Evalvacija slovensko-srbskih strojnih prevodov v projektu SUMAT

Mirjam Sepesy Maučec

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko
Smetanova 17, SI-2000 Maribor
mirjam.sepesy@um.si

Povzetek

V prispevku predstavljamo postopek in rezultate evalvacije prevodov statističnega strojnega prevajalnika podnapisov, ki smo ga razvili v okviru evropskega projekta SUMAT. Cilj projekta je bil razviti orodje za strojno prevajanje podnapisov, ki bi olajšalo delo profesionalnih prevajalcev. V projektu smo predlagali uporabo strojnih prevodov kot osnovo za kvalitetno prevajanje, ki vključuje tudi popravljanje strojnih prevodov, ki ga opravi profesionalni prevajalec. Uporaba strojnih prevodov je smiselna, če prevajalcu olajša delo oz. mu prihrani čas, zato je bila ključna naloga v projektu evalvacija prevodov, tako v smislu njihove kvalitete kot prihranka časa. V prispevku prikažemo rezultate evalvacije za prevajanje iz slovenščine v srbščino. Ta je pokazala, da je lahko sistem SUMAT učinkovito orodje za prevajalce podnapisov. V projektu je bila opravljena tudi evalvacija za prevajanje v obratni smeri, ki je dala podobne rezultate, saj gre za prevajanje med dvema sorodnima jezikoma.

Evaluation of machine translation for Slovenian – Serbian in the project SUMAT

This article describes the evaluation of statistical machine translation as carried out during the SUMAT project. The goal of this project was to build a tool for the automatic translation of subtitles that would help professional translators. Machine translation is useful if it makes subtitler's job easier and saves him/her time. The idea of the project was to use the translations obtained by the tool as the basis for post-editing, which should be done by professional translators in order to obtain translations of required quality. The crucial part in the project was the evaluation of machine translations quality, and the measurement of productivity gain/loss. The results for Slovenian – Serbian translation show that SUMAT system could be a useful tool for professional translators. This article only presents those results for translation from Slovenian to Serbian are presented. Similar results were however obtained for translation in opposite direction.

1. Uvod

Danes je večina besedilnega gradiva na voljo v elektronski obliki. To daje korpusnim in statističnim pristopom v jezikovnih tehnologijah veliko prednost, tudi na področju strojnega prevajanja. Statistično strojno prevajanje se je skozi številne raziskave pokazalo kot najučinkovitejši pristop k avtomatskemu prevajanju. Dodaten razlog za njegov uspeh je tudi ta, da za razvoj prevajalnika ni potrebno poglobljeno znanje o jezikih, med katerimi prevajamo.

Zahtevnost strojnega prevajanja je odvisna od žanra in domene besedil, ki jih prevajamo. Sprva je kazalo, da je prevajanje podnapisov, s katerim smo se ukvarjali v projektu SUMAT, za statistično strojno prevajanje zelo hvaležno področje, saj so povedi praviloma kratke. Toda podnapisi prinašajo tudi številne probleme. Ker gre za prevajanje podnapisov video vsebin, so nekateri problemi blizu problemom govorjenega jezika. Še večji problem pa je, da so se mora dolžina besedila podrejati dolžinam podnapisa, kar privede do številnih postopkov krajšanja izvornega besedila.

V projektu SUMAT smo razvili statistični strojni prevajalnik za prevajanje podnapisov med 14 jezikovnimi pari, med temi je tudi za slovenščino-srbščino, ki ga obravnavamo v tem članku. Ideja projekta je bila uporabiti že razvite metode statističnega strojnega prevajanja in zgraditi prevajalsko orodje, ki bo v pomoč profesionalnim prevajalcem pri generiranju kvalitetnih prevodov. Cilj ni bil, da bi prevajalnik tvoril brezhibne prevode, ampak da bi prevajalcu ponudil prevod, ki ga bo le-ta s čim manj dela preoblikoval v prevod željene oz. zahtevane kakovosti.

V projektu so bili uporabljeni prevodi profesionalnih prevajalcev, ki so v lasti prevajalskih podjetij, partnerjev v projektu in izven projekta niso dostopni.

2. Korpus SUMAT

Osnovno gradivo statističnega prevajalnika predstavlja vzporedni korpus. Od kvalitete vzporednega korpusa je neposredno odvisna uspešnost prevajanja, saj je uporabljen kot učni korpus prevajalnika. V projektu SUMAT so gradivo za vzporedni korpus iz svojih arhivov prispevala mednarodna podjetja, ki se profesionalno ukvarjajo s prevajanjem podnapisov. Izvorno gradivo ni neposredno uporabno, ampak ga je treba obdelati. Koraki procesiranja so: pretvorba v enoten format; poravnavanje dokumentov; tokenizacija; poravnavanje podnapisov in normalizacija (zapis z malimi črkami). Procesiranje korpusa SUMAT za slovensko-srbski jezikovni par je podrobneje predstavljeno v Maučec et al. (2012). Vzporedni korpus obsega 167.700 poravnanih podnapisov, kar je okrog 1,7 mio. besed v slovenskem jeziku in 2 mio. besed v srbskem.

Korpus takega obsega je za gradnjo »uporabnega« prevajalnika premajhen, zlasti pri visoko pregibnih jezikih, kot sta tako slovenščina kot srbščina; za tovrstne jezikovne pare potrebujemo za doseganje zadovoljive pokritosti besedišča čim večje korpuse, vsaj 10 mio. besed ali več. V projektu smo zato učni korpus dopolnili še z neprečiščenim gradivom iz prosto dostopnega korpusa OpenSubtitles (Tiedemann, 2009) (1,9 mio poravnanih podnapisov) in z interno zbirko prevodov popularnih filmov (44.500 podnapisov). Celoten korpus, ki smo ga uporabili za učenje prevajalnika, je vseboval 2,1 mio. podnapisov (16,8 mio besed za slovenski in 17,6 mio besed za srbski jezik).

Nepogrešljiva komponenta prevajalnika je tudi jezikovni model. Za njegovo učenje uporabimo enojezično gradivo. V projektu smo uporabili vse razpoložljivo gradivo iz prej omenjenih virov, tj. tudi tiste segmente, ki jih v pripravi vzporednega korpusa nismo uspeli poravnati. Slovenski korpus je obsegal 4,35 mio. podnapisov oz. 36 mio. besed, srbski korpus pa 4,56 mio. podnapisov oz. 42 mio. besed.

Slovar prevajalnika (tj. besede, ki jih prevajalnik prevaja) smo izluščili iz vzporednega korpusa. Tako je slovenski slovar vseboval 394.000 besed, srbski pa 570.000 besed.

3. Gradnja prevajalnika SUMAT

Prevajalnik SUMAT ima klasično strukturo. Kot osnovna enota prevajanja se običajno uporablja poved, v projektu pa je bilo opravljenih nekaj preliminarnih testov, ki so vodili v odločitev, da kot osnovno enoto uporabimo podnapis.

Prevajalnik sestavljajo 3 osnovne komponente: model prevajanja, model preurejanja in jezikovni model. Prvi dve komponenti smo zgradili s pomočjo orodja Moses (Koehn et al., 2007), jezikovni model pa z orodjem SRI LM (Stolcke, 2002).

Učni vzporedni korpus izhaja iz treh različnih virov, zato smo modela prevajanja in preurejanja gradili za vsak vir posebej in jih potem sestavili po principu adaptacije na domeno (Sennrich, 2012). Kot vzorec ciljne domene smo uporabili razvojno množico, ki je obsegala 2000 podnapisov.

Za izgradnjo jezikovnih modelov smo uporabili celotni korpus podnapisov. Uporabili smo 3-gramski jezikovni model z Good-Turingovim odštevanjem in sestopanjem po Katz. Perpleksnost slovenskega jezikovnega modela na testni množici je znašala 206, na srbski pa 230. Preizkusili smo tudi dodajanje gradiva iz enojezičnega korpusa pisanega jezika, ki rezultata ni izboljšalo.

Uteži komponent prevajalnika smo optimirali po MERT (Och, 2003) na razvojni množici 2000 podnapisov.

4. Evalvacija

V prvem delu smo izvedli avtomatsko evalvacijo s 4000 naključno izločenimi podnapisi, ki niso bili ročno pregledani. Uporabili smo metrike avtomatske evalvacije BLEU in TER (Papineni et al., 2002; Snover et al, 2006). Zanimal nas je tudi delež podnapisov, ki se 100% ujemajo z referenco (Equal), in delež podnapisov, pri katerih je, da dosežemo ujemanje, potrebnih največ 5 korakov preurejanja (Lev5). Rezultati evalvacije so v prvi vrstici v tabeli 1. Rezultati so slabi. Vzrok za to je v nastanku slovensko-srbskih SUMAT poravnanih podnapisov. Le-ti niso bili generirani kot neposredni slovensko – srbski prevodi, ampak oboji neposredno iz video signala v angleščini. To je prevajalcem iz angleščine v srbščino in slovenščino ponujalo veliko mero svobodne pri izbiri besed. Testno množico SUMAT smo podrobneje analizirali v (Verdonik & Maučec, 2013).

	BLEU	TER	Equal	Lev 5
Testna	17,80	66,10	4,00	11,60
Faza 1	46,30	36,20	13,80	38,90
Faza 2	57,40	26,30	22,10	47,90
Faza 3	69,20	17,30	37,40	69,10
SUMAT povprečje	39,69	44,88	20,1	35,69

Tabela 1: Rezultati evalvacije v različnih fazah projekta

V drugem delu evalvacije so bili v ocenjevanje kvalitete prevodov vključeni profesionalni prevajalci. Evalvacija s prevajalci je potekala v dveh sklopih. V prvem sklopu smo z njihovo pomočjo izboljševali sistem, v drugem delu pa merili, ali se učinkovitost prevajanja ob uporabi sistema izboljša, če se torej čas prevajanja z uporabo strojnih prevodov kaj skrajša.

4.1. Ocenjevanje s pomočjo prevajalcev

Prvi sklop evalvacije je potekal v treh fazah. Za vsako fazo je bila izbrana datoteka, ki je predstavljala zaključeno celoto, npr. podnapise celotnega filma, dokumentarca, pogovorne oddaje ipd. Datoteka je bila prevedena s pomočjo sistema in dana prevajalcu v pregled. Pregled je vključeval: rangiranje prevoda glede na kvaliteto, klasifikacijo napak in popravljanje prevoda v pravilnega. Prevajalci so v vsaki fazi v posebni datoteki podali tudi predloge popravkov.

4.1.1. Avtomatska evalvacija z ročno popravljenimi strojnimi prevodi za referenco

Strojne prevode dokumentov vseh treh faz smo avtomatsko evalvirali, tako da smo kot referenco uporabili popravljene prevode, ki so jih zapisali prevajalci. Rezultati so zbrani v vrsticah od 2 do 4 v tabeli 1. Vidimo, da so rezultati neprimerno boljši kot v primeru testne množice SUMAT. Razvidno je tudi, da so se rezultati iz faze v fazo izboljševali. V zadnji vrstici v tabeli 1 so povprečni rezultati avtomatskih metrik po vseh jezikovnih parih projekta. Tudi iz te primerjave lahko razberemo, da so bili v primerjavi z drugimi jezikovnimi pari za prevajanje slovenščina-srbščina doseženi zelo dobri rezultati. To je pričakovano, saj gre za sorodna jezika.

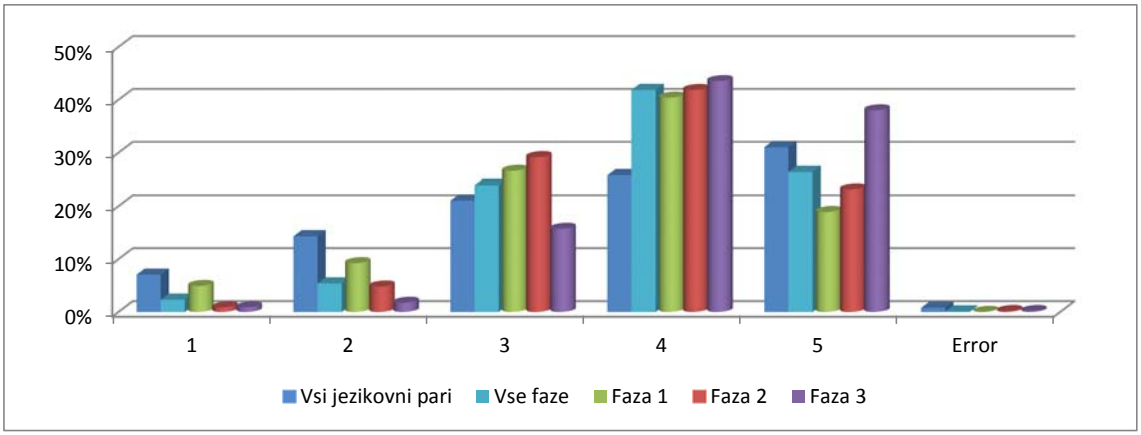
4.1.2. Rangiranje prevodov

Prevajalci so vsak podnapis v strojnem prevodu rangirali glede na kvaliteto oz. zahtevnost popravljanja. Pri tem smo uporabili skalo, definirano v "WMT 2012 Shared Task on MT quality estimation", po kateri je vsak podnapis rangiran z vrednostjo od 1 do 5. Ocena 1 pomeni neuporaben in nerazumljiv prevod, ocena 5 pa brezhiben prevod, ki ne potrebuje nobenega popravka. Rezultati za prevode slovenščina-srbščina so zbrani v tabeli 2. Vidimo, da je največ prevodov dobilo oceno 4. Več kot 20 % prevodov ima oceno 5, kar pomeni, da je petina prevodov neposredno uporabnih. Tudi iz te tabele je razvidno, da so se rezultati iz faze v fazo izboljševali.

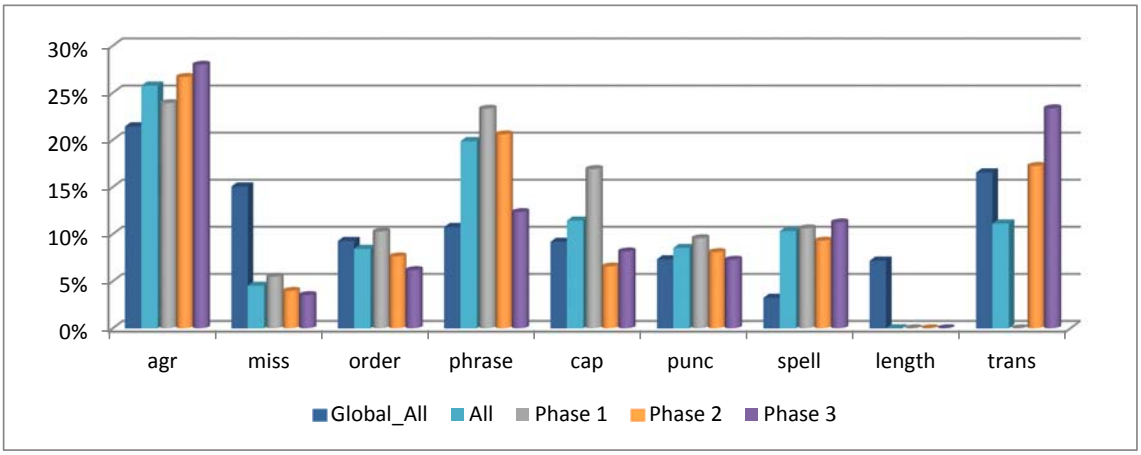
4.1.3. Klasifikacija napak

Prevajalci so napake v prevodih klasificirali v razrede:

- *agr*: slovnično neujemanje,
- *miss*: manjka polnopomenska beseda ali odsek,
- *order*: napačni vrstni red besed,
- *phrase*: večbesedna zveza napačno prevedena kot ločene, nepovezane besede,
- *cap*: napačen zapis velike/male črke,
- *punc*: napačno ločilo,
- *spell*: napačno črkovanje,



Slika 1: Rangiranje prevodov glede na oceno kvalitete



Slika 2: Deleži napak v različnih fazah projekta

- *length*: predolg prevod glede na omejeno dolžino podnapisa,
- *trans*: napačen prevod.

Slika 1 prikazuje deleže napak po fazah. Vidimo lahko, da se je delež nekaterih napak skozi faze manjšal (npr. *cap*, *punct*, *phrase*, *order*), nekaterih pa celo povečal (npr. *trans*).

4.1.4. Popravki v sistemu

Na osnovi klasifikacije napak in predlogov prevajalcev smo sistemu dodali nekaj korakov naknadne obdelave strojnih prevodov. Glede na končna ločila smo popravili velike začetnice besed. Dodali smo nekaj 100 pravil za popravljanje slovničnega neujemanja. Definirali smo tudi nekaj pravil za zapis števil. Brisali smo presledke pred ločili.

Napačnih prevodov nismo uspeli popravljati. Vzrok za nekatere napačne prevode je neupoštevanje konteksta. Prevajalnik obravnava podnapis kot zaključeno celoto, ne glede na vsebino predhodnih podnapisov.

Reševanje določenih napak je pogojeno z uporabo dodatnih jezikovnih virov, ki jih zaradi komercialne naravnanosti projekta nismo dodajali, saj je za vsak uporabljen vir potrebno dovoljenje za komercialno rabo.

Nekatere napake smo odpravili tudi s tem, da smo dodali še en korak optimizacije uteži z MERT, tako da smo razvojno množico nadomestili z datotekami iz drugega dela evalvacije.

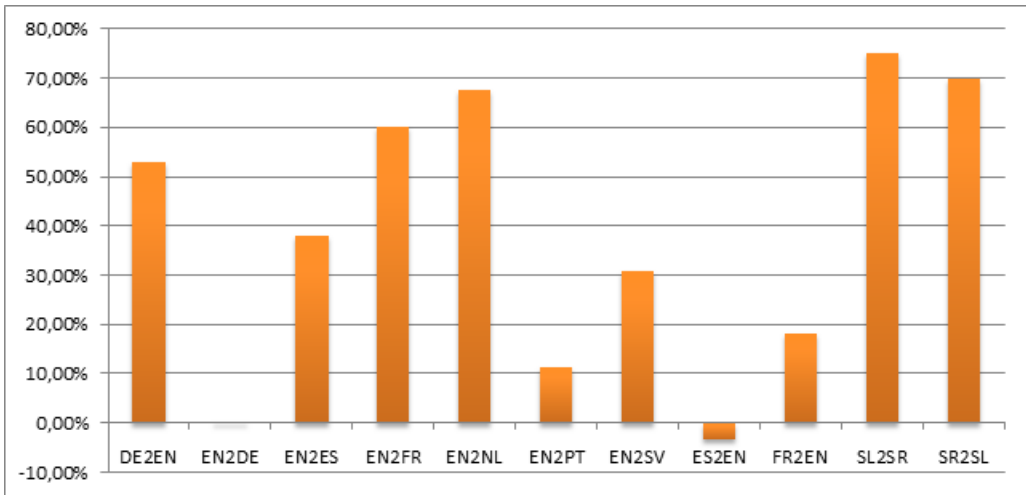
V fazi izboljševanja sistema učnega korpusa nismo spreminjali, čeprav smo ugotovili, da OpenSubtitles korpus vsebuje veliko šuma. V okviru projekta bi bilo čiščenje korpusa časovno preveč zahtevno.

4.2. Merjenje produktivnosti

V drugem delu evalvacije s prevajalci smo merili učinkovitost uporabe strojnih prevodov, ki jih je generiral izboljšan sistem. Primerjali smo čas, ki ga potrebuje prevajalec, če neposredno prevaja dokument iz izvornega v ciljni jezik, s časom, ki ga potrebuje za naknadno obdelavo strojnih prevodov. Menimo, da je tovrstna primerjava zelo jasen in neposreden pokazatelj uporabnosti sistemov strojnega prevajanja.

Pred izvedbo drugega dela evalvacije smo v sistem prevajanja vpeljali še dodaten postopek filtriranja strojnih prevodov. V razdelku 5.1 smo opisali rangiranje prevodov glede na kvaliteto. Na osnovi teh ocen smo učili binarni klasifikator, ki prevode klasificira v dva razreda, v razred dobrih in razred slabih prevodov. Za učenje klasifikatorja in klasifikacijo smo uporabili orodje QuEst, ki je podrobneje opisano v (Specia et al., 2013). Strojne prevode, ki jih je klasifikator označil kot slabe, smo odstranili, kar je pomenilo, da jih mora prevajalec tvoriti iz podnapisa v izvornem jeziku.

Za vsak jezikovni par oz. za vsako smer prevajanja sta sodelovala dva profesionalna prevajalca. Izjema je jezikovni par slovenščina-srbščina, kjer je za vsako smer prevajanja sodeloval le en prevajalec. Vsak prevajalec je tvoril tri datoteke. V prvi je prevajal iz izvornega jezika, v drugi je popravljal strojne prevode in v tretji je popravljal le filtrirane strojne prevode. Pri tem je vsak prevajalec



Slika 3: Rast produktivnosti pri uporabi strojnih prevodov v prevajalskem procesu

uporabil programsko okolje, ki ga tudi sicer uporablja pri svojem delu. Razlika je bila le v tem, da se je v ozadju meril čas efektivnega dela.

Iz primerjave časov, potrebnih za generiranje prevodov v prvi in drugi datoteki oz. prvi in tretji datoteki, smo izračunali rast/padec produktivnosti (ang. productivity gain/loss) pri prevajalskem procesu. Rezultati učinkovitosti uporabe strojnih prevodov za vse jezike v projektu SUMAT so prikazani na sliki 2. Vidimo, da je pri prevajanju srbščina-slovenščina in obratno prihranek časa največji. To je za sorodna jezika pričakovano. Strojni prevodi so lahko zelo učinkovita vmesna faza pri prevajanju tudi za večino drugih jezikov.

Omenimo še en vidik uporabe strojnih prevodov. Za prevajalce popravljanje strojnih prevodov ni najbolj »všečen« proces in nekateri do tega čutijo določen odpor. V tem oziru so lahko prikazani rezultati do neke mere popačen prikaz, subjektivna percepcija strojnega prevajanja profesionalnih prevajalcev.

5. Zaključek

V članku smo predstavili sistem strojnega prevajanja za podnapise, ki smo ga razvili v projektu SUMAT. Podrobneje smo opisali evalvacijo kvalitete prevodov in učinkovitosti uporabe strojnih prevodov v prevajalskih procesih profesionalnih prevajalcev. Evalvacija s pomočjo prevajalcev je pokazala, da je kvaliteta prevodov za jezikovni par slovenščina-srbščina na visoki ravni. Povprečna ocena prevodov je več kot 3,5. Kar 40 % prevodov je dobilo oceno 4, kar pomeni, da je bilo potrebnih le malo popravkov za zagotavljanje običajne kvalitete prevodov.

Tudi produktivnost prevajalca se lahko z uporabo strojnih prevodov poveča, zahteva pa od prevajalca prilagajanje na nov način dela. Zaenkrat je popravljanje strojnih prevodov še relativno nepoznan postopek med prevajalci. Da bi bilo strojno prevajanje pozitivno sprejeto med njimi, bi bilo treba učenje tehnik popravljanja vključiti tudi v učne procese v prevajalstvu.

6. Literatura

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., 2007. Moses: Open source toolkit for statistical machine translation. *Zbornik 45th Annual Meeting of the ACL*, 177–180.

Och, F. J., 2003. Minimum error rate training in statistical machine translation, *Zbornik 41st Annual meeting of the Association for Computational Linguistics,* Sapporo, Japan.

Papineni, K., Roukos, S., Ward, T., Zhu. W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. *40th Annual meeting of the Association for Computational Linguistics*, Philadelphia, 311–318.

Sennrich, R., 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. *Zbornik 13th Conference of the European Chapter of the Association for Computational Linguistics,* 539–549.

Sepesy Maučec, M., Presker, M., Zimšek, D., Rojc, M., Vlaj, D., Verdonik, D., Kačič, Z., 2012. Izdelava slovensko-srbskega vzporednega korpusa podnapisov za razvoj strojnega prevajanja v projektu SUMAT. *V: ERJAVEC, T. (ur.), ŽGANEC GROS, J. (ur.). Zbornik Osme konference Jezikovne tehnologije*, 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.

Specia, L., Shah, K., de Souza, J. G., Cohn, T., Kessler, F. B., 2013. QuEst–a translation quality estimation framework. Zbornik 51st *Annual meeting of the Association for Computational Linguistics* : System Demonstrations, 79–84.

Stolcke, A., 2002. SRILM: an extensible language modeling toolkit. *Proceedings of the Int. Conf. on Spoken Language Processing*, 901–904.

Tiedemann, J., 2009. News from OPUS - a collection of multilingual parallel corpora with tools and interfaces, *Recent Advances in Natural Language Processing*, vol. V, 237–248.

Verdonik, D., Sepesy Maučec, M., 2013. O avtomatski evalvaciji strojnega prevajanja. *Slovenščina 2.0*, 2013, št. 1, 111–133.