O AVTOMATSKI EVALVACIJI STROJNEGA PREVAJANJA

Darinka VERDONIK
Mirjam SEPESY MAUČEC
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko

Verdonik, D., Sepesy Maučec, M. (2013): O avtomatski evalvaciji strojnega prevajanja. Slovenščina 2.0, 1 (1): 111–133.
URL: http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_06.pdf.

Stalen del razvoja strojnega prevajanja je evalvacija prevodov, pri čemer se v glavnem uporabljajo avtomatski postopki. Ti vedno temeljijo na referenčnem prevodu. V tem prispevku pokažemo, kako zelo različni so lahko referenčni prevodi za področje podnaslavljanja ter kako lahko to vpliva na oceno – ista metrika lahko isti prevajalnik oceni kot neuporaben ali kot zelo uspešen samo na podlagi tega, da uporabimo referenčne prevode, ki so pridobljeni po različnih postopkih, vendar vedno jezikovno in pomensko povsem ustrezni.

Ključne besede: strojno prevajanje, evalvacija, referenčni prevod, BLEU, TER

1 UVOD

Začetkom strojnega prevajanja lahko sledimo nazaj vse do časa po drugi svetovni vojni ali še dlje. Danes lahko delimo pristope k strojnemu prevajanju v grobem na prevajanje s pravili in statistično prevajanje. Prevajanje s pravili temelji na množici jezikovnih pravil, katerih definiranje je dolgotrajno in zahtevno opravilo, vendar je po drugi strani tudi izdelava ustrezno velikega vzporednega korpusa, ki je osnova za statistično prevajanje, lahko precej dolgotrajna. Prevajanje s pravili je med drugim osnova edinega slovenskega komercialnega prevajalnika Presis,[1] ki pokriva slovenščino v paru z nemščino in angleščino. Na osnovi statističnega prevajanja delujeta na primer Google

[1] http://presis.amebis.si

Translate[2] in Bing,[3] ki med številnimi jezikovnimi pari pokrivata tudi slovenščino. Od večjih svetovnih prevajalnikov omenimo še Systran,[4] ki ima že 40-letno zgodovino. Sprva je temeljil samo na pravilih, zadnji čas pa vključuje tudi statistične metode in je tako postal hibridni sistem. Kot orodje, ki omogoča hibridni pristop k strojnemu prevajanju, čeprav v osnovi izhaja iz statističnih pristopov, se označuje tudi Moses.[5] V slovenskem prostoru se s področjem strojnega prevajanja ukvarjajo seveda tudi v akademskem okolju, na primer Sepesy Maučec, Kačič (2006), s stališča terminologije tudi Vintar idr. (2012), akademski prevajalnik, ki temelji na pravilih plitkega prenosa in deluje za slovensko-srbski par, je bil predstavljen v okviru doktorata v Vičič (2012), z evalvacijo strojnega prevajanja pa se je ukvarjal Vrščaj (2011).

V prispevku se bomo ukvarjali s prevodi statističnega prevajanja, zato nekaj več o statističnih pristopih. Prvi sistemi statističnega prevajanja so temeljili na IBM-ovih modelih (Brown idr. 1993), ki so za osnovno prevodno enoto uporabljali besedo. Kasneje so skušali mnogi IBM-ove modele izboljšati ali nadgraditi (He 2007; Ma idr. 2007) ali pa izboljšati ujemanja besed z entropijskimi modeli (Ittycheriah, Roukos 2005) ali dodajanjem jezikoslovnih informacij (in sicer z lemo – Popović, Ney 2004; pregibnostnimi stopnjami – Niessen, Ney 2004; ali redukcijo morfoloških enot – Sepesy Maučec, Brest 2010). Sčasoma se je pokazalo, da beseda ni najbolj ustrezna prevodna enota, in tako je sledil razvoj v smeri statističnega prevajanja na podlagi fraze kot prevodne enote (angl. *phrase-based machine translation*) (Och, Ney 2004; Koehn 2010). Pri tem je kot fraza razumljeno katerokoli zaporedje pojavnic (vključno z ločili), ki se pojavlja v besedilih z določeno pogostostjo, in ne z jezikoslovnega stališča, torej kot besedna zveza, ki ima stalen pomen. Na takem pristopu temelji tudi orodje Moses (Koehn idr. 2007), s katerim smo tvorili prevode za naše eksperimente.

[2] http://translate.google.com
[3] http://www.microsofttranslator.com
[4] http://www.systran.co.uk
[5] http://www.statmt.org/moses/

Stalen del razvoja strojnega prevajanja je evalvacija prevodov, saj lahko le tako ocenimo, ali smo bili s svojim postopkom uspešni ali ne. Pri tem se v glavnem uporabljajo avtomatski postopki evalvacije, le ko zares pričakujemo, da smo dosegli bistveno izboljšavo, oz. ko želimo zares oceniti stanje določenega prevajalnika, se vključi tudi ročna evalvacija, saj je slednja stroškovno in časovno veliko zahtevnejša in jo lahko izvedemo samo s sodelovanjem strokovnjaka – prevajalca. Ne ena ne druga evalvacija pa ni povsem neproblematična.

Praviloma šteje ročna evalvacija prevodov za referenčno in cilj različnih pristopov k avtomatski evalvaciji je, da bi se čim bolj približali rezultatom, ki bi jih dala ročna evalvacija. Slednja običajno opazuje prevod vsaj z dveh stališč: kako razumljiv in tekoč je ter kako natančen je (koliko informacij izvornega teksta ohrani). Praktična opazovanja in merjenja ročnih evalvacij pa so pokazala, da je izrazito nestrinjanje o kvaliteti prevoda med različnimi prevajalci pogosto (o tem npr. poročajo Turian idr. 2003 ter Callison - Burch idr. 2007) oz. da je v določenih primerih lahko celo večja skladnost med rezultati avtomatske in ročne evalvacije kot med rezultati več različnih prevajalcev ter da je celo s strani istega prevajalca, ki je posamezen prevod ocenjeval dvakrat, zaznati odstopanja v rezultatih (pri čemer pa je praviloma neskladnost vendarle manjša kot med različnimi prevajalci) (Callison - Burch idr. 2007). Avtorji avtomatskih metrik za evalvacijo prevodov zato upravičeno opozarjajo, da je treba to dejstvo upoštevati tudi pri ocenjevanju uspešnosti avtomatske evalvacije, še vedno pa vsi sprejemajo dejstvo, da edino ročna evalvacija da pravo sliko o kvaliteti strojnega prevoda.

V nadaljevanju se bomo osredotočili predvsem na avtomatske postopke evalvacije strojnih prevodov. Obstajajo različne metrike za evalvacijo (opisane so v razdelku 2), ki pa vse temeljijo na primerjavi prevoda strojnega prevajalnika z referenčnim prevodom istega besedila s strani prevajalca. Nobena od poznanih evalvacijskih metrik ni idealna in mnogi avtorji se ukvarjajo z njihovim vrednotenjem (Turian idr. 2003; Callison - Burch idr. 2007; Callison - Burch idr. 2008; v slovenskem prostoru sta se do zdaj tega problema dotaknila predvsem Vičič (2010) in Vrščaj (2011)). V nasprotju z

večino tovrstnih preskusov, ki v glavnem evalvirajo uspešnost posameznih metrik, se tukaj ne bomo ukvarjali s primerjavo rezultatov avtomatske in ročne evalvacije, ampak bomo kot problematičnega izpostavili referenčni prevod prevajalca ter skušali pokazati, kako lahko rezultati avtomatske evalvacije strojnega prevoda variirajo glede na referenčni prevod. Pri tem izhajamo s stališča, da je isti stavek mogoče prevesti na različne načine. Podobne poskuse so predstavili Snover idr. (2006), tukaj pa jih bomo prenesli na področje podnapisov, kjer lahko pričakujemo, da bo ta problem še posebej izpostavljen. Pri prevajanju podnapisov so namreč prevajalci omejeni z dolžino prevoda, tj. podnapisa, zato se srečujejo predvsem z vprašanji, kaj izpustiti in kaj prevesti ter kako nekaj prevesti s čim manj besedami (o tem med drugim piše Kovačič (1996)). Pokazali bomo, kako različni pristopi k izdelavi evalvacijskega gradiva za področje podnapisov vplivajo na napovedovanje kvalitete strojnega prevoda s pomočjo avtomatske evalvacije.

2 POSTOPKI AVTOMATSKE EVALVACIJE

Osnovno vodilo avtomatske evalvacije je najti metriko, pri kateri bodo ocene prevodov čim bolj podobne ocenam, ki bi jih dali prevajalci, saj so nenazadnje ti najzahtevnejši končni uporabniki strojnega prevajanja. Banerjee (2005) je definiral pet lastnosti dobre metrike: skladnost, doslednost, občutljivost, zanesljivost in splošnost. Dobra metrika je skladna z ročnim ocenjevanjem. Je tudi dosledna, kar pomeni, da daje podobne rezultate za podobne prevajalne sisteme na istih testnih vzorcih. Občutljivost metrike pomeni, da odraža tudi manjše spremembe v zasnovi sistema. Metrika naj bi bila zanesljiva, torej lahko pričakujemo, da je sistem, ki mu metrika pripisuje visoke ocene, tudi zares dober. Splošnost metrike pomeni, da je uporabna za besedila različnih žanrov pisnega in govorjenega jezika. V nadaljevanju bomo opisali nekaj najbolj uveljavljenih metrik.

2.1 BLEU

Metrika BLEU[6] je bila prva, ki jo je odlikovala visoka korelacija z ročno evalvacijo. Danes velja za najpogosteje uporabljeno metriko. BLEU neodvisno oceni vsak posamezen segment, tj. poved oz. v našem primeru podnapis. Ocena segmenta je dejansko izračunana natančnost (ang. *precision*) na nivoju n-gramov. Ocena dokumenta oz. celotnega testnega vzorca je geometrijsko povprečje ocen posameznih segmentov. Ocena posameznega segmenta temelji na ujemanju n-gramov (n-gram je niz besed dolžine n) ocenjevanega segmenta in referenčnega prevoda. Privzeto se uporabljajo n-grami do reda štiri (do štiri zaporedne besede). Iskanje ujemanj je nekoliko prilagojeno v primerih, ko imamo na voljo več referenčnih prevodov. Ocena ujemanj posredno kaznuje predolge segmente, nima pa vgrajenega mehanizma za kaznovanje prekratkih prevodov. Kazen prekratkih prevodov temelji na dolžini celotnega dokumenta oz. vzorca in uporablja eksponentno funkcijo razmerja med dolžino avtomatskega prevoda in referenčnega dokumenta. V praktični uporabi se ne osredotočamo na ocene posameznih segmentov, ampak uporabljamo ocene metrike BLEU za celoten dokument oz. vzorec.

2.2 NIST

Metrika NIST[7] je izpeljana iz metrike BLEU. Medtem ko BLEU preprosto išče ujemanja n-gramov, NIST računa tudi, kako informativen je določen n-gram. To pomeni, da je najdeni n-gram, ki se redko pojavlja, bolj obtežen (ima pripisano večjo vrednost) kot tisti, ki se pojavlja bolj pogosto. Metrika NIST računa ujemanje za n-grame do reda pet. Tudi kazen prekratkih prevodov je modificirana, tako da manjša odstopanja v dolžini prevodov ne vplivajo na končno oceno.

[6] Angl. *bilingual evaluation understudy*.
[7] Kratica za ameriški *National Institute of Standards and Technology*.

2.3 WER, PER in TER

Metrika WER[8] je bila najprej uporabljena za evalvacijo sistemov razpoznave govora, a je uporabna tudi za ocenjevanje prevodov. Temelji na računanju Levenshteinove razdalje[9] med strojnim prevodom in referenčnim prevodom. Levenshteinova razdalja določa najmanjše število osnovnih operacij, potrebnih, da strojni prevod pretvorimo v referenčni prevod. Te osnovne operacije so: izbris, vstavitev in zamenjava. Variacija metrike WER je PER,[10] ki poleg osnovnih operacij dovoljuje tudi preurejanje besed in zaporedij besed, vendar tega ne šteje kot napako. Metrika TER[11] je enaka metriki PER, s to razliko, da preurejanje besed in zaporedij besed šteje za napako. Od metrik iz tega sklopa je najpogosteje uporabljen TER.

2.4 Ostale metrike

V zadnjem času se je uveljavila tudi metrika METEOR.[12] Temelji na štetju besed, ki se ujemajo, besed, ki se ujemajo v korenu besede, in besed, ki so sinonimne. Dodatno kaznuje neujemanje vrstnega reda besed. Metrika POSBleu računa BLEU na osnovi POS-oznak. Metrika TERp je izpeljanka metrike TER, ki dodatno upošteva parafraziranje, krnjenje in sinonime. Krnjenje pri metrikah METEOR in TERp temelji na Porterjevem algoritmu (ta besedo z avtomatskim postopkom razdeli na osnovo in »končnico«). Metriki METEOR in TERp črpata informacije o sinonimih iz semantičnega leksikona (angl. *WordNet*). Kot manj uspešne so se pokazale metrike, ki temeljijo na primerjavi odvisnostnih drevesnic.

[8] Angl. *word error rate*.
[9] http://en.wikipedia.org/wiki/Levenshtein_distance
[10] Angl. *position-independent word error rate*.
[11] Angl. *translation error rate*.
[12] Angl. *metric for evaluation of translation with explicit ordering*.

2.5 Metrike na podlagi popravljenih strojnih prevodov (angl. *human-targeted metrics*)

Snover idr. (2006) so predstavili eksperimente, v katerih so zgoraj navedene metrike izračunavali na podlagi novega referenčnega prevoda, ki so ga dobili tako, da so strojne prevode popravili po načelu, da so vnesli minimalno število potrebnih popravkov, da je bil prevod tekoč (jezikovno pravilen). Metrike so poimenovali »human-targeted«: HTER, HBLEU, HMETEOR … Tako izračunane rezultate so primerjali z rezultati istih metrik na podlagi referenčnih prevodov iz evalvacijskega gradiva ter z rezultati človeške evalvacije. Zaključili so, da so metrike na podlagi popravljenih strojnih prevodov najbolj skladne z ročno evalvacijo ter da v primeru metrike HTER dosegajo celo večjo skladnost z ročno evalvacijo kot več ročnih evalvacij različnih prevajalcev med seboj.

3 OPIS EKSPERIMENTOV

V eksperimentih, s katerimi smo preskušali odvisnost avtomatskih metrik od evalvacijskega gradiva, smo uporabljali gradivo in sistem strojnega prevajanja v okviru projekta SUMAT. Evropski projekt SUMAT[13] (angl. *An Online Service for SUbtitling by MAchine Translation*; projekt financira EU po pogodbi ICT-PSP-270919) se je začel v letu 2011, njegov namen pa je izdelati spletno aplikacijo za prevajanje podnapisov za različne evropske jezike. Podnaslavljanje je namreč priljubljen način za posredovanje tujejezičnih multimedijskih vsebin v veliko evropskih državah in za večino žanrov. Trenutna evropska politika (European Commission 2010) podpira podnaslavljanje v javnih televizijskih mrežah in posledično se je potreba po podnaslavljanju v avdiovizualni industriji v preteklih letih povečala (MCG 2007). Hkrati se prevajanje podnapisov srečuje s pomembnimi problemi, kot so visoki stroški, časovna potratnost in posledično vprašanje kvalitete podnapisov. SUMAT skuša zato razviti sistem, ki bi prevajalcem pomagal s

[13] http://www.sumat-project.eu

predhodnim strojnim prevodom podnapisov, tako da bi prevajalec prevode le popravil.

Za potrebe projekta SUMAT so bili izdelani vzporedni korpusi podnapisov za vse jezikovne pare, ki jih projekt pokriva, med drugim tudi za par slovenščina-srbščina, ki bo predmet analiz v tem prispevku. Zbrano gradivo SUMAT za ta jezikovni par je opisano v Sepesy Maučec idr. (2012) in obsega približno 110.000 poravnanih podnapisov. Poleg gradiva iz projekta SUMAT so bila za učenje strojnega prevajalnika uporabljena tudi gradiva iz korpusa OPUS OpenSubtitles (Tiedemann 2009) za ta jezikovni par, od katerih smo po dodatnem čiščenju[14] uporabili približno 2 milijona poravnanih podnapisov. Skupaj je učno gradivo prevajalnika obsegalo dobrih 15 milijonov besed za vsak jezik.

Iz SUMAT-ovega gradiva je bilo izločenih 4.000 parov podnapisov, ki so služili kot osnovno evalvacijsko gradivo – v nadaljevanju imenovano SUMATeval. Pri tem evalvacijskem gradivu smo opažali nekatere lastnosti, ki so pri strojnem prevajanju težko premostljive. V uvodu smo že omenili, da je prevajanje podnapisov oz. podnaslavljanje specifičen način prevajanja, saj mora prevajalec upoštevati časovne omejitve in omejitve dolžine podnapisov, hkrati mora upoštevati tudi to, da lahko mnoge informacije gledalec dobi iz video- in avdiovsebine, prevodi morajo velikokrat ustrezno prenašati nekatere značilnosti govorjenega jezika v pisni medij, izraziti pa so tudi vplivi različnih kultur na jezik. Poleg navedenega je posebnost SUMAT-ovega gradiva ta, da podnapisi niso neposredni prevodi eden drugega, ampak so nastali neodvisno drug od drugega s prevodom iz angleškega avdio-video gradiva (pisni angleški izvirnik ni bil na voljo) v srbščino oz. na drugi strani v slovenščino. Podobno lahko sklepamo za korpus OpenSubtitles. Razlog je seveda ta, da televizije večinoma predvajajo produkcijo iz angleško govorečih okolij, tako da gre za lastnost, ki jo lahko pripišemo podnapisom na splošno in je značilna za večino jezikovnih parov, ki ne vključujejo angleščine. To je vsekakor oteževalna

[14] Zaradi nepreverjenega izvora in kvalitete gradiva, zajetega v korpus OpenSubtitles, je vmes veliko nepravilno poravnanih prevodnih enot in drugih napak.

okoliščina tudi za strojno prevajanje: vemo namreč, da se prevod razlikuje od izvirnika – pri podnapisih toliko bolj, ker je izvirnik govorjeno besedilo, ki ga spremlja slika, in ker je dolžina podnapisa omejena. Kot primer različnih odločitev, ki sta jih sprejela prevajalca, lahko recimo zelo pogosto opazimo angleški »you« preveden v ednino v enem in množino v drugem jeziku. Tipični primeri prevodnih parov podnapisov v SUMATEval so na primer takšni (prikazani so pari celotnih podnapisov, ki se pojavijo na televizijskem zaslonu):[15]

slovenski prevod SUMATeval	**srbski izvirnik**
teh zvokcev ne morejo slišati . zakaj si tako jezen name ?	*ovo se ni ne čuje . zašto ste tako ljuti na mene ?*
namenjena si bila na sever . na kakšen določen kraj ?	*idete na sever . imate li neki razlog ?*
a tam le drživa vse niti	*ovo nam je samo mesto odakle rukovodimo poslovima*

Kot vidimo, se prevodni pari tudi pomensko pogosto razlikujejo, saj je srbski prevajalec očitno drugače povzemal pomen angleškega izvirnika kot slovenski. Takšno evalvacijsko gradivo je zelo težavno za avtomatsko evalvacijo, saj pri tej, kot smo videli, štejejo le razlike ter ujemanja med referenčnim in strojnim prevodom. V zgornjih primerih pa lahko ugotovimo, da bi celo človeški prevajalec zelo verjetno srbske vhodne podnapise prevedel v slovenščino drugače, kot so prevedeni v referenčnem gradivu, še toliko bolj velja to za strojni prevod. Posledično pričakujemo zelo slabe ocene strojnih prevodov s pomočjo avtomatskih metrik.

V drugem koraku smo iz obstoječih 4.000 parov podnapisov naključno izbrali 1.000 srbskih podnapisov in jih na novo prevedli v slovenščino, tokrat samo na podlagi srbskega besedila (brez angleškega izvirnika in brez videa, saj ne

[15] Vsi primeri v prispevku so zapisani na enak način kot v evalvacijskem gradivu, to je z malimi začetnicami in ločili kot samostojnimi pojavnicami. Prevajalni sistem sicer pretvori zapis velikih začetnic in ločil v običajen zapis pri postprocesiranju izhoda iz prevajalnika.

eno ne drugo ni bilo na voljo) – novo evalvacijsko gradivo smo poimenovali FERIeval. Prevajalec je bila oseba z dobrim razumevanjem srbščine in strokovnjak za slovenski jezik. Ob tem je bilo navodilo, naj bo slovenski prevod pomensko in jezikovno neoporečen, vendar naj prevajalec v primeru izbire med različnimi prevodnimi možnostmi daje prednost dobesednemu prevodu. Za iste pare podnapisov kot zgoraj smo tako dobili naslednje besedilo:

slovenski prevod FERIeval	**srbski izvirnik**
to se niti ne sliši . zakaj ste tako jezni name ?	*ovo se ni ne čuje . zašto ste tako ljuti na mene ?*
greste na sever . imate kakšen razlog ?	*idete na sever . imate li neki razlog ?*
s tega mesta samo vodiva posle	*ovo nam je samo mesto odakle rukovodimo poslovima*

Kot navajamo v razdelku 2.5, so najbližje rezultatom ročne evalvacije rezultati avtomatskih metrik, če jih računamo na podlagi popravljenih strojnih prevodov. To je bil tretji korak naših eksperimentov, zato smo slovenske prevode strojnega prevajalnika (vseh 1.000 podnapisov iz evalvacijskega gradiva) popravili, tako da smo naredili vse potrebne popravke, da smo dobili jezikovno pravilen in pomensko ustrezen[16] prevod. Popravljanje prevodov je opravila oseba z dobrim razumevanjem srbščine in strokovnjak za slovenski jezik. Ta nabor imenujemo v nadaljevanju (H)FERIeval. Zgornji prevodni pari podnapisov so bili tako iz naslednjega izhoda prevajalnika:

slovenski prevod prevajalnika SUMAT	**srbski izvirnik**
to se ne sliši . zakaj ste tako jezni name ?	*ovo se ni ne čuje . zašto ste tako ljuti na mene ?*
gresta na sever . imate kakšen razlog ?	*idete na sever . imate li neki razlog ?*

[16] Pri tem gremo celo nekoliko dlje kot Snover idr. (2006), saj so tam strojni prevod popravljali le s pomočjo drugih referenčnih prevodov, ne s pomočjo izvirnika.

to nam je samo mesto , od kod rukovodimo poslih	*ovo nam je samo mesto odakle rukovodimo poslovima*

popravljeni v različice:

slovenski prevod (H)FERIeval	**srbski izvirnik**
to se ne sliši . zakaj ste tako jezni name ?	*ovo se ni ne čuje . zašto ste tako ljuti na mene ?*
greste na sever . imate kakšen razlog ?	*idete na sever . imate li neki razlog ?*
to je mesto , od koder vodimo posle	*ovo nam je samo mesto odakle rukovodimo poslovima*

Ob tem delu smo dobili natančnejši vpogled tudi v tipe napak, ki jih prevajalnik dela. Za celostno sliko kvalitete prevajalnika je praviloma priporočljiva tudi ročna analiza napak, zato smo v zadnjem koraku naredili še to. V razdelku 4 opisujemo rezultate vseh opravljenih evalvacij.

4 REZULTATI EVALVACIJE

4.1 Rezultati avtomatskih metrik

Evalvirali smo prevode, ki jih je tvorila osnovna različica sistema za strojno prevajanje iz srbščine v slovenščino brez upoštevanja oblikoskladenjskih informacij. Pri gradnji modela prevajanja smo uporabili privzete nastavitve orodja Moses (Koehn idr. 2007). Prevajalni sistem je bil učen na gradivu, opisanem v razdelku 3. Sistem temelji na poravnavi besed učnega gradiva, iz katerega izlušči pare prevodnih fraz. Vsaki prevodni frazi pridruži tudi verjetnost iz 3-gramskega jezikovnega modela slovenskega jezika (Sepesy Maučec idr. 2004).

Rezultate prevajanja smo ovrednotili z metrikami NIST, BLEU in TER. Najvišja možna vrednost metrik BLEU in TER je 100 %. Vrednosti metrike NIST niso normalizirane, zato ne moremo govoriti o njeni zgornji meji. Za metriki BLEU in NIST velja, da višji, kot sta, boljši je prevod. Tipično se vrednosti BLEU gibljejo okrog 40 ali 50 % in NIST okrog 8. Za vrednost TER

velja, da nižja, kot je, boljši je prevod. Tipično se vrednosti gibljejo okrog 30–40 %. Pri TER nas je poleg osnovne vrednosti metrike zanimalo tudi, koliko vstavitev, izbrisov, zamenjav in premikov je bilo potrebnih za preslikavo avtomatskih prevodov v referenčni prevod. Tabela 1 prikazuje vrednosti izbranih metrik za izhod strojnega prevajalnika SUMAT.

	NIST	**BLEU**	**TER**	**Vstavitev**	**Izbris**	**Zamenjava**	**Premik**
SUMAT eval	5,05	19,47 %	65,27 %	3.102 10,27 %	3.125 10,36 %	12.564 41,60 %	1156 3,83 %
FERI eval	7,78	43,10 %	32,91 %	372 3,84 %	400 4,14 %	1915 20,83 %	218 2,26 %
	(H)NIST	(H)BLEU	(H)TER	(H)Vstavitev	(H)Izbris	(H)Zamenjava	(H)Premik
(H)FERIeval	10,62	71,6 %	14,1 %	154 1,78 %	189 2,18 %	768 8,85 %	97 1,12 %

Tabela 1: Rezultati avtomatske evalvacije.

Rezultati avtomatske evalvacije na gradivu SUMATeval po vseh uporabljenih metrikah sporočajo, da sistem tvori neuporabne prevode. Če nekoliko poenostavimo, BLEU pove, da se v povprečju le 19,47 % n-gramov v prevodih ujema z referenco. Pri tem so upoštevani n-grami do reda 4, tj. od unigramov do 4-gramov. O slabem rezultatu poroča tudi TER. Podrobnejši vpogled v rezultate TER kaže, da je bilo treba 41 % besed zamenjati, 10 % besed izbrisati in ravno toliko vstaviti.

Rezultati evalvacije istega izhoda prevajalnika na gradivu FERIeval pokažejo povsem drugačno sliko. Po vseh treh metrikah dosega sistem ocene, ki so tipične za prevajalnike s sprejemljivo kvaliteto, prevode lahko po tej ocenitvi štejemo za uporabne. Tudi deleži vstavitev, izbrisov in zamenjav pri TER so se več kot razpolovili.

Pri evalvaciji prevodov na podlagi popravljenega strojnega prevoda (hFERIeval) so rezultati ponovno pri vseh treh metrikah še bistveno boljši. Iz njih sledi, da je delež pravilnih prevodov zelo visok in da smo dobili zelo

uporaben prevajalni sistem. Ti rezultati sicer tudi sovpadajo z rezultati, ki so običajni za jezikovne pare sorodnih jezikov, kot sta slovenščina in srbščina. Praviloma je namreč strojno prevajanje med sorodnimi jeziki uspešnejše od prevajanja med jeziki iz različnih jezikovnih skupin.

Razlike med rezultati vseh treh nizov evalvacije so izredne: od ocene, da imamo neuporaben sistem, smo prišli do ocene, da imamo uspešen sistem. Kako je to mogoče? Poglejmo še enkrat primer, ki smo ga navedli zgoraj:

slovenski prevod prevajalnika SUMAT	**srbski izvirnik**
to se ne sliši . zakaj ste tako jezni name ?	*ovo se ni ne čuje . zašto ste tako ljuti na mene ?*

V prvem nizu evalvacij je metrika primerjala ujemanje med tema prevodoma:

slovenski prevod prevajalnika SUMAT	**slovenski prevod SUMATeval**
to se ne sliši . zakaj ste tako jezni name ?	*teh zvokcev ne morejo slišati . zakaj si tako jezen name ?*

Rezultati avtomatske metrike so v takšnem primeru skrajno slabi, saj ni skoraj nobene skupne točke (skupne besede ali n-grama besed) med obema prevodoma.

V drugem nizu evalvacij je metrika primerjala med prevodoma:

slovenski prevod prevajalnika SUMAT	**slovenski prevod FERIeval**
to se ne sliši . zakaj ste tako jezni name ?	*to se niti ne sliši . zakaj ste tako jezni name ?*

Tukaj bo – nasprotno kot zgoraj – metrika prevodu pripisala zelo visoko pravilnost, saj se razlikuje samo v besedici *niti*.

V tretjem nizu evalvacij je metrika primerjala med prevodoma:

slovenski prevod prevajalnika SUMAT	**slovenski prevod (H)FERIeval**
to se ne sliši . zakaj ste tako jezni name ?	*to se ne sliši . zakaj ste tako jezni name ?*

Kot vidimo, sta tokrat prevoda identična, ocena metrike BLEU bo za ta podnapis 100 % oz. metrike TER 0 %. Samo človeški prevajalec pa lahko iz vseh treh prevodov prepozna, da gre za pomensko ustrezne in pravilne, čeprav različne prevode.

4.2 Ročna analiza tipov napak

Zaradi izrazite neskladnosti metrik na različnih evalvacijskih gradivih za isti izhod iz prevajalnika smo v zadnjem koraku naredili ročno analizo napak, ki se pojavljajo v prevajalniku.

4.2.1 SLOVNIČNE NAPAKE

a) Neujemanje oblik in izbira napačnih oblik

V izhodnem besedilu prevajalnika se (verjetno na mejah zlepljenih večbesednih enot, lahko tudi zaradi kalkiranja) pogosto zgodi neujemanje oblik v spolu, sklonu in številu (v nadaljevanju vedno navajamo samo izseke podnapisov, pri katerih se pojavi opazovani problem):

slovenski prevod prevajalnika SUMAT	**srbski izvirnik**	**slovenski prevod (H)FERIeval**
prevejani pošast	*prevejani monstrum*	*prevejana pošast*
na polici v shrambo	*na polici u špajzu*	*na polici v shrambi*
moji ljudje jih bo osvobodila	*moji ljudi će da ih oslobode*	*moji ljudje jih bodo osvobodili*

b) Kalkiranje srbskih slovničnih oblik

V izhodni prevod se lahko prenaša srbski oblikoskladenjski vzorec:

slovenski prevod prevajalnika SUMAT	**srbski izvirnik**	**slovenski prevod (H)FERIeval**
žival v glavnem , od papirja	*životinju uglavnom , od papira*	*v glavnem žival , iz papirja*
celo so nam tudi hrana , ki jo ješ , in steklenica , poklonjeni	*čak su nam i hrana koju jedeš , i ta boca , poklonjeni*	*celo hrana , ki jo ješ , in ta steklenica sta nam poklonjeni*
ali bo vse dobro , da mine	*ili će sve dobro da prodje*	*ali pa bo vse dobro minilo*

c) Neustrezen besedni red

slovenski prevod prevajalnika SUMAT	**srbski izvirnik**	**slovenski prevod (H)FERIeval**
hvala , moji dragi	*hvala vam , dragi moji*	*hvala , dragi moji*
zakaj potem jim vzamemo hrano ?	*zašto onda da im oduzmemo hranu ?*	*zakaj jim potem vzamemo hrano ?*
ste bili dober prijatelj	*bili ste divan prijatelj*	*bili ste dober prijatelj*

4.2.2 LEKSIKALNE NAPAKE

a) Večpomenskost

Kadar je za isto besedo v različnih kontekstih primeren različen prevod, prevajalnik pogosto ne izbere pravilnega prevoda:

slovenski prevod prevajalnika SUMAT	**srbski izvirnik**	**slovenski prevod (H)FERIeval**
morali večji orkester .	*trebale veći orkestar .*	*potreben večji orkester .*
vi mu boste rekli	*vi ćete mu reći*	*vi mu boste povedali*
ne želim posneti videl	*ne želim da snimimo video*	*ne želim posneti videa*

b) Ohranjanje besede iz prevajanega besedila

Kadar prevajalnik ne najde prevoda, ohrani besedo neprevedeno:

slovenski prevod prevajalnika SUMAT	**srbski izvirnik**	**slovenski prevod (H)FERIeval**
je zasvirao	*je zasvirao*	*je zaigral*
prejakim akcentima	*prejakim akcentima*	*preveč izrazitim naglasom*
v mace iz crtanog filma	*u mace iz crtanog filma*	*v muce iz risank*

4.2.3 PRAVOPISNE NAPAKE

Od pravopisnih napak so opazne predvsem napake pri vejici in kraticah; drugim v obstoječi verziji prevajalnika ni mogoče slediti, saj je celotno besedilo postavljeno v male začetnice, tudi lastna imena, ločila pa so samostojne pojavnice:

slovenski prevod prevajalnika SUMAT	**srbski izvirnik**	**slovenski prevod (H)FERIeval**
televizijskim , in filmskimi igralcem	*televizijskim i filmskim glumcima*	*televizijskim in filmskim igralcem*
st . james	*sent džejms*	*st. james*
če veš kaj mislim	*ako znaš šta mislim*	*če veš , kaj mislim*

5 ZAKLJUČEK

V prispevku smo se ukvarjali z vprašanjem avtomatske evalvacije strojnega prevajanja. Pri tem smo ves čas ocenjevali isti sistem strojnega prevajanja, spreminjali pa smo evalvacijsko gradivo, na podlagi katerega smo izvajali evalvacije. Običajno se pri avtomatski evalvaciji opozarja na pomanjkljivosti posameznih metrik, zlasti pogosto je tarča kritik najpogosteje uporabljana metrika BLEU. Kritike v zvezi z njo so na primer, da zanemari dejstvo, da so nekatere besede pomembnejše od drugih (če je na primer v prevodu izpuščena beseda *ne*, lahko to bolj vpliva na pravilnost prevoda kot izpust katere druge besede), ne ocenjuje celostne slovnične strukture povedi (problematično zlasti za daljše prevodne enote) in je manj primerna za pregibne jezike (kot v našem primeru), ker so mnoge napake povezane z napačnimi oblikami in so kot

takšne manj problematične kot napačne leksikalne enote. Kot povzema Koehn (2011), so eksperimenti pokazali še, da so rezultati metrike BLEU za prevode, narejene s strani prevajalca in torej brez dvoma veliko kvalitetnejše prevode, kot so strojni, le minimalno boljši od rezultatov evalvacije strojnih prevodov.

Z eksperimenti v tem prispevku smo pokazali, da bolj kot izbrana metrika na avtomatsko evalvacijo prevajalnika vpliva evalvacijsko gradivo in način, kako je bilo narejeno. To je še posebej izrazito pri izbranem besedilnem tipu v tej razpravi, tj. pri podnapisih, pri katerih je način prevajanja zelo specifičen: gre bolj za povzemanje izvirnega govorjenega besedila, ne za prevajanje, kot velja sicer za pisna besedila. Še dodatno je problem našega izvornega evalvacijskega gradiva izrazit, ker ni nastalo z neposrednim prevodom. Rezultati vseh treh izbranih metrik, NIST, BLEU in TER, so se tako v naših eksperimentih pri vseh treh evalvacijskih gradivih gibali dokaj skladno, medtem ko so rezultati posamezne metrike poskočili od ugotovitve, da imamo neuporaben prevajalnik, do ugotovitve, da imamo uspešen prevajalnik, in to samo na podlagi sprememb v evalvacijskem gradivu. Merilo, naj bo metrika »zanesljiva, torej lahko pričakujemo, da je sistem, ki mu metrika pripisuje visoke ocene, tudi zares dober« (razdelek 2), je tako bolj kot od same metrike vsaj v našem primeru odvisno od evalvacijskega gradiva in načina, kako je bilo pridobljeno. Ker so podnapisi precej specifično prevodno gradivo, ostaja predmet nadaljnjih eksperimentov, v kolikšni meri je ta problem prisoten tudi pri evalvaciji strojnega prevajanja drugih besedilnih tipov.

Avtorji avtomatsko evalvacijo najpogosteje uporabljajo kot povratno informacijo, ali so s posamezno spremembo v sistemu strojnega prevajanja dosegli izboljšavo ali ne. Ob tem se nujno postavlja vprašanje, ali bi dobili ne glede na to, katero od treh različnih gradiv bi uporabili za evalvacijo, pravilno informacijo o morebitni (ne)izboljšavi s posamezno spremembo v sistemu. Povedano s primerom: ali bi evalvacija na našem prvotnem evalvacijskem gradivu, SUMATeval, ki je po naši oceni zelo problematično, vrnila pravilno informacijo, če bi na primer v prevajalni sistem dodali informacije o lemi in želeli izvedeti, ali smo s tem sistem izboljšali, poslabšali ali pa lematizacija sploh nima vpliva. Upamo si trditi, da ne, vendar bo to treba še

eksperimentalno potrditi. Ob tem naj pripomnimo, da je razen dobre izbire evalvacijskega gradiva pri uporabi avtomatskih metrik ocenjevanja seveda nujno tudi upoštevanje statističnih lastnosti, tj. preverjanje, ali je izboljšava prevodov z vneseno spremembo v sistemu zgolj naključna ali pa ji lahko zaupamo, ker je statistično signifikantna (kar ocenimo s t. i. aproksimacijo naključnosti – angl. *approximate randomization*). Vse navedeno je lahko predmet nadaljnjega dela.

Ob treh povsem različnih rezultatih evalvacije našega sistema strojnega prevajanja se seveda sprašujemo, kateri rezultati so najbolj pravilni. Odgovor lahko da samo temeljita ročna evalvacija s strani več strokovnjakov (kot vemo, so lahko tudi med njihovimi ocenami velike razlike), ki pa je nismo izvedli, saj gre za zahtevno in obsežno delo. Poleg tega mora v primeru podnapisov ročna evalvacija kvalitete prevoda upoštevati tudi avdio- in videogradivo ter pravila podnaslavljanja. Prav obsežnost in zahtevnost dobre ter objektivne ročne evalvacije je tudi razlog za uporabo avtomatskih metrik. Eksperimente (seveda na povsem drugem gradivu) v tej smeri so naredili Snover idr. (2006), ki trdijo, da je evalvacija, ki temelji na popravljenem strojnem prevodu, najbolj skladna z rezultati ročne evalvacije. Nenazadnje je takšen zaključek logičen, saj je vprašanje, koliko dela ima prevajalec, da iz izhoda prevajalnika naredi jezikovno in pomensko pravilen prevod, ključno za oceno uporabnosti prevajalnika. Seveda pa je tudi pri tem še vedno treba upoštevati, da bi različni prevajalci izhod prevajalnika popravili različno in da imajo tudi metrike mnoge pomanjkljivosti. Tako morda vemo, koliko zamenjav, izbrisov, vrivanj in premikov bi moral narediti prevajalec pri popravljanju prevoda, še vedno pa ne vemo, kako kvaliteten je strojni prevod sam po sebi.

LITERATURA

Brown, P. F. , Pietra, S. A. D. , Pietra V. J. D., in Mercer, R. L. (1993): The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19 (2): 263–311.

Callison-Buch, C., Fordyce, C., Koehn, P., Monz, C., in Schroeder, J. (2007):

(Meta)evaluation of Machine Translation. *Proceedings of ACL-2007 Workshop on Statistical Machine Translation*.

Callison-Buch, C., Fordyce, C. , Koehn, P. , Monz, C., in Schroeder, J. (2008): Further Meta-evaluation of Machine Translation. *Proceedings of ACL-2008 Workshop on Statistical Machine Translation*.

European Commision (2010): Audiovisual Media Services Directive (AVMSD – 2010/13/EU). *Official Journal of the European Union*, 10. marec 2010.

He, X. (2007): Using Word-dependent Transition Models in HMM-based Word Alignment for Statistical Machine Translation. *Proceedings of the Second Workshop on Statistical Machine Translation*: 80–87. Praga.

Ittycheriah, A., in Roukos, S. (2005): A Maximum Entropy Word Aligner for Arabic-English Machine Translation. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing:* 89–96. Vancouver .

Koehn, P. (2010): *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Koehn, P., idr. (2007): Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the Annual Meeting of the Association for Computational linguistics*: 177–180. Association for Computational Linguistics.

Kovačič, I. (1996): Subtitling Strategies: A Flexible Hierarchy of Priorities. Traduzione multimediale per il cinema, la televisione e la scena / Multimediale Übersetzung für Film, Fernsehen und Bühne / Multimedia translation for film, television and the stage: atti del convegno internazionale: 297–305. Forlì.

Ma, Y., Stroppa, N., in Way, A. (2007): Bootstrapping Word Alignment via Word Packing. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*: 304–3011. Praga.

MCG – Media Consulting Group (2007): Study on Dubbing and Subtitling Needs and Practices in the European Audiovisual Industry. *On behalf of the Information Society and Media Directorate General and the Culture Directorate of the European Commission*.

Niessen, S., in Ney, H. (2004): Statistical Machine Translation with Scarce Resources Using Morpho-Syntactic Information. *Computational Linguistics*, 30 (2): 181–204.

Och, F. J., in Ney, H. (2004): The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 29 (1): 19–52.

Popović, M., in Ney, H. (2004): Improving Word Alignment Quality using Morpho-syntactic Information. *Proceedings of 20th International Conference on Computational Linguistics*: 310–314. Ženeva.

Maučec, M. S., Kačič, Z., in Horvat, B. (2004): Modelling Highly Inflected Languages. *Information Sciences*, 166 (1/4): 249–269. Dostopno prek: http://dx.doi.org/10.1016/j.ins.2003.12.004.

Maučec, M. S., in Kačič, Z. (2006): Statistical Machine Translation Using the IJS-ELAN Corpus. V: WILIAMS, B. (ur.). 5th International Conference On Language Resources And Evaluation, Genova, W06 Strategies for Developing Machine Translation for Minority Languages: Satelite Workshop: 87–90.

Maučec, M. S., in Brest, J. (2010): Reduction of Morpho-syntactic Features in Statistical Machine Translation of Highly Inflective Language. *Informatica (Vilnius)*, 21 (1): 95–116.

Maučec, M. S., idr. (2012): Izdelava slovensko-srbskega vzporednega korpusa podnapisov za razvoj strojnega prevajanja v projektu SUMAT. V: ERJAVEC, Tomaž (ur.), ŽGANEC GROS, Jerneja (ur.): *Zbornik Osme konference Jezikovne tehnologije, zvezek C*: 167–172.

Snover, M. G., Madnani, N., Dorr, B., in Schwartz, R. (2006): TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate. *Machine Translation*, 23 (2–3): 117–127.

Tiedemann, J. (2009): News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. V: Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R. (ur.): *Recent Advances in Natural Language Processing* (vol. V): 237–248. Amsterdam, Philadelphia: John Benjamins.

Turian, J. P., Shen, L. , in Melamed, I. D. (2003): Evaluation of Machine Translation and its Evaluation. *Proceedings of the Machine Translation Summit IX*. New Orleans.

Vičič, J. (2010): Strojno prevajanje in slovenščina. V: ERJAVEC, Tomaž (ur.), ŽGANEC GROS, Jerneja (ur.): *Zbornik Sedme konference Jezikovne tehnologije, zvezek C:* 47–52.

Vičič, J. (2012): Hitra postavitev prevajalnih sistemov na osnovi pravil za sorodne naravne jezike: *doktorska disertacija*. Ljubljana. Dostopno prek: http://eprints.fri.uni-lj.si/1778/.

Vintar, Š., Fišer, D., in Vrščaj, A. (2012): Were the Clocks Striking or Surprising? Using WSD to Improve MT Performance. *Proceedings of the 13th Conference on the European Chapter of the Association for Computational Linguistics*: 87–92. Avignon. Dostopno prek: http://aclweb.org/anthology-new/W/W12/W12-01.pdf.

Vrščaj, A. (2011): *Evalvacija strojnih prevajalnikov: Diplomsko delo*. Ljubljana: Filozofska fakulteta.

ON AUTOMATIC MACHINE TRANSLATION EVALUATION

An important task of developing machine translation (MT) is evaluating system performance. Automatic measures are most commonly used for this task, as manual evaluation is time-consuming and costly. However, to perform an objective evaluation is not a trivial task. Automatic measures, such as BLEU, TER, NIST, METEOR etc., have their own weaknesses, while manual evaluations are also problematic since they are always to some extent subjective.

In this paper we test the influence of a test set on the results of automatic MT evaluation for the subtitling domain. Translating subtitles is a rather specific task for MT, since subtitles are a sort of summarization of spoken text rather than a direct translation of (written) text. Additional problem when translating language pair that does not include English, in our example Slovene-Serbian, is that commonly the translations are done from English to Serbian and from English to Slovenian, and not directly, since most of the TV production is originally filmed in English.

All this poses additional challenges to MT and consequently to MT evaluation. Automatic evaluation is based on a reference translation, which is usually taken from an existing parallel corpus and marked as a test set. In our experiments, we compare the evaluation results for the same MT system output using three types of test set. In the first round, the test set are 4000 subtitles from the parallel corpus of subtitles SUMAT. These subtitles are not direct translations from Serbian to Slovene or vice versa, but are based on an English original. In the second round, the test set are 1000 subtitles randomly extracted from the first test set and translated anew, from Serbian to Slovenian, based solely on the Serbian written subtitles. In the third round, the test set are the same 1000 subtitles, however this time the Slovene translations were obtained by manually

correcting the Slovene MT outputs so that they are correct translations of the Serbian subtitles.

The results of MT evaluation were calculated for the metrics NIST, BLEU and TER. They were strikingly diverse, even though the system output was always the same: when calculated on the original translations from the parallel corpus, BLEU was 19.47%, TER 65.27% and NIST 5.05; when calculated on directly translated subtitles from Serbian to Slovenian, BLEU was 43.10%, TER 32.91% and NIST 7.78; when calculated on the manually corrected MT output, BLEU (also so-called hBLEU) was 71.6%, (h)TER 14.1% and (h)NIST 10.62.

Keywords: machine translation, evaluation, reference translation, BLEU, TER