AVTOMATSKO LUŠČENJE HRVAŠKO-SLOVENSKEGA LEKSIKONA IZ PRIMERLJIVIH KORPUSOV

Darja Fišer
Filozofska fakulteta, Ljubljana

Nikola Ljubešić
Filozofski fakultet, Zagreb

UDK 81'322.4=163.42=163.6:81'374:004.91

V prispevku predstavljamo metodo za avtomatsko luščenje hrvaško-slovenskega leksikona iz primerljivega časopisnega korpusa s predpostavko, da se besede in njihove prevodne ustreznice pojavljajo v podobnih sobesedilih. Izhodiščni leksikon za primerjavo kontekstnih vektorjev z izkoriščanjem podobnosti med jezikoma zgradimo kar iz korpusa, nato pa opravimo še razvrščanje rezultatov glede na stopnjo sorodnosti med izvorno besedo in njenimi prevodnimi kandidati. Rezultati so zelo spodbudni in odpirajo številne možnosti uporabe za druge sorodne jezike.

primerljivi korpusi, spletni korpusi, dvojezični leksikoni, luščenje prevodnih ustreznic, sorodnice

In this paper we present a method for extracting a bilingual lexicon for closely related languages from comparable corpora. We take advantage of the similarities between languages to build a seed lexicon to compare context vectors in both languages and use cognates for reranking translation candidates. The results are very encouraging, suggesting that other similar languages could benefit from the same approach.

comparable corpora, web corpora, bilingual lexica, extraction of translation equivalents, cognates

1 Uvod

Za večino večjezičnih računalniških aplikacij (npr. strojno prevajanje) nujno potrebujemo dvojezične leksikone, njihova izdelava pa je izjemno dolgotrajna in draga. Zato so se v računalniškem jezikoslovju uveljavile metode, ki dvojezične leksikone avtomatsko luščijo iz vzporednih korpusov (Och, Ney 2000). Vendar to še vedno ni mogoče za jezikovne pare in strokovna področja, za katere ne obstajajo vzporedni korpusi. Za te primere se je v zadnjem desetletju pojavil pristop, ki za avtomatsko iskanje prevodnih ustreznic uporablja primerljive korpuse (Fung 1998; Rapp 1999). Pristop je privlačen, ker je veliko več besedil, ki govorijo o isti temi, so bila objavljena v istem obdobju, imajo isti komunikacijski namen in jih je lažje zbrati kot vzporedna besedila, še posebej z vse bogatejšega svetovnega spleta (Xiao, McEnery 2006).

Pristop temelji na predpostavki, da se beseda in njen prevod pojavljata v podobnih sobesedilih, npr.:

> *Zbog opasnosti od bijega sudac […] odbio je zahtjev za **jamčevinom** […] od milijun dolara koji ostaje u pritvoru do daljnjega*. (www.jutarnji.hr)
> *Sodniki so zaradi begosumnosti zavrnili ponudbo […] odvetnikov po plačilu **varščine** v višini milijon dolarjev*. (www.delo.si)

To pomeni, da lahko najdemo prevod neke besede z iskanjem besede v ciljnem korpusu, ki ima najbolj podoben kontekstni vektor kot beseda v izvornem korpusu. Vendar neposredna avtomatska primerjava vektorjev

v dveh različnih jezikih ni mogoča, zato potrebujemo dvojezični slovar, s katerim izvorni kontekstni vektor najprej prevedemo v ciljni jezik in nato primerjamo dva vektorja v istem jeziku. To pa je tudi najbolj paradoksalen del na primerljivih korpusih temelječega pristopa luščenja prevodnih ustreznic, saj sploh ne bi bilo smiselno uporabljati tako kompleksnega postopka, če bi že imeli na voljo obsežen dvojezični slovar; s tem se sorodne raziskave praviloma ne ukvarjajo.

V pričujoči raziskavi predstavljamo nadgradnjo opisane metode, ki za iskanje prevodnih ustreznic ne potrebuje nobenega že obstoječega slovarskega vira, temveč omogoča avtomatsko gradnjo izhodiščnega leksikona za prevajanje kontekstnih vektorjev neposredno iz korpusa, in sicer z izkoriščanjem podobnosti med izvornim in ciljnim jezikom. Pristop preizkusimo na primerljivem časopisnem korpusu za jezikovno kombinacijo hrvaščina-slovenščina, za katero ni bilo doslej za raziskovalne namene na voljo nobenega dvojezičnega leksikona v računalniško berljivi obliki.

2 Pregled sorodnih raziskav

Večina sorodnih raziskav uporablja za prevajanje kontekstnih vektorjev kar obstoječe dvojezične slovarje; alternativnih pristopov, ko takšen slovar ni na voljo, ni veliko. Enega izmed prvih poskusov sta izvedla Koehn in Knight (2002), ki sta za prevajanje vektorjev namesto slovarja uporabila besede, ki so identične v nemškem in angleškem delu korpusa. Nekoliko drugače so se naloge lotili Al-Onaizan in Knight (2002) ter Shao in Ng (2004), ki so za prevajanje iz angleščine uporabili transliteracijska pravila za arabščino oz. kitajščino, kar je še posebej učinkovito za avtomatsko prevajanje lastnih imen in novih besed, ki jih še ni v slovarjih. Podobno so Markó idr. (2005) uporabili substitucijska pravila za luščenje medicinskih sorodnic med španščino in portugalščino (npr. *ph* → *f*, *ph*os*ph*or → *f*os*f*or). Učinkovita nadgradnja standardnega pristopa je tudi ponovno razvrščanje prevodnih kandidatov na podlagi podobnosti znakovnih nizov prevodnih kandidatov in izvorne besede (Saralegi idr. 2008).

Predstavljena raziskava je podobna pristopu Koehna in Knighta (2002), saj tudi mi za gradnjo izhodiščnega leksikona uporabljamo identične besede v obeh jezikih. Za razliko od njiju pa izhodiščni leksikon dopolnimo še s sorodnicami in prevodi najpogostejših besed v korpusu. Na koncu izvedemo še razvrščanje prevodnih kandidatov na podlagi sorodnic, podobno kot Saralegi idr. (2008). Pričujoča raziskava se od sorodnih razlikuje po tem, da v njej uporabljamo bistveno večji korpus in precej bolj podobna jezika, kar vpliva na boljši priklic in natančnost izluščenih prevodnih ustreznic, tako zgrajen dvojezični leksikon je zato tudi bolj uporaben v praksi. Pomembna prednost prestavljenega pristopa je tudi ta, da se ne omejujemo zgolj na samostalnike (kot večina sorodnih raziskav), temveč luščimo prevodne ustreznice za vse polnopomenske besede.

3 Gradnja uporabljenih virov

Naš cilj je avtomatsko luščenje prevodnih ustreznic za splošno besedišče, zato smo zgradili hrvaško-slovenski primerljivi časopisni korpus, ki pokriva različne splošne teme. Besedila za hrvaški del primerljivega korpusa smo pridobili iz hrvaškega spletnega korpusa hrWaC, ki vsebuje 1 milijardo pojavnic; za slovenski del smo uporabili sorodni slovenski spletni korpus slWaC, ki vsebuje 380 milijonov pojavnic (Ljubešić, Erjavec 2011). Iz njiju smo izluščili vse dokumente, objavljene na domenah www.jutranji.hr in www.delo.si; tj. spletne izdaje dnevnih časopisov s primerljivo visoko naklado in podobno ciljno publiko. Dokumenti so že bili tokenizirani, oblikoskladenjsko označeni in lematizirani ter vsebujejo 13,4 milijona različnic v hrvaščini in 15,8 milijona različnic v slovenščini.

Za hrvaščino in slovenščino za zdaj žal ni nobenega slovarja v računalniško berljivi

obliki, zato smo morali izhodiščni leksikon za prevajanje kontekstnih vektorjev izdelati sami. Za olajšanje dela smo se odločili izkoristiti visoko stopnjo podobnosti med jezikoma. Da njuna podobnost ni zgolj navidezna, je dokazal Scannell (2007), ki je primerjal številne svetovne jezike, tako da je s kosinusom meril podobnost trigramov v posameznih jezikih. Hrvaščina in slovenščina sta dosegli 74-odstotno podobnost; podobnost med njima je enkrat večja kot 34-odstotna podobnost med angleščino in nemščino, ki sta ju v svoji raziskavi uporabljala Koehn in Knight (2002). Podobne rezultate kot hrvaščina in slovenščina so dosegle še češčina in slovaščina (70-odstotna podobnost) ter španščina in portugalščina (76-odstotna podobnost). To pomeni, da bi bilo mogoče tudi pri teh jezikovnih parih pomanjkanje slovarskih virov do neke mere nadoknaditi z izkoriščanjem jezikovnih podobnosti med njimi.

Glede na visoko stopnjo podobnosti med hrvaščino in slovenščino se nam je zdelo smiselno izdelati izhodiščni leksikon za prevajanje kontekstnih vektorjev kar iz primerljivega časopisnega korpusa. V izhodiščni leksikon smo tako vključili vse leme iz korpusa, ki so bile identične v obeh jezikih in so imele pripisano isto besednovrstno oznako (npr. *agresor*, *litij*, *nuklearka*). Med njimi je tudi veliko lastnih imen (npr. *Sisak*, *Kučan*, *Thyssenkrupp*) in tujih besed, ki so bile v korpusu rabljene citatno (npr. *chill*, *mortal*, *gypsy*).

Preglednica 1: Ročna evalvacija izhodiščnega leksikona

Besedna vrsta	Št. vnosov	Natančnost
samostalniki	25.703	88 %
pridevniki	4042	76 %
glagoli	3315	69 %
prislovi	435	54 %
skupaj	33.495	69 %

Kot kaže Preglednica 1, vsebuje zgrajeni izhodiščni leksikon 33.495 vnosov, med katerimi je 77 % samostalnikov. Ročni pregled vzorca naključnih 100 vnosov za vsako besedno vrsto kaže, da so med avtomatsko izluščenimi vnosi najkakovostnejši samostalniki (88-odstotna natančnost) in pridevniki (76-odstotna natančnost), precej manj pa glagoli (69-odstotna natančnost) in prislovi (54-odstotna natančnost). Napake, na katere smo naleteli pri ročnem vrednotenju, so večinoma hrvaške besede, ki se pojavljajo v slovenskem delu korpusa (npr. *baka*, *svjetski*, *usuditi se*). Najverjetneje izvirajo iz komentarjev, ki jih bralci pogosto pišejo v pogovornem jeziku, v katerega vključujejo tudi nekatere hrvaške izraze. V prihodnje nameravamo te besede natančneje filtrirati. Resnejše težave bi utegnili pri nadaljnjih korakih povzročati lažni prijatelji (npr. pridevnik *neslužben* v hrvaščini pomeni »neuraden«, v slovenščini pa »ki ni povezan s službo«). Tovrstne primere nameravamo v prihodnje prepoznavati z merjenjem podobnosti sobesedil, v katerih se pojavljajo.

4 Luščenje prevodnih ustreznic

S pomočjo zgrajenega izhodiščnega leksikona smo v primerljivem korpusu iskali slovenske prevodne ustreznice za hrvaške polnopomenske besede. Vektorje smo zgradili za vse besede, ki se v korpusu pojavljajo vsaj 50-krat, pri čemer je bila širina upoštevanega konteksta 7 besed, sopojavitve pa smo normirali z logaritmom verjetnosti. Hrvaške kontekstne vektorje smo nato prevedli s pomočjo izhodiščnega leksikona, jih primerjali s slovenskimi kontekstnimi vektorji in s pomočjo Jensen-Shannonove divergence izračunali podobnost med njimi.

Pri avtomatskem in ročnem vrednotenju rezultatov smo za vsako hrvaško besedo upoštevali 10 najbolje uvrščenih prevodnih kandidatov. Ročno vrednotenje smo opravili na 100 naključnih prevodnih ustreznicah za vsako besedno vrsto, za avtomatsko vrednotenje pa smo izdelali referenčni leksikon, v

katerega smo vključili 500 naključnih slovarskih vnosov za posamezno besedno vrsto iz srbohrvaško-slovenskega slovarja (Jurančič 1986). V celotnem eksperimentu smo vselej uporabljali isti korpus in frekvenčni kriterij za gradnjo kontekstnih vektorjev, tako da je bil priklic vselej 54-odstoten; to pomeni, da smo za dobro polovico vnosov iz referenčnega leksikona z izbranimi nastavitvami našli vsaj eno prevodno ustreznico.

Natančnost merimo s srednjo vzajemno uvrstitvijo (ang. *mean reciprocal rank*; Vorhees 2001), ki poleg preverjanja, ali 10 najbolje uvrščenih prevodnih kandidatov vsebuje pravilen prevod, upošteva še, na katerem mestu se ta pojavi.[1] Natančnost izhodiščnega leksikona, v katerem so zgolj identične besede, je 0,597. Natančnost pa se je spreminjala z nadgraditvami pristopa – z razširjanjem izhodiščnega leksikona in ponovnim razvrščanjem prevodnih ustreznic.

4.1 Razširitev izhodiščnega leksikona s sorodnicami

Podobnosti med hrvaščino in slovenščino smo dodatno izkoristili tako, da smo v izhodiščni leksikon vključili še sorodnice (ang. *cognates*). To so besede, ki so si podobne po zunanji podobi in ponavadi tudi po pomenu (npr. *t**u**žba* → *t**o**žba*, *brani**telj*** → *brani**lec***, ***u**druženje* → ***z**druženje*). Sorodnice smo v korpusu identificirali avtomatsko, s funkcijo BI-SIM (Kondrak, Dorr 2004), ki za par besed v obeh jezikih izračuna najdaljši skupni sklop bigramov. Šumu smo se izognili z dodatnim pogojem, ki ni dovoljeval katerihkoli sorodnic v korpusu, temveč samo tiste, ki so se pojavljale v dovolj podobnih sobesedilih.

Kot kaže Preglednica 2, smo z opisanim postopkom identificirali 3159 kontekstno potrjenih sorodnic, med katerimi je približno polovica samostalnikov. Ročno vrednotenje rezultatov pokaže, da je luščenje sorodnic najbolj natančno pri pridevnikih (92-odstotna natančnost, npr. *digital**an*** → *digital**en***). Do napak prihaja, kadar končnica ne pomeni razlike med jezikoma, temveč izkazuje besedotvorni postopek (npr. *doktor* → *doktor**at***, *funkcija* → *funkci**onar***, *ekonomist* → *ekonomi**ja***). Zanimivo je, da je izluščenih sorodnic bistveno manj kot identičnih besed, vendar je po drugi strani njihova kakovost precej višja (v povprečju za 14 %). To je mogoče utemeljiti z različnima metodama, ki smo ju uporabili za luščenje: medtem ko je bil edini pogoj za luščenje identičnih besed popolno ujemanje črkovnih nizov, so morale sorodnice poleg pogoja zadostnega ujemanja črkovnih nizov zadostiti še pogoju kontekstne potrditve.

Preglednica 2: Ročna evalvacija sorodnic

Besedna vrsta	Št. vnosov	Natančnost
samostalniki	1560	84 %
pridevniki	779	92 %
glagoli	706	74 %
prislovi	114	85 %
skupaj	3159	83 %

Preglednica 3 vsebuje rezultate avtomatskega vrednotenja luščenja prevodnih ustreznic z izhodiščnim leksikonom, ki smo ga razširili s sorodnicami. Če v leksikon dodamo sorodnice za vsako besedno vrsto posebej, imajo največji prispevek k razširitvi leksikona samostalniške sorodnice (1560), kakovost izluščenih prevodov pa najbolj dvignejo pridevniške sorodnice (0,657). Najboljše rezultate dosežemo, če v leksikon dodamo vse sorodnice, saj s tem natančnost dvignemo za 0,088 (z 0,597 na 0,685).

4.2 Razširitev izhodiščnega leksikona s prevodnimi ustreznicami najpogostejših besed

Že v raziskavi Fišer idr. (2011) smo pokazali, da je prevajanje zelo pogostih besed v korpusu s to metodo zelo zanesljivo. Izhodiščni

1 Rezultat za mero srednje vzajemne uvrstitve je izražen v decimalkah, ker ne gre za odstotek natančnosti rezultatov, temveč za povprečno uvrstitev pravilnih odgovorov na seznamu več možnih kandidatov.

Preglednica 3: Avtomatsko vrednotenje luščenja prevodnih ustreznic z izhodiščnim leksikonom, razširjenim s sorodnicami

Besedna vrsta	Št. vnosov	Št. novih vnosov	Natančnost
izhodiščni leksikon	33.495	/	0,597
sorodnice (sam.)	34.089	1560	0,626
sorodnice (prid.)	33.999	779	0,657
sorodnice (gl.)	33.655	706	0,621
sorodnice (prisl.)	33.565	114	0,598
sorodnice (vse)	34.823	3159	0,685

leksikon smo se odločili dopolniti še s prevodnimi ustreznicami najpogostejših besed in preveriti, ali lahko z njihovo pomočjo izboljšamo luščenje dvojezičnih leksikonov. Upoštevali smo samo prve prevodne kandidate besed, ki se v korpusu pojavijo najmanj 200-krat (npr. *lanac → veriga*, *smještaj → bivanje*, *zabrana → prepoved*).

Preglednica 4: Ročno vrednotenje prvih prevodnih kandidatov najpogostejših besed

Besedna vrsta	Št. vnosov	Natančnost	Delež sorodnic
samostalniki	2510	71 %	48 %
pridevniki	957	57 %	38 %
glagoli	1002	63 %	30 %
prislovi	325	59 %	26 %
skupaj	4794	62 %	34 %

Z dodajanjem prevodnih ustreznic najpogostejših besed v korpusu smo izhodiščni leksikon razširili s 1635 vnosi več kot pri sorodnicah (tj. skupaj s 4794 vnosi). Vendar ročno vrednotenje rezultatov pokaže, da so precej slabše kakovosti (povprečno za 21 %). Skoraj 53 % izluščenih prevodov najpogostejših besed je samostalnikov, ki so tudi najboljši (71-odstotna natančnost).

Zanimivo je, da je med ročno pregledanimi prevodnimi ustreznicami veliko sorodnic, še posebej med samostalniki (48 %), kar še dodatno utemeljuje smiselnost uporabe sorodnic za luščenje prevodnih ustreznic med podobnimi jeziki (npr. *rijedak → redek*, *vlasnički → lastniški*, *nesretan → nesrečen*). Z analizo napak smo ugotovili, da je 23 % prevodov, ki sicer niso pravilni, semantično tesno povezanih z izvorno besedo (npr. *brat → sestra*, *hajdukov → olimpijin*, *pustiti → zapustiti*) in verjetno prav tako pozitivno vplivajo na modeliranje sobesedila, s tem pa tudi na kakovost luščenja dvojezičnega leksikona.

Preglednica 5 podaja rezultate avtomatskega vrednotenja luščenja prevodnih

Preglednica 5: Avtomatsko vrednotenje luščenja prevodnih ustreznic z izhodiščnim leksikonom, razširjenim s prvimi prevodi najpogostejših besed v korpusu

Besedna vrsta	Št. vnosov	Št. novih vnosov	Natančnost
izhodiščni	33.495	/	0,597
sorodnice (sam.)	33.964	2510	0,662
647 sorodnice (prid.)	33.967	957	0,652
sorodnice (gl.)	33.695	1002	0,641
sorodnice (prisl.)	33.818	325	0,611
sorodnice (vse)	34.817	4794	0,714

ustreznic s pomočjo izhodiščnega leksikona, ki smo ga razširili s prvimi prevodnimi ustreznicami najpogostejših besed v korpusu. Kot smo opazili že pri sorodnicah, k leksikonu največ prispevajo samostalniške ustreznice (2510 novih vnosov). Za razliko od sorodnic pa samostalniki v tem primeru tudi najbolj vplivajo na dvig kvalitete izluščenih prevodnih ustreznic (za 0,065). Najverjetneje je izboljšanje večje kot pri sorodnicah (za 0,036) zato, ker je med najpogostejšimi besedami v korpusu največ samostalnikov, iskanje prevodnih ustreznic zanje pa je najbolj uspešno, kar vpliva tudi na dvig celotne natančnosti.

Ko smo izhodiščnemu leksikonu dodali prve prevode najpogostejših polnopomenskih besed v korpusu, se je natančnost izboljšala za 0,117 (z 0,597 na 0,714) oz. za 0,029 več kot pri sorodnicah. To pomeni, da so najpogostejše besede v korpusu pomembnejše za prevajanje kontekstnih vektorjev, s tem pa tudi za kakovost izluščenega dvojezičnega leksikona; kljub temu, da je natančnost prevodov za najpogostejše besede nekoliko slabša.

4.3 Razširitev izhodiščnega leksikona s kombinacijo sorodnic in prevodov najpogostejših besed

Vpliv dodajanja novih informacij v izhodiščni leksikon smo preverili še s kombinacijo sorodnic in prevodov najpogostejših besed. V tem primeru smo v leksikon dodali 2303 nove vnose, tako da je bilo po razširitvi v leksikonu 35.798 vnosov; s tem smo dosegli 0,731 natančnost, kar je veliko izboljšanje izhodiščnih rezultatov (0,597). Ročno vrednotenje rezultatov pokaže, da smo za 88 besed našli ustrezen prevod med desetimi najbolje uvrščenimi prevodnimi kandidati; od teh vsebuje 64 besed ustrezen prevod na prvem mestu, 24 pa na preostalih devetih mestih. Opazili smo, da je med predlaganimi prevodnimi ustreznicami pogosto več kot ena ustrezna prevodna varianta. Med pravilnimi prevodi je prav tako kar 59 sorodnic, kar kaže na to, da bi bilo rezultate mogoče še dodatno izboljšati s ponovnim razvrščanjem predlaganih prevodnih kandidatov s pomočjo merjenja stopnje sorodnosti.

4.4 Ponovno razvrščanje prevodnih kandidatov glede na stopnjo sorodnosti

Deset najbolje uvrščenih prevodnih kandidatov smo ponovno razvrstili glede na stopnjo sorodnosti, ki jo izkazujejo v primerjavi z izhodiščno hrvaško besedo. Stopnjo sorodnosti smo tudi tukaj izračunali s funkcijo BI-SIM. Preglednica 6 vsebuje primerjavo rezultatov luščenja prevodnih ustreznic za posamezne besedne vrste z izhodiščnim leksikonom, z razširjenim leksikonom in s ponovnim razvrščanjem prevodnih kandidatov. Ponovno razvrščanje glede na stopnjo sorodnosti izboljša rezultate pri vseh besednih vrstah, razen pri prislovih, še posebej učinkovito pa je pri pridevnikih in samostalnikih (za 0,131 oz. 0,086 boljši rezultat kot pri razširjenem leksikonu).

S tem postopkom smo izboljšali rezultate za tiste besede, ki so med desetimi kandidati

Preglednica 6: Avtomatsko vrednotenje luščenja prevodnih ustreznic za posamezne besedne vrste po ponovnem razvrščanju

Besedna vrsta	Izhodiščni leksikon	Razširjeni leksikon	Ponovno razvrščanje
samostalniki	0,597	0,731	0,817
pridevniki	0,625	0,654	0,785
glagoli	0,453	0,479	0,516
prislovi	0,598	0,598	0,598
povprečje	0,559	0,601	0,655

vsebovale pravilen prevod, vendar ne na prvem mestu. Npr.: hrvaška beseda *poezija* je bila najprej prevedena s tesno povezano, vendar neustrezno slovensko ustreznico *proza*, po razvrščanju rezultatov glede na sorodnost pa je bila izbrana pravilna *poezija*; podobno je bila hrvaška beseda *dopuna* sprva prevedena kot *zakon*, potem pa z razvrščanjem popravljena v *dopolnitev*. Razvrščanje se je izkazalo kot zelo koristno tudi v primerih, ko neustrezen prevod sploh ni bil pomensko povezan s hrvaškim izvirnikom, npr. *evolucija* je bila sprva napačno prevedena z *revolucijo*, potem pa popravljena v *evolucijo*.

Neredko imamo po razvrščanju na prvih dveh mestih oba ustrezna prevoda: tujko, ki je sorodnica s hrvaško besedo, in slovensko dvojnico (npr. *ekran* → *ekran*, *zaslon*; *princip* → *princip*, *načelo*; *proizvod* → *proizvod*, *izdelek*). Rezultati ponovnega razvrščanja tako potrjujejo našo izhodiščno hipotezo, da izkoriščanje podobnosti med jeziki pozitivno vpliva na luščenje dvojezičnih leksikonov iz primerljivih korpusov brez uporabe zunanjih leksikalnih virov.

5 Sklep

V prispevku smo predstavili pristop za avtomatsko gradnjo dvojezičnih leksikonov iz primerljivih korpusov za sorodne jezike, ki ne temelji na obstoječih leksikalnih virih. Ko smo pristop preizkusili na časopisnem korpusu za slovenščino in hrvaščino, smo dosegli boljše rezultate kot sorodne raziskave, tako z vidika natančnosti (0,801 za samostalnike in pridevnike) kot priklica (54 %). Za razliko od večine sorodnih raziskav, ki se ukvarjajo samo s samostalniki, smo v predstavljenem eksperimentu luščili prevode vseh polnopomenskih besed, pri čemer smo z izluščenimi prevodi sproti dopolnjevali tudi izhodiščni leksikon za prevajanje kontekstnih vektorjev. Rezultat raziskave je prvi prosto dostopen računalniško berljivi hrvaško-slovenski leksikon, ki bo v kratkem objavljen na spletu, za zdaj pa ga je mogoče dobiti pri avtorjih. Predstavljen pristop je uporaben tudi za številne druge pare sorodnih jezikov, za katere še ni na voljo primernih dvojezičnih leksikalnih virov. V prihodnje ga nameravamo razširiti še na večbesedne zveze, ki so pomemben sestavni del računalniške obdelave naravnega jezika. Poleg tega se želimo z izboljšavami prevajanja kontekstnih vektorjev in merjenja podobnosti med njimi na eno- in večjezični ravni podrobneje posvetiti tudi večpomenskosti.

Viri

AL-ONAIZAN, Yaser, KNIGHT, Kevin, 2002: Translating Named Entities Using Monolingual and Bilingual Resources. *ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics. 400–408.

FIŠER, Darja idr., 2011: Building and using comparable corporafor domain-specific bilingual lexicon extraction. *4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web: Proceedings of the Workshop*. Stroudsburg: The Association for Computational Linguistics.

FUNG, Pascale, 1998: A statistical view on bilingual lexicon extraction: From parallel corpora to nonparallel corpora. *AMTA '98 Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*. London: Springer-Verlag. 1–17.

JURANČIČ, Janko, 1986: *Srbskohrvatsko-slovenski slovar.* Ljubljana: DZS.

KOEHN, Philipp, KNIGHT, Kevin, 2002: Learning a translation lexicon from monolingual corpora. *ULA '02 Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition. Volume 9*. Stroudsburg: Association for Computational Linguistics. 9–16.

KONDRAK, Grzegorz, DORR, Bonnie, 2004: Identification of Confusable Drug Names: A New Approach and Evaluation Methodology. *COLING '04 Proceedings of the 20th international conference on Computational Linguistics*. Stroudsburg: Association for Computational Linguistics.

LJUBEŠIĆ, Nikola, ERJAVEC, Tomaž, 2011: Compiling web corpora for Croatian and Slovene. *Zbornik konference SlaviCorp*.

MARKÓ, Kornél, SCHULZ, Stefan, HAHN, Udo, 2005: Multilingual Lexical Acquisition by Bootstrapping Cognate Seed Lexicons. *Proceedings of the 5th International Conference on Recent Advances in Natural Language Processing*. Borovets. 301–307.

OCH, Franz Josef, NEY, Hermann, 2000: Improved Statistical Alignment Models. *ACL '00 Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics. 440–447.

RAPP, Reinhard, 1999: Automatic identification of word translations from unrelated English and German corpora. *ACL '99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Stroudsburg: Association for Computational Linguistics. 519–526.

SARALEGI, Xabier, SAN VICENTE, Iñaki, GURRUTXAGA, Antton, 2008: Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. *Proceedings of the 1st Workshop on Building and Using Comparable Corpora*.

SCANNELL, Kevin P., 2007: *Language similarity table*. http://borel.slu.edu/crubadan/table.html

SHAO, Li, NG, Hwee Tou, 2004: Mining New Word Translations from Comparable Corpora. *COLING '04 Proceedings of the 20th international conference on Computational Linguistics*. Stroudsburg: Association for Computational Linguistics.

SHEZAF, Daphna, RAPPOPORT, Ari, 2010: Bilingual Lexicon Generation Using Non-Aligned Signatures. *ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics. 98–107.

VORHEES, Ellen M., 1999: The TREC-8 Question Answering Track Report. *Proceedings of the 8th Text Retrieval Conference*. 77–82.

XIAO, Richard, MCENERY, Tony, 2006: Collocation, semantic prosody and near synonymy: a cross-linguistic perspective. *Applied Linguistics* 27/1. 103–129.