Razdvoumljanje besednega pomena pri strojnih prevajalnikih Amebis Presis, Google Translate in MT@EC

Jure Škerl

Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
jureskerl@gmail.com

Povzetek

V članku je predstavljena raziskava, katere namen je bil izmeriti natančnost razdvoumljanja besednega pomena za tri strojne prevajalnike: Amebis Presis, Google Translate in MT@EC. Na voljo so različni specializirani algoritmi za razdvoumljanje besednega pomena, vendar je njihova implementacija v prakso redka. Od treh preizkušanih sistemov ima le Amebis Presis vgrajene eksplicitne mehanizme za razdvoumljanje, medtem ko se ostala dva zanašata na statistični frazni model, pri katerem je razdvoumljanje posameznih besed posledica statističnega prevajanja večjih kosov besedil naenkrat. Primerjava natančnosti razdvoumljanja med prvim in drugim pristopom je dala nekaj uporabnih uvidov v prednosti in slabosti obeh. Druga raven primerjave je bila med rezultati na različnih jezikovnih področjih: književnost, publicistika, spletni dnevniki.

Word Sense Disambiguation in MT Systems: Amebis Presis, Google Translate, MT@EC

The article presents a research in which accuracy of word sense disambiguation was measured for three MT systems: Amebis Presis, Google Translate and MT@EC. There are various specialised word sense disambiguation algorithms available, however few of them are actually implemented into practical systems. Out of the three tested MT systems, only Amebis Presis uses explicit disambiguation techniques, whereas the other two rely on the statistical phrase-based model, in which disambiguation of words is the result of statistically translating longer text segments. The comparison of disambiguation accuracy between the first and the second approach provided some useful insights into advantages and disadvantages of both. Another level of comparison was done by testing disambiguation on texts from different language domains: literary, journalistic and blogs.

1 Uvod

Razdvoumljanje besednega pomena (RBP) je eden najstarejših in tudi najtežje rešljivih problemov v okviru strojnega prevajanja. Poleg tega se z njim v naravnih besedilih srečujemo pogosteje, kot je očitno na prvi pogled. Za 121 najpogostejših angleških samostalnikov, denimo, ki predstavljajo približno petino vseh besed v poljubnem besedilu, angleški WordNet v povprečju našteje 7,8 različnih pomenov (Agirre in Edmonds, 2007). Če k temu dodamo še 70 najpogostejših glagolov, ki imajo v WordNetu povprečno po dvanajst pomenov (Palmer et al., 2007), lahko zaključimo, da bomo primere večpomenskosti srečali v slehernem naravnem besedilu.

To je pomembno poudariti, ker povprečen bralec večine večpomenskih besed niti ne zazna; človeški možgani razdvoumljanje opravljajo tako rekoč trenutno in večinoma povsem podzavestno. To počnejo na osnovi vsega znanja o svetu, ki so ga nabrali tekom svojega obstoja, načini, na katerega je to znanje shranjeno, pa so abstraktni in nikakor ne na voljo v oblikah, ki bi bile dostopne računalniškim sistemom. Računalnik se zato lahko spotakne že ob tako osnovne primere večpomenskosti, kot je tista v stavku »Ana je jabolko.« Kako lahko strojni prevajalnik ve, ali gre v tem stavku za glagol »jesti« ali glagol »biti«? Brez znanja o svetu, ki človeškemu bralcu omogoča brez pomislekov odgovoriti na to vprašanje, je vse, kar mu preostane, kontekst, se pravi okoliške besede.

Načini, na katere lahko slednjega analizira in ga uporabi kot osnovo za odločanje med pomeni, so raznoliki in segajo od uporabe pomenskih virov, kot so slovarji in tezavri, pa vse do statistične analize ogromnih količin vzporednih besedil v dvojezičnih korpusih. Ta prispevek ni poskus predstavitve vseh možnih pristopov, temveč le poskus analize praktičnih rezultatov, ki jih pri razdvoumljanju dosegajo trije v naslovu omenjeni strojni prevajalniki.

Namen članka je podrobna predstavitev rezultatov razdvoumljanja treh strojnih prevajalnikov (Presis, Google Translate in MT@EC) pri smeri prevajanja iz slovenščine v angleščino. Eksperiment je bil izveden na več besedilih v skupnem obsegu 869 besed, ki so bila izbrana tako, da so bila žanrsko, tematsko in slogovno raznolika. Nabrana so bila s treh jezikovnih področij: knjižna, publicistična in spletna besedila. Natančnosti razdvoumljanja so bile izmerjene za vse skupaj in za vsak sklop posebej. To je omogočilo primerjavo ne samo med posameznimi strojnimi prevajalniki, temveč tudi med rezultati, ki jih vsak od njih dosega na različnih jezikovnih področjih. Od treh prevajalnikov ima le Presis vgrajene specializirane mehanizme za razdvoumljanje (Holozan, 2011), druga dva, ki sta statistična, pa ne. Zato bi rezultati morali razkriti tudi učinek prisotnosti oziroma odsotnosti takšnih mehanizmov za razdvoumljanje ter razlike med njimi in golim statističnim pristopom.

2 Pregled pristopov

Pristope k RBP lahko v grobem razdelimo na dva dela: korpusne in nekorpusne (Agirre in Edmonds, 2007). Prednost korpusnih pristopov v primerjavi z nekorpusnimi je višja natančnost razdvoumljanja, slabost pa, da znajo razdvoumljati le besede, ki se znajdejo v uporabljenih korpusih. Nekorpusne metode, ki se naslanjajo na slovarje in pomenske vire, kot je WordNet, po drugi strani omogočajo razdvoumljanje precej širšega besedišča.

Med nekorpusne pristope uvrščamo denimo algoritme, ki računajo *semantično sorodnost*: iz predpostavke, da naravna besedila težijo h koherentnosti, lahko sklepamo, da bodo pomeni besed iz enega besedila med seboj bolj ali

manj sorodni. Na podlagi te ugotovitve za pravilen pomen dvoumnih besed izbiramo tiste, ki izkazujejo najvišjo semantično sorodnost z okoliškimi pomeni (Mihalcea, 2007). Enega prvih postopkov, delujočih na tem principu, je razvil Philip Resnik (1995). Za začetek je vpeljal pojem specifičnosti koncepta, kar je definiral kot verjetnost, da se nek koncept pojavi v relativno obsežnem korpusu. Resnik nato definira semantično bližino med dvema besedama tako, da kvantificira razdaljo do najnižjega skupnega vozlišča, do katerega pridemo, če potujemo od obeh besed po hierarhiji navzgor.

Njegovo enačbo sta nekoliko prilagodila Jiang in Conrath (1997), in sicer tako, da sorodnost med dvema konceptoma merita z razliko v informacijski vsebini (ang. information content). Hirst in St-Onge (1998) pa sta v svojo verzijo enačbe za izračun semantične sorodnosti integrirala še smer, v katero tečejo povezave. Ta pristop temelji na ideji, da je semantična sorodnost med dvema konceptoma tem višja, čim manjkrat povezava med njima spremeni smer.

Druga skupina znotraj nekorpusnih pristopov k RBP so *hevristične metode*, ki izrabljajo določene naravne zakonitosti jezikov. Njihova prednost je v preprostosti, zaradi katere ne zahtevajo veliko računske moči, slabost pa na splošno nižja natančnost razdvoumljanja, čeprav so lahko v specifičnih primerih zelo uspešne. Najosnovnejši hevristični postopek RBP je izbiranje najpogostejšega pomena. Naravna zakonitost, ki se jo tu izrablja, je dejstvo, da distribucije besednih pomenov sledijo Zipfovemu zakonu (1949), tj. statistični distribuciji, v kateri je ena kategorija dominantna, frekvence vseh ostalih pa so bistveno nižje. Na osnovi tega lahko izdelamo zelo preprost in robusten sistem za razdvoumljanje, ki vsaki dvoumni besedi pripiše pomen, ki je v naravnih besedilih najpogostejši. Še en primer hevristične metode je izbiranje enega pomena na diskurz, ki so ga vpeljali Gale in sodelavci (1992b) in temelji na predpostavki, da večpomenska beseda ohrani isti pomen skozi celotno besedilo, v katerem se pojavlja. To pomeni, da je razdvoumljanje večjega števila njenih pojavitev trivialna naloga, če pravilno identificiramo njen pomen v najmanj eni od njih.

Druga obsežna skupina pristopov k RBP temelji na izrabi korpusov. Vanjo spadajo številne metode, tu pa se bomo na kratko dotaknili le ene: RBP na osnovi *prevodne vzporednosti*. Ta metoda je sposobna razlikovati med pomeni izključno glede na prevodne ustreznice, ki so zelo uporabna indikacija za razdvoumljanje. Pogoj za njeno delovanje je dvojezični korpus, ki mora izpolnjevati eno poglavitno zahtevo, in sicer to, da so besedila v obeh jezikih poravnana. To pomeni, da ima vsaka fraza ali vsaka beseda v izvirnem besedilu svojo ustreznico v ciljnem besedilu. Do tega pridemo tako, da najprej poravnamo stavke, nato pa še posamezne besede/fraze. Ves postopek se izvaja samodejno, saj bi bila ročna poravnava preveč zamudna. Samo razdvoumljanje nato poteka s pomočjo podatkov o besedišču in skladnji v okolici razdvoumljane besede in preverjanjem, kako je bila ta beseda prevedena v podobnih kontekstih, najdenih v vzporednem korpusu. Eni prvih, ki so preizkušali to metodo, so bili Gale in sodelavci (1992a), kasneje pa tudi Chklovski in sodelavci (2004).

Tako za omenjene kot druge tehnike RBP sicer velja, da so bile v glavnem razvite in preizkušane zgolj v raziskovalnem okolju, njihova implementacija v praksi pa je redka. V nadaljevanju predstavljen eksperiment je tako predvsem poskus splošne primerjave natančnosti razdvoumljanja med statističnimi in na pravilih temelječimi prevajalniki.

3 Metodologija

3.1. Predstavitev prevajalnikov

Preizkušani so bili trije prevajalniki, dva statistična (Google Translate in MT@EC) in eden, ki temelji na pravilih (Presis). Slednjega razvija slovensko podjetje Amebis in je specializiran za jezikovna para slovenščina-angleščina ter slovenščina-nemščina (pri tem le v smeri iz nemščine v slovenščino). Pri stavčni analizi se opira na ročno izdelano podatkovno zbirko Ases, ki vsebuje podatke o besedah, besednih zvezah, skupinah, predlogih in pomenih (Holozan, 2011), za reprezentacijo pomena izhodiščnega besedila pa uporablja vmesni jezik (interlingua).

MT@EC je interni statistični prevajalnik prevajalske službe EU, ki je bil zgrajen na odprtokodnem sistemu Moses. Prevajati zna med vsemi kombinacijami 24 uradnih jezikov EU, pri tem pa uporablja korpuse v skupnem obsegu 1,65 milijarde besed. Sklepamo lahko, da je močno specializiran za zakonodajna in tehnična besedila, ki se tičejo delovanja EU, kar se je deloma potrdilo tudi v eksperimentu.

Zadnji izmed treh preizkušanih prevajalnikov je Google Translate, ki je ravno tako kot MT@EC statističen, vendar temelji na lastniški programski opremi. Z možnostjo prevajanja iz in v 103 jezike sveta je trenutno najbolj obsežen javno dostopen strojni prevajalnik, katerega storitve dnevno uporablja prek 200 milijonov uporabnikov. Besedila za svoje čedalje obsežnejše korpuse podobno kot MT@EC črpa iz dokumentov EU, poleg tega pa še iz uradnih besedil Združenih narodov, ki so praviloma objavljana v šestih uradnih jezikih te organizacije, ter drugih eno- in dvojezičnih spletišč.

3.2. Predobdelava

Preizkušanje razdvoumljanja je potekalo na odlomkih besedil v skupnem obsegu 869 besed, ki so bila izbrana tako, da so pokrila tri različna jezikovna področja – knjižni, publicistični in spletni jezik. Za knjižna besedila so bili odlomki vzeti iz Cankarjeve povesti Krčmar Elija ter iz Mansarde Slavka Gruma. Odlomki publicističnih besedil so bili nabrani z novičarske spletne strani MMC[1], in sicer iz dveh različnih člankov z notranje- in zunanjepolitično tematiko. Zadnja tretjina besedil vključuje objave v manj formalnem jeziku iz dveh spletnih dnevnikov[2]. Cilj takšnega izbora je bil pokriti nekoliko širšo sliko jezikovne realnosti, navkljub relativno majhnemu obsegu testnih besedil.

Na besedilih je bila najprej opravljena ročna semantična analiza pomenov z upoštevanjem njihovega vpliva na izbor prevodnih ustreznic v angleščini. Subjektivnost takšne ročne analize je bila omiljena z doslednim naslanjanjem na več slovarskih ter korpusnih virov, ki so dostopni na spletu in v knjižni obliki. Glavna

[1] www.rtvslo.si
[2] http://tomazjakofcic.com/blog in http://heliopolis.si

uporabljena orodja v tej fazi so bila pregibnik Amebis Besana 4.12, semantični leksikon sloWNet 3.1 in spletni zbirki slovarjev fran.si ter thefreedictionary.com. V prvem koraku semantične analize je bila vsaka beseda vnešena v Amebisov javno dostopni sistem Besana, ki pozna vse pregibne oblike leksemov in zna za vsako poiskati vse izvirne leme. Na ta način je bilo mogoče odkriti vse obstoječe leme, ki se pregibajo v posamezno morfološko obliko. Denimo, v stavku *naročil je še en bokal* gre lahko, če ne upoštevamo skladnje, pri besedi *naročil* za samostalnik v rodilniku dvojine/množine ali pa za pretekli deležnik. V tem primeru gre torej za dvoumnost, ki je posledica dejstva, da dva leksema vsebujeta isto pregibno obliko. Po tej analizi je za vsako besedo nastal seznam vseh možnih lem, iz katerega je bilo nato mogoče izpeljati vse možne pomenske interpretacije. Določanje potencialnih pomenov je bilo izvedeno z vnašanjem lem v slovarski iskalnik fran.si in analizo vrnjenih slovarskih vnosov.

Glede na to, da je predmet raziskave razdvoumljanje za potrebe strojnega *prevajanja*, se je v zadnjem, ključnem koraku semantične analize vse dobljene potencialne pomene navzkrižno preverilo z leksikalno podatkovno bazo sloWNet 3.1 (Fišer in Sagot, 2015) ter angleškim spletnim slovarjem thefreedictionary.com. V tem koraku so bili vsi pomeni, ki imajo v angleškem jeziku isto prevodno ustreznico, združeni v eno enoto, saj je cilj uspešnega razdvoumljanja v strojnem prevajanju dejansko najti "pravilen" prevod in ne nujno "pravilen" pomen.

Končni rezultat takšne obdelave je bil seznam vseh besed v besedilih s pripisanimi števili možnih prevodov (in ne pomenov) ter seznamom le-teh za vsako besedo posebej. Vse to je bilo vnešeno v tabelo, tako da je bilo možno v naslednjem koraku vzporediti besede, njihove možne prevode ter dejanske prevode, ki so jih izbrali preizkušani strojni prevajalniki. Pred tem pa je bilo potrebno seznam še prečistiti, in sicer so bile iz njega izločene naslednje kategorije, ki so bile bodisi nerelevantne bodisi bi povzročile izkrivljene rezultate:

- vse besede, ki imajo v ciljnem jeziku le eno prevodno ustreznico
- vsi osebni zaimki
- vse pojavitve glagola biti, v katerih le-ta nastopa kot pomožni glagol
- vse pojavitve prislova "pa" (če je nastopal kot veznik, je bil ohranjen)

Poleg tega so bili nekateri večdelni vezniki in redke idiomatične fraze združeni v eno pomensko enoto, kjer je bilo to smiselno, in sicer tako, da je bila ohranjena le ključna dvoumna beseda v frazi oziroma večdelnem vezniku. Ta se je štela za pravilno razdvoumljeno, če je prevajalnik uspel pravilno razbrati skupen pomen. Po tej fazi je na seznamu iz knjižnih besedil ostalo 106 večpomenskih enot, na seznamu iz publicističnih besedil 173 in na seznamu iz spletnih dnevnikov 116 (skupno 395). Tako prečiščen seznam predstavlja celoto podatkov, na podlagi katere je bilo moč v nadaljevanju eksperimenta izračunati natančnost razdvoumljanja za vsak preizkušani sistem posebej.

3.3. Kriteriji

Ocene natančnosti razdvoumljanja so bile izračunane ločeno za vsak prevajalnik in za vsako jezikovno področje posebej (knjižna, publicistična in spletna besedila) ter na koncu še za vsa besedila skupaj. Ocena uspešnosti pomenskega razdvoumljanja je razdeljena v štiri kategorije:

(1) Pravilno razdvoumljen pomen
(2) Napačno razdvoumljen pomen
(3) Pravilno razdvoumljen pomen s prevodno napako
(4) Neprevedeno

Pod kategorijo (1) so se uvrstili vsi primeri, v katerih je strojni prevajalnik razdvoumljanje večpomenske besede izvedel pravilno, tj. kjer je našel ustrezen prevod v ciljnem jeziku in ga tudi slovnično pravilno vtkal v prevedeno besedilo.

V kategoriji (2) so združeni vsi čisti primeri napačnega razdvoumljanja, tj. primeri, kjer je prevajalnik izbral očitno pomensko napačen prevod.

Kategorija (3) je zajela primere, kjer so prevajalniki pomensko razdvoumljanje sicer opravili pravilno, vendar je pri tem prišlo do neke prevodne napake. Tipični primeri v tej kategoriji so denimo odločitev za prislov, kjer bi moral stati pridevnik, ter obratno, pa neujemanje pravilno razdvoumljane besede z izvirnikom v sklonu, številu ali osebi, izbrano deležje namesto starinske oblike glagola v 3. os. mn. itd.

Zadnja kategorija (4) zajema vse primere, v katerih je prevajalnik odpovedal in dvoumno besedo pustil neprevedeno, bodisi z izpustitvijo bodisi tako, da jo je pustil v izvirni, slovenski obliki.

Primeri, v katerih je bil pravilno razdvoumljen pomen zajet v okoliških besedah, izvirna beseda pa ni bila neposredno prevedena, so bili uvrščeni v kategorijo (1). Ravno tako primeri, ko je bila pravilno razdvoumljena beseda umeščena na napačno mesto v stavku, tako da je bil prevod napačen s skladenjskega vidika.

4 Rezultati

4.1. Razčlemba

4.1.1. Knjižna besedila

Pri knjižnih besedilih se je najbolje odrezal Presis, ki je dosegel 70,7-odstotno natančnost razdvoumljanja, v kar se štejejo primeri iz kategorije 1 (67,9 odstotka) in primeri, v katerih je bilo razdvoumljanje uspešno, vendar je prišlo do neke druge prevodne napake (kategorija 3: 2,8 odstotka). Nasprotno so bili rezultati, ki jih je na knjižnih besedilih dosegel MT@EC, z naskokom najslabši, saj je dosegel le 49,1-odstotno natančnost (kategorija 1: 43,4 odstotka, kategorija 3: 5,7 odstotka). To je hkrati tudi najnižja posamezna vrednost v celotnem eksperimentu, se pravi izmed vseh kombinacij treh besedilnih žanrov in treh prevajalnikov. S tem se delno potrjuje domneva, da je MT@EC tematsko najožje specializiran prevajalnik, kar ni presenetljivo glede na to, da so ga razvili za potrebe prevajalskih služb EU. Za ilustracijo tega dejstva lepo služi primer napačnega razdvoumljanja glagola *tožiti* v pomenu *pritoževati se*, ki ga je MT@EC prevedel z glagolom *to sue*, kar bi bilo v kontekstu uradnih evropskih

dokumentov skoraj zagotovo pravilno razdvoumljanje, v našem, knjižnem primeru pa ne. V tem primeru je bil Presis edini, ki je razdvoumljanje opravil uspešno in izbral glagol *to moan*. Kar se Googlovega prevajalnika tiče, je v kategoriji knjižnih besedil dosegel 68,8-odstotno natančnost razdvoumljanja (66,0 odstotka v kategoriji 1 in 2,8 odstotka v kategoriji 3), torej le nekoliko manj od Presisa.

4.1.2 Publicistična besedila

Najvišjo natančnost razdvoumljanja publicističnih besedil je dosegel Google Translate, in sicer 83,8 odstotka (kategorija 1: 80,3 odstotka, kategorija 3: 3,5 odstotka). MT@EC se je z 80,9-odstotno natančnostjo uvrstil tesno za njim (kategorija 1: 79,2 odstotka, kategorija 3: 1,7 odstotka). Nekoliko bolj je zaostal Presis z 72,2-odstotno natančnostjo (kategorija 1: 69,9 odstotka, kategorija 3: 2,3 odstotka). Zanimivo si je pogledati prevajanje večpomenske besede *stran*, ki se je v tem delu pojavila trikrat. Prav tolikokrat je njeno razdvoumljanje spodletelo Presisu, ki je dvakrat izbral prevod *page* in enkrat prevod *direction*. Nasprotno je Google Translate vse tri pojavitve besede razdvoumil pravilno, dvakrat s prevodom *side* in enkrat z idiomatično frazo *on the other hand*. MT@EC se je uvrstil med njiju z enim spodletelim poskusom, ko je *na drugi strani morja* prevedel z *on the other hand, the sea*. Ti rezultati namigujejo na dober potencial za razdvoumljanje, ki ga imajo statistični prevajalniki. Vendar pa se, če upoštevamo vse rezultate skupaj, razkrije tudi njihova slabost, namreč dejstvo, da natančno razdvoumljajo le besedila s tistih tematskih področij, na katerih so bili trenirani (v tem primeru so to politične teme uradnih dokumentov EU, ki jih oba uporabljata za trenažne korpuse). To domnevo potrjuje tudi primerjava natančnosti razdvoumljanja po besedilnih kategorijah, kjer vidimo, da Presis kot predstavnik na pravilih temelječih prevajalnikov dosega najbolj konsistentne rezultate skozi vse kategorije, Googlovi in zlasti MT@EC-jevi pa od kategorije do kategorije občutno bolj nihajo (slika 1).

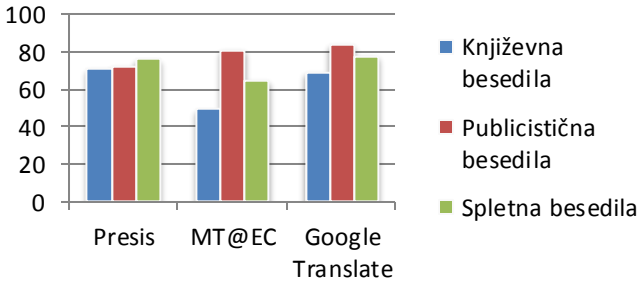
4.1.3 Spletna besedila

Tretjo kategorijo sestavljajo spletna besedila, ki so bila nabrana iz dveh slovenskih spletnih dnevnikov. V prvem je bil jezik bolj pogovoren, v drugem pa bolj knjižen; za oba je sicer značilno pisanje v prvi osebi. Najvišjo natančnost razdvoumljanja v tej kategoriji je dosegel Google Translate s 77,6 odstotki (kategorija 1: 75,9 odstotka, kategorija 3: 1,7 odstotka), sledil je Presis s 75,9 odstotki (kategorija 1: 73,3 odstotka, kategorija 3: 2,6 odstotka), precej zadaj se je uvrstil MT@EC s 64,6 odstotka (kategorija 1: 61,2 odstotka, kategorija 3: 3,4 odstotka). Slabši rezultat slednjega znova namiguje, da gre za prevajalnik, ki je ozko specializiran za potrebe prevajalskih služb EU. Za to se najde tudi konkreten primer: besedna zveza *enoglasna pritrditev*, ki je sicer povzročila težave vsem trem prevajalnikom. Nobeden je ni v celoti razdvoumil pravilno, vendar pa je bil MT@EC temu še najbližje. Medtem ko je besedo *pritrditev* preprosto prezrl, je za prvo besedo pravilno izbral pomen *unanimous*. Tu lahko sklepamo, da je MT@EC pravilen prevod našel v korpusu dokumentov EU, kjer se lema *enoglasen* verjetno res najpogosteje pojavlja v povezavi s prevodom *unanimous*. Presis je celo frazo prevedel z *enoglasna consent*, torej je pravilno razdvoumil del, v katerem je MT@EC odpovedal. Najzanimivejšo rešitev pa je ponudil Google Translate, ki je frazo (povsem napačno) razdvoumil v *monophonic mounting*.

4.1.4 Skupni rezultati

Po skupnem rezultatu se je najvišje uvrstil Google Translate (tabela 3) z 78-odstotno natančnostjo razdvoumljanja, sledi mu Presis (tabela 1) z 72,9-odstotno natančnostjo, temu pa MT@EC (tabela 2) s 67,6-odstotno natančnostjo. Že zgoraj je bilo na kratko omenjeno, da Presis izkazuje največjo konsistentnost skozi vse kategorije, kar je najbrž posledica dejstva, da temelji na pravilih in je kot tak manj občutljiv na to, s katerih tematskih področij so razdvoumljana besedila – za razliko od drugih dveh, ki temeljita na statističnem modelu in posledično prej odpovesta na besedilih s tematskih področij, na katerih ju niso trenirali.

Po drugi strani lahko glede na to, da je MT@EC precej bolj nekonsistenten kot Google, sklepamo, da je mogoče to pomankljivost omiliti s preprosto strategijo gole sile, z drugimi besedami s tem, da statistični model hranimo s karseda velikimi količinami vzporednih in enojezičnih korpusov, saj Google tu prednjači pred MT@EC.



Slika 1: Konsistentnost Presisa v primerjavi z drugima dvema prevajalnikoma.

Presis	KNJIŽNA BESEDILA	PUBLICISTIČNA BESEDILA	SPLETNA BESEDILA	SKUPAJ
Pravilno	67,9 %	69,9 %	73,3 %	70,4 %
Napačno	23,6 %	24,3 %	19,8 %	22,8 %
Pravilno s prevodno napako	2,8 %	2,3 %	2,6 %	2,5 %
Neprevedeno	5,7 %	3,5 %	4,3 %	4,3 %

Tabela 1: Rezultati za Presis.

MT@EC	KNJIŽNA BESEDILA	PUBLICISTIČNA BESEDILA	SPLETNA BESEDILA	SKUPAJ
Pravilno	43,4 %	79,2 %	61,2 %	64,3 %
Napačno	28,3 %	12,2 %	21,6 %	19,2 %
Pravilno s prevodno napako	5,7 %	1,7 %	3,4 %	3,3 %
Neprevedeno	22,6 %	6,9 %	13,8 %	13,2 %

Tabela 2: Rezultati za MT@EC.

Google Translate	KNJIŽNA BESEDILA	PUBLICISTIČNA BESEDILA	SPLETNA BESEDILA	SKUPAJ
Pravilno	66,0 %	80,3 %	75,9 %	75,2 %
Napačno	22,7 %	11,6 %	14,6 %	15,4 %
Pravilno s prevodno napako	2,8 %	3,5 %	1,7 %	2,8 %
Neprevedeno	8,5 %	4,6 %	7,8 %	6,6 %

Tabela 3: Rezultati za Google Translate.

5 Zaključek

Cilj eksperimenta je bil v praksi preveriti stanje tehnologije razdvoumljanja pri treh različnih strojnih prevajalnikih. Ob pregledu tega področja se izkaže predvsem, da je kljub obstoju številnih različnih strategij za razdvoumljanje besednega pomena njihova praktična implementacija redka. Le malokateri trenutno aktualen strojni prevajalnik pri izdelavi prevodov uporablja kakšno od obstoječih specializiranih tehnologij razdvoumljanja besednega pomena. To se zdi posledica dejstva, da so trenutno glavni trendi razvoja usmerjeni v statistične pristope s fraznimi modeli strojnega prevajanja, ki kot osnovne pomenske enote ne analizirajo posameznih besed, temveč daljše kose besedil. V tem primeru se razdvoumljanje v bistvu zgodi kot naravna posledica iskanja prevodno ustreznih daljših kontekstov v ciljnem jeziku. Z drugimi besedami, dovolj dolgi segmenti besedil se v fraznem modelu strojnega prevajanja razdvoumijo kar sami od sebe.

Takšna taktika zagotavlja sorazmerno visoko natančnost razdvoumljanja pri večini besedil, s katerimi se srečujemo pri strojnem prevajanju, vendar pa hitro odpove pri razdvoumljanju manj pogostih pomenov. V tovrstnih primerih se uporabnost različnih algoritmov za razdvoumljanje besednega pomena izkaže za nesporno. To se je potrdilo tudi v eksperimentu, ki je zajel Google Translate in MT@EC kot predstavnika statističnih pristopov ter Presis kot predstavnika na pravilih temelječih pristopov, ki ima tudi edini od trojice vgrajene eksplicitne postopke za razdvoumljanje (Holozan, 2011). Temu najbrž lahko pripišemo vsaj del zaslug za Presisovo visoko natančnost razdvoumljanja na književnih besedilih, v katerih se praviloma večkrat kot pri drugih besedilnih žanrih srečamo z redkejšimi besednimi pomeni. Druga očitna posledica pa je višja konsistentnost razdvoumljanja skozi različne besedilne žanre, ki jo izkazuje Presis v primerjavi s statističnima prevajalnikoma.

Za prihodnji razvoj in doseganje višjih natančnosti razdvoumljanja besednega pomena bi bilo morda smiselno preizkušati kombiniranje statističnih pristopov z opisanimi specializiranimi algoritmi za razdvoumljanje. Modularno dodajanje teh algoritmov že obstoječim statističnim strojnim prevajalnikom bi verjetno povzročilo povišanje natančnosti razdvoumljanja, vendar je to nekaj, kar bi zaradi specifičnosti delovanja statističnih prevajalnikov najbrž zahtevalo veliko dela pri implementaciji. Vseeno pa bi raziskovanje v to smer morda obrodilo koristne rezultate.

6 Zahvala

Zahvaljujem se prof. dr. Špeli Vintar za nasvete pri pripravi na izvedbo eksperimenta in pisanju članka, mag. Petru Holozanu za dostop do Presisa in Zoranu Zakiću za dostop do MT@EC.

7 Literatura

Eneko Agirre in Philip Edmonds. 2007. *Word Sense Disambiguation*. Springer, Dordrecht.

Amebis Besana – Pregibanje. http://besana.amebis.si/pregibanje/

Tim Chklovski, Rada Mihalcea, Ted Pedersen in Amruta Purandare. 2004. The Senseval-3 multilingual English-Hindi lexical sample task. V: *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, str. 5-8. Barcelona, Španija.

Dictionary, Encyclopedia and Thesauraus – The Free Dictionary. www.thefreedictionary.com

Darja Fišer in Benoît Sagot. 2015. Constructing a poor man's wordnet in a resource-rich world. *Language Resources and Evaluation 49*: 601–35.

Fran. www.fran.si

William Gale, Ken Church in David Yarowsky. 1992a. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. V: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, str. 249–56. Newark, ZDA.

William Gale, Ken Church in David Yarowsky. 1992b. One sense per discourse. V: *Proceedings of the DARPA Speech and Natural Language Workshop*, str. 233–37. New York, ZDA.

Graeme Hirst in David St-Onge. 1998. Lexical chains as representations of context in the detection and correction of malaproprisms. V: *WordNet: An electronic lexical database*, str. 305–32. MIT Press, Massachusetts, ZDA.

Peter Holozan. 2011. *Samodejno izdelovanje besedilnih logičnih nalog v slovenščini.* Magistrsko delo, Univerza v Ljubljani.

Jian Jiang in David Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. V: *Proceedings of the International Conference on Research in Computational Linguistics*. Taipei, Tajvan.

Leposlovje. http://lit.ijs.si/leposl.html

Philip Resnik. 1995. Using information content to evaluate semantic similarity. V: *Proceedings of the International Joint Conference on Artificial Intelligence*, str. 448–53. Montreal, Kanada.

sloWTool. http://nl.ijs.si/slowtool/

George Kingsley Zipf. 1949. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley.