prejšnje verzije sloWNeta. Za pripisovanje pravega pomena večpomenskim besedam smo si pomagali s klasifikatorjem, ki je za odločanje uporabljal različne značilke, predvsem pa distribucijsko podobnost. S širitvijo smo število nepraznih sinsetov v sloWNetu povečali s 17.817 na 42.919, število parov (literal, sinset) pa je z 24.081 poskočilo na 82.721 (+244 %).

Ročno in avtomatsko vrednotenje tako razširjenega wordneta pokaže 85 % natančnost in ima bistveno večji priklic v primerjavi s prejšnjo različico. Razširjen sloWNet smo naložili v prosto dostopno spletno orodje za brskanje, editiranje in vizualizacijo wordneta, sloWTool (Fišer in Novak 2011), s čimer je na voljo študentom, prevajalcem in drugim jezikoslovcem, celotna podatkovna zbirka pa je dostopna pod licenco Creative Commons BY-SA (s priznanjem avtorstva in deljenjem pod enakimi pogoji): http://nl.ijs.si/sloWNet.

V prihodnje želimo izboljšati luščenje leksikosemantičnih informacij iz Wikipedije in Wiktionarija, s čimer bi lahko v sloWNet dodali definicije in primere rabe. Pristop želimo razširiti tudi na primerljive korpuse (Fišer idr. 2012), ki jih je veliko lažje pridobiti s spleta kot vzporedne. Prav tako pa si pristop, ki se je že izkazal kot učinkovit za gradnji francoskega wordneta (Sagot in Fišer 2008), želimo preizkusiti še na hrvaščini.

Literatura

H. Daume III. 2004. *Notes on CG and LM-BFGS optimization of logistic regression*.

T. Declerck, A.G. Pérez, O. Vela, Z. Gantner, D. Manzano-Macho. 2006. Multilingual lexical semantic resources for ontology translation. *Zbornik konference International Conference on Language Resources and Evaluation (LREC '06)*. Genova, Italija.

M. Diab. 2004. The feasibility of bootstrapping an Arabic wordnet leveraging parallel corpora and an English wordnet. *Zbornik konference Arabic Language Technologies and Resources.*

H. Dyvik. 2002. Translations as semantic mirrors: from parallel corpus to wordnet. *Zbornik konference ICAME '02*. Gothenburg, Švedska.

T. Erjavec, D. Fišer. 2006. Building Slovene wordnet. *Zbornik konference International Conference on Language Resources and Evaluation (LREC '06).* Genova, Italija.

D. Fišer, B. Sagot. 2008. Combining Multiple Resources to Build Reliable Wordnets. *Zbornik konference Text, Speech and Dialogue (TSD '08)*. Brno, Češka.

D. Fišer, N. Ljubešić, O. Kubelka. 2012. Addressing polysemy in bilingual lexicon extraction from comparable corpora. *Zbornik konference International Conference on Language Resources and Evaluation (LREC '12)*. Istanbul, Turčija.

Fišer, D., Novak, J. 2011. Visualizing sloWNet. *Zbornik konference Electronic lexicography in the 21st century: new applications for new users (eLEX '11)*. Bled, Slovenija.

P. Fung. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. *Zbornik konference Association for Computational Linguistics*. Stroudsburg, PA, ZDA.

N. Ide, T. Erjavec, D. Tufis. 2002. Sense discrimination with parallel corpora. *Zbornik konference Association for Computational Linguistics, Workshop on word sense disambiguation: recent successes and future directions.* Stroudsburg, PA, ZDA.

K. Knight, S. K. Luk. 1994. Building a large-scale knowledge base for machine translation. *Zbornik konference Artificial intelligence*. Menlo Park, CA, ZDA.

D. Lin. 1998. An information-theoretic definition of similarity. *Zbornik konference Machine Learning*. Madison, Wisconsin, ZDA.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *Zbornik conference 5th annual international conference on systems documentation (SIGDOC '86)*, 24-26.

G. de Melo, G. Weikum. 2009. Towards a universal wordnet by learning from combined evidence. *Zbornik konference Information and knowledge management, (CIKM '09)*.

R. Mihalcea, R. Sinha, D. McCarthy. 2010. Semeval-2010 task 2: Cross-lingual lexical substitution. *Zbornik delavnice 5th international workshop on semantic evaluation.*

R. Navigli, S. P. Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. *Zbornik konference Association for Computational Linguistics*, Uppsala, Švedska.

R. Navigli, S. P. Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193, 217–250.

R. Navigli, S. P. Ponzetto. 2009. Large-scale taxonomy mapping for restructuring and integrating wikipedia. *Zbornik konference Artifical intelligence (IJCAI '09)*. San Francisco, CA, USA.

N. Reiter, M. Hartung, A. Frank. 2008. A Resource-Poor Approach for Linking Ontology Classes to Wikipedia Articles. *Zbornik konference Semantics in Text Processing (STEP '08)*.

P. Resnik, D. Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. *Zbornik delavnice Tagging Text with Lexical Semantics: Why, What, and How?*. Washington, D.C., ZDA.

M. Ruiz-Casado, E. Alfonseca, P. Castells. 2005. Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. *Zbornik konference Advances in Web Intelligence*.

B. Sagot, D. Fišer. 2008. Building a free French wordnet from multilingual resources. *Zbornik konference Ontolex 2008*. Marakeš, Maroko.

V. Sornlertlamvanich. 2010. Asian wordnet: Development and service in collaborative approach. *Zbornik konference Global WordNet Association (GWC '10)*. Mumbaj, Indija.

F. M. Suchanek, G. Kasneci, G. Weikum. 2008. Yago: A large ontology from Wikipedia and WordNet. *Journal of Web Semantics 6(3)*, 203–217.

D. Tufis. BalkaNet – Design and Development of a Multilingual Balkan WordNet. 2000. *Romanian Journal of Information Science and Technology Special Issue* 7(1–2).

P. Vossen. 1999. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.

D. Widdows, K. Ferraro. 2008. Semantic vectors: a scalable open source package and online technology management application. *Zbornik konference International Conference on Language Resources and Evaluation (LREC '08)*. Marrakech, Morocco.

T. Yokoi. 1995. The EDR electronic dictionary. *Zbornik konference ACM 38 (11)*, 42–44.