PRIMERJAVA OBIČAJNIH IN FAKTORSKIH MODELOV PRI STATISTIČNEM STROJNEM PREVAJANJU IZ ANGLEŠČINE V SLOVENŠČINO Z ORODJEM MOSES

Sašo KUNTARIČ

Simon KREK
Filozofska fakulteta Univerze v Ljubljani, Inštitut »Jožef Stefan«

Marko ROBNIK ŠIKONJA
Fakulteta za računalništvo in informatiko Univerze v Ljubljani

Kuntarič, S., Krek, S., Robnik Šikonja, M. (2017): Primerjava običajnih in faktorskih modelov pri statističnem strojnem prevajanju iz angleščine v slovenščino z orodjem Moses. Slovenščina 2.0, 2017 (1): 1–26.
DOI: http://dx.doi.org/10.4312/slo2.0.2017.1.1-26.

Strojno prevajanje je področje računalniške lingvistike, ki raziskuje uporabo programske opreme za prevajanje besedila iz enega jezika v drugega. Faktorsko statistično strojno prevajanje je različica statističnega, pri katerem besedilu dodamo jezikoslovne oznake na ravni besed in jih spremenimo v vektorje. Tako želimo izboljšati kakovost dobljenih prevodov. V prispevku opišemo uporabo odprtokodnega sistema Moses za faktorsko statistično strojno prevajanje iz angleščine v slovenščino. Iz besedilnega korpusa smo ustvarili več faktorskih in nefaktorskih prevajalnih modelov. Z njimi smo prevedli dve besedili s področja informacijskih tehnologij. Prvo je usmerjeno tržno in ima kompleksnejšo zgradbo, drugo pa je bolj tehnične narave. Prevode, ki smo jih dobili, smo na dva načina primerjali z dvema neodvisnima človeškima prevodoma in s prevodom, ki smo ga ustvarili s storitvijo Google Translate. Za prvi način primerjave smo uporabili metriko BLEU, za drugega pa so prevode pregledali človeški pregledovalci in podali subjektivno oceno, ki je pri prevajanju še vedno zelo pomembna. Čeprav rezultatov ne moremo primerjati neposredno zaradi različnih metrik, se gibanje ocen kakovosti pri obeh besedilih dobro ujema. Edina

občutna razlika med računalniško in človeško oceno se pojavi pri prehodu na faktorske modele pri drugem besedilu. Analizirali smo zanesljivost ocenjevalcev in rezultate ocenjevanja. Ugotovili smo, da so naši modeli primernejši za tehnična besedila in da uporaba faktorskih modelov vidneje izboljša prevajanje kompleksnejših besedil.

Ključne besede: statistično strojno prevajanje, faktorsko strojno prevajanje, sistem Moses, BLEU, človeška evalvacija

1 UVOD

Prevajanje in lokalizacija sta panogi, ki v zadnjih desetletjih doživljata velike spremembe, saj se v spletu pojavlja vedno več uporabniško ustvarjenih vsebin (družabna omrežja, forumi itd.), za katere natančni in do potankosti pregledani prevodi niso potrebni, so predragi, za njihovo prevajanje pa ni na voljo dovolj časa, saj uporabniki te vsebine ustvarjajo neverjetno hitro. Uporabniku je v večini primerov dovolj, da približno razume smisel sporočila, popolna pravilnost in skladnost s slovnico pa mu nista tako pomembni. Vse to je vodilo k razmahu strojnega prevajanja, kjer bi želeli računalnik naučiti prevajati dovolj dobro, da bi obvladovali velike količine takšnih besedil, hkrati pa vseeno posredovali pomen sporočila uporabniku. Podobno kot na drugih področjih tudi pri strojnem prevajanju pomembno vlogo igra velikost tržišča. Tako je bilo za močno razširjene jezike narejeno veliko glede čim boljše kakovosti strojnega prevajanja. Slovenščina med te jezike žal ne spada, zato na tem področju nekoliko zaostajamo. Eden od razlogov za šibko podprtost in hkrati ovira za doseganje dobrih rezultatov je, da je slovenščina morfološko bogat jezik, zato je strojne prevajalnike težko naučiti slovnično pravilne slovenščine. Namen naše raziskave je ustvariti angleško-slovenski jezikovni korpus s tematiko informacijskih tehnologij, nato pa s tem korpusom v strojnem prevajalniku Moses (Koehn in dr. 2007) preizkusiti, ali lahko z uporabo faktorskega strojnega prevajanja izboljšamo kakovost dobljenih prevodov.

Korpus, ki smo ga uporabili, je rezultat desetletnega prevajanja prevajalske

agencije. Iz tega korpusa smo ustvarili različne modele za strojni prevajalnik – od osnovnega nefaktorskega modela brez prerazporejanja besed v dobljenih prevodih do bolj zapletenih faktorskih modelov s prevajanjem, prerazporejanjem in ustvarjanjem novih besednih zvez v prevodu. Vsakega od modelov smo preizkusili na dveh različnih besedilih. Eno je usmerjeno bolj tržno in ima zapleteno zgradbo, drugo je bolj tehnične narave. Dobljene prevode smo primerjali z dvema različnima človeškima prevodoma in s prevodom, ki smo ga ustvarili s storitvijo Google Translate. Za prvi način primerjave smo uporabili metriko BLEU, za drugega so prevode pregledali človeški pregledovalci in podali subjektivno oceno, ki je pri prevajanju še vedno zelo pomembna. Sklepali smo, da bomo višje ocene dobili pri prevajanju drugega, bolj tehničnega besedila, prehod na faktorske modele pa bo vidneje izboljšal oceno prevoda prvega besedila, saj bo imelo dodajanje oblikoslovnih informacij večji vpliv pri bolj tekočem besedilu.

V nadaljevanju si bomo v 2. poglavju pobližje ogledali vrste strojnega prevajanja, ki jih bomo uporabljali, in v 3. poglavju na kratko opisali strojni prevajalnik Moses. V 4. poglavju si bomo ogledali prevajalne modele, ki jih bomo uporabljali. V 5. poglavju bomo podali analizo besedil in evalvacijo rezultatov, v zadnjem poglavju pa bomo povzeli narejeno, povedali, kako bi lahko naše rezultate izboljšali, ter navedli predloge za nadaljnje delo.

2 STROJNO PREVAJANJE

Strojno prevajanje, ki ga ne smemo zamenjevati z računalniško podprtim prevajanjem, je postopek, pri katerem računalniški program analizira besedilo in brez posredovanja človeka ustvari prevedeno besedilo. Vloga uporabnika pri takih sistemih ni vselej povsem odpravljena, pri interaktivnih sistemih je za razreševanje večpomenskosti predvidena človekova pomoč, skoraj pri vseh sistemih pa je potrebna predpriprava izvornega besedila in poprava prevedenega besedila (Vintar 1999).

2.1 Statistično strojno prevajanje

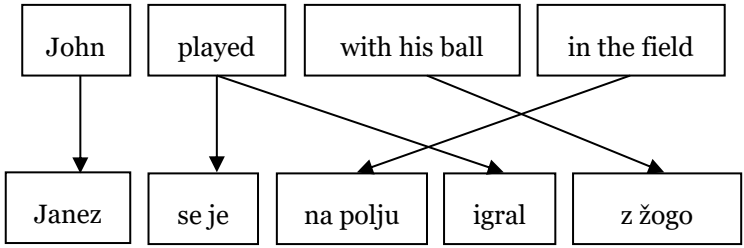
Statistično strojno prevajanje je vrsta strojnega prevajanja, ki temelji na analizi večje količine vzporednih besedil, iz katerih se izračunavajo verjetnosti prevajalne ustreznosti za posamezne jezikovne možnosti. Zamisel izhaja iz informacijske teorije. Dokument se prevede v skladu z verjetnostmi $p(e|f)$, da je niz *e* v ciljnem jeziku prevod niza *f* v izvornem jeziku (Koehn 2010). Najprej moramo ustvariti mehanizem, ki vsakemu slovenskemu stavku *e* dodeli verjetnost $p(e)$. Temu pravimo jezikovni model. Za računalnik je stavke najlažje razdeliti na podnize, ki so lahko različnih dolžin. Podniz z *n* besedami se imenuje n-gram. Če so deli ustrezni in se združujejo na pravilne načine, potem pravimo, da je niz slovenski. V naslednjem koraku se moramo osredotočiti na $p(f|e)$, verjetnost angleškega niza *f*, če imamo slovenski niz *e*. Temu pravimo ustvarjanje prevajalnega modela.

2.2 Statistično strojno prevajanje po besednih zvezah

V orodju Moses, ki ga uporabljamo, smo se odločili za prevajanje po besednih zvezah in za njegovo razširitev, faktorsko prevajanje. Cilj takega prevajanja je zmanjšati omejitve prevajanja po besedah. Prednosti sta predvsem dve:

- Včasih se ena beseda v tujem jeziku prevede v več besed v ciljnem jeziku in obratno (na primer *played – se je igral*). Strojno prevajanje po besedah v takih primerih pogosto ne deluje pravilno.
- Prevajanje skupin besed nam pomaga, da razrešimo dvoumne prevode (beseda žvižgati ima na primer lahko (vsaj) dva pomena – *glasno je žvižgal pesem* ali *veter kar žvižga okoli ušes*).

Slika 1 prikazuje postopek prevajanja po besednih zvezah. Vhodni stavek se razdeli na nize zaporednih besed (ki se imenujejo besedne zveze). Vsaka angleška besedna zveza se prevede v slovensko, vrstni red dobljenih besednih zvez pa se lahko po prevajanju prerazporedi.

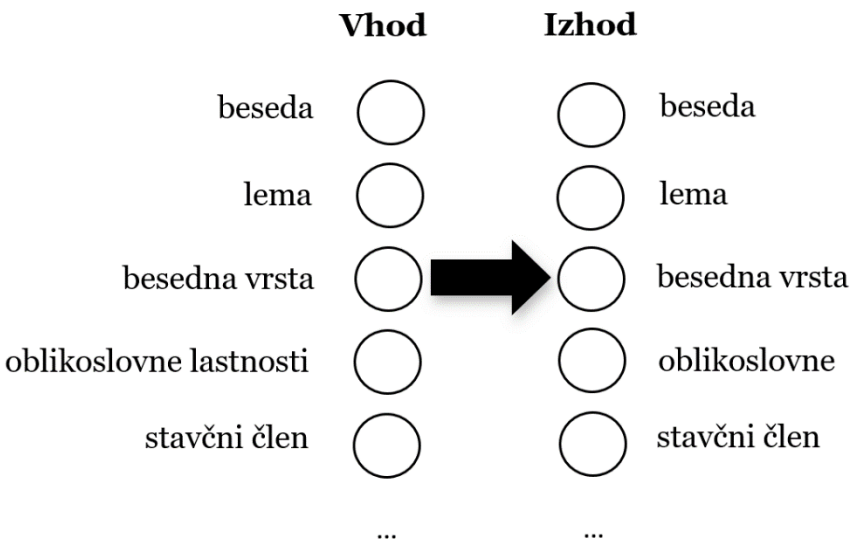


Slika 1: Pri prevajanju po besednih zvezah prevajamo besedne zveze in ne več posameznih besed.

Dokazano je bilo, da uporaba besednih zvez, daljših od treh besed, v korpusu za statistično strojno prevajanje na učinkovitost ne vpliva močno (Koehn in dr. 2003). Zaradi tega smo se odločili, da uporabimo trigramski model.

2.3 Faktorski prevajalni modeli

Faktorski prevajalni modeli dodajajo jezikoslovne oznake na ravni besed. Vsaka vrsta dodatnih informacij na ravni besede se imenuje faktor (Slika 2). Ločen prevod leme in dodani morfološki faktorji nam lahko pomagajo pri razdvoumljanju, poleg tega pa nam lahko dodatne informacije o stavčnih členih pomagajo pri preurejanju in zagotavljanju slovnične skladnosti. Prisotnost morfoloških značilnosti na ciljni strani omogoča preverjanje ustreznosti samostalniških besednih zvez ali odnosov med predmetom in povedkom.



Slika 2: Faktorski prevajalni model, kjer so besede predstavljene kot vektorji faktorjev.

Učenje faktorskih modelov

Za učenje faktorskih modelov moramo običajnim vzporednim korpusom na vhodni in izhodni strani dodati faktorske oznake. To običajno naredimo z avtomatiziranimi orodji, saj je ročno označevanje korpusov drago in počasno. V našem primeru smo za angleški del korpusa uporabili program MXPOST (Ratnaparkhi 1996), za slovenski del pa program Obeliks (Grčar in dr. 2012).

2.4 Ocenjevanje strojnih prevodov

Pri strojnem prevajanju običajno delamo z velikimi korpusi besedil, pri katerih bi bilo človeško pregledovanje rezultatov preveč časovno potratno in predrago. Zaradi tega so se pojavile pobude, da bi za to uporabljali računalniške algoritme. Ti bi iz ocenjevanja odstranili subjektivnost (kar je hkrati dobro in slabo), predvsem pa bi zadevo pospešili in pocenili. Izbiramo lahko med več različnimi ocenjevalnimi metrikami, kot sta METEOR (Banerjee in Lavie 2005) in LEPOR (Han in dr. 2012), mi pa smo se odločili za najpogosteje uporabljani BLEU

(Papineni in dr. 2002).

BLEU (*bilingual evaluation understudy*) je metrika za oceno kakovosti strojno prevedenega besedila iz enega naravnega jezika v drugega. Kakovost ocenjuje s primerjavo strojnega prevoda s človeškim. Osrednja ideja metrike BLEU je: »Bolj kot je strojni prevod podoben človeškemu, boljši je.« Je ena prvih metrik, ki je dosegla visoko korelacijo s človeško presojo kakovosti.

Za določanje ocene BLEU se strojni prevod primerja z enim ali več človeškimi (referenčnimi) prevodi istega stavka. Metrika BLEU ima, podobno kot druge samodejne ocene, veliko nasprotnikov, ki kot glavne slabosti navajajo naslednje:

- BLEU ne upošteva relativne pomembnosti posameznih besed. Nekatere besede so bolj pomembne od drugih. Primer je beseda *ne*, ki lahko močno spremeni prevod, če jo izpustimo.
- BLEU deluje na lokalni ravni in ne upošteva splošne slovnične doslednosti. Izhodni prevod je lahko videti dobro, če upoštevamo samo n-grame, širša primernost pa je vprašljiva.
- Dejanske ocene BLEU so brez pomena. Nihče v resnici ne ve, kaj pomeni rezultat 30, saj je metrika sestavljena iz velikega števila dejavnikov, na primer števila referenčnih prevodov, jezikovnega para, domene in drugih.

Za metriko BLEU smo se kljub pomanjkljivostim odločili zato, ker v praksi dosega visoko korelacijo s človeško oceno kakovosti, poleg tega pa je ena od najbolj priljubljenih samodejnih in nezahtevnih metrik.

3 STROJNI PREVAJALNIK MOSES

Moses je začel leta 2005 kot naslednika strojnega prevajalnika Pharaoh razvijati Hieu Hoang. Gre za implementacijo statističnega strojnega prevajanja, trenutno prevladujočega pristopa na tem področju.

Za učenje uporablja Moses vzporedne korpuse, za ugotavljanje ustreznih

prevodov med dvema izbranima jezikoma pa uporablja skupne pojavitve besed in besednih zvez. V Moses je vključena tudi razširitev strojnega prevajanja po besednih zvezah, ki se imenuje faktorsko strojno prevajanje. Moses je sestavljen iz dveh glavnih komponent: cevovoda za učenje in dekodirnika (Koehn in dr. 2007).

Moses in slovenščina

Kljub temu da strojno prevajanje za slovenščino ni tako dobro razvito kot za večje svetovne jezike, si Moses in slovenščina nista tujca. Tako je bil med drugim razvit sistem Asistent, ki omogoča prevajanje med angleščino, slovenščino, hrvaščino in srbščino (Arčan in dr. 2016), omeniti pa je treba tudi raziskavo o glavnih ovirah pri statističnem strojnem prevajanju morfološko bogatih južnoslovanskih jezikov (Arčan in Popović 2015). Na druge slovanske jezike so se med drugim osredotočile raziskava o večjezikovnem statističnem prevajanju sorodnih južnoslovanskih jezikov (Popović in Ljubešić 2014), raziskava o ponovni oceni vpliva tehnik statističnega strojnega prevajanja s človeško evaluacijo na angleško-hrvaški kombinaciji (Toral in dr. 2016) ter raziskava o spopadanju s pomanjkanjem podatkov za hrvaščino s faktorskimi modeli in morfološko razširitvijo (Sanchez-Cartagena in dr. 2016).

4 PRIPRAVA KORPUSA IN PREVAJALNI MODELI

Pri statističnem strojnem prevajanju je zelo pomembna količina podatkov, ki jih imamo na voljo. Običajno velja – več primernega besedila imamo, boljši bodo strojni prevodi. Naš korpus je vseboval 30 zbirk prevodov s področja informacijskih tehnologij, ki so plod desetletnega prevajanja ene od prevajalskih agencij (običajno ima agencija različno zbirko prevodov za vsako stranko). Podrobnejša statistika korpusa je navedena v Tabeli 1.

	Angleški korpus	Slovenski korpus
Št. segmentov	2.395.472	2.395.472
Št. besed	27.480.126	25.411.837
Št. znakov (s presledki)	162.579.039	173.952.193

Tabela 1: Statistika angleškega in slovenskega korpusa.

Če želimo ustvarjati kakovostne prevode, je seveda pomembno, da so prevodi v uporabljenem korpusu kakovostni. Večino prevodov sta v postopku prevajanja pregledali dve osebi (prevajalec in pregledovalec), zato lahko s precejšnjo verjetnostjo trdimo, da so podatki, ki smo jih uporabili za učenje modelov, kakovostni.

Da bi analizirali zmožnosti statističnega strojnega prevajanja v sistemu Moses in naredili primerjavo med običajnimi ter faktorskimi modeli statističnega strojnega prevajanja, bomo izvedli več poskusov na različnih prevajalnih modelih. Poskuse bomo izvajali na dveh sorodnih besedilih. Prvo besedilo je kompleksnejše in vsebuje nekaj prenesenih pomenov, s katerimi imajo strojni prevajalniki običajno težave, v drugem pa so stavki preprostejši, vključili pa smo tudi nekaj tehničnih lastnosti ene od opisanih naprav. Sklepamo, da bo prevajalnik bolje prevedel drugo besedilo, pri prvem pa se bo bolj poznal vpliv faktorskega prevajanja, saj je vrstni red besedila pomembnejši.

Izdelali smo spodaj navedene modele strojnega prevajanja. Faktorske modele smo označili, kot je to definirano v raziskavi o taksonomiji za faktorske modele za strojno prevajanje po besednih zvezah (Bojar in dr. 2012). V našem primeru smo z *F* označili površinsko obliko besede, z *L* lemo, s *P* besedno vrsto in s črko *T* oblikoskladenjske informacije.

1. Osnovni nefaktorski model: začeli bomo z osnovnim nefaktorskim modelom, ki bo vseboval samo privzeto Mosesovo prerazporejanje besed glede na razdaljo. Pri tem modelu je premik besede za dve mesti dvakrat dražji kot premik za eno mesto.

2. Nefaktorski model s prerazporejanjem: prvi nefaktorski model bomo nadgradili z leksikalnim prerazporejanjem prevoda, pri katerem upoštevamo, da se nekatere besedne zveze prerazporedijo pogosteje kot druge. Ta bo omogočal različne usmerjenosti besednih zvez – usmerjenost prevedene besedne zveze je lahko monotona, zamenjana ali prekinjena. V prvem primeru je vrstni red besed v zvezi enak kot v izvirniku, v drugem se besede med seboj zamenjajo, v zadnjem pa se med njih vrinejo druge besede (P. Koehn, 2010). Model bo omogočal prerazporejanje v obe smeri in bo upošteval tako besedilo izvirnika kot prevoda.

3. Osnovni faktorski model (*tF-FaP*): ustvarili bomo faktorski korpus, v katerem bomo površinskim oblikam besed dodali oznake – besedno vrsto v izvirniku ter lemo, besedno vrsto in oblikoskladenjske informacije v prevodu. Pri prvem faktorskem modelu bomo površinsko obliko izvirnika prevedli v površinsko obliko in besedno vrsto.

4. Faktorski model s prerazporejanjem (*tF-FaP, leksikalno prerazporejanje*): pri četrtem modelu nas bo zanimalo, kako leksikalno prerazporejanje vpliva na faktorske modele, zato bomo 3. modelu dodali prerazporejanje po površinski obliki izvirnika in prevoda.

5. Faktorski model s korakom ustvarjanja (*tF-L+gL-FaP, leksikalno prerazporejanje*): ta model bomo zasnovali tako, da bomo koraku prevajanja dodali še korak generiranja novih besed, s katerim bomo poskusili izboljšati kakovost dobljenih prevodov. Poleg tega izvirnih površinskih oblik ne bomo prevajali v površinsko obliko, pač pa v lemo. Angleščina v nasprotju s slovenščino ni morfološko bogat jezik, zato so besede v imenovalniku pogostejše, te pa je bolj smiselno prevajati v osnovno obliko slovenskih besed – leme. Prevedene besede bomo v koraku ustvarjanja povezali s površinsko obliko in besedno vrsto, da bi tako pridobili boljše prevode. Še vedno bomo uporabljali enak model leksikalnega prerazporejanja, le da bomo ta postopek izvajali glede na besedno vrsto izvirnika in prevoda.

6. Faktorski model z dvema korakoma prevajanja (*tF-L+gL-F+tP-P, leksikalno prerazporejanje*): kot zadnjega bomo ustvarili najkompleksnejši model, v katerem bomo ustvarili dve tabeli prevodov. To pomeni, da bomo imeli poleg koraka ustvarjanja in leksikalnega prerazporejanja dva koraka prevajanja. V prvem koraku bomo površinske oblike besed prevedli v leme, v koraku ustvarjanja leme povezali s površinskimi oblikami prevedenega besedila, nato pa besedne vrste izvirnika prevedli v besedne vrste prevoda. S tem želimo doseči višjo natančnost prevodov.

7. Google Translate: vse naše modele bomo primerjali s prevodom, ki smo ga ustvarili s storitvijo Google Translate. Ta omogoča brezplačno prevajanje krajših besedil. Naši besedili smo vnesli v storitev, ki nam je vrnila prevod. Tega smo razdelili na osnovne enote, kar je priporočeno pred uporabo Mosesove skripte za ocenjevanje po metriki BLEU.

Vse modele bomo primerjali med seboj ter z dvema neodvisnima prevodoma človeških prevajalcev. Za vse primerjane pare bomo izračunali podobnost BLEU, podmnožico posameznih prevedenih stavkov vsakega besedila pa bo z ocenami od 1 do 5 ocenilo tudi pet neodvisnih pregledovalcev, za katere bomo izračunali ujemanje. Tako želimo na dva načina, objektivnega in subjektivnega, oceniti prednosti in slabosti možnosti, ki jih ponujajo različni prevajalni modeli sistema Moses. Z metriko BLEU se bomo osredotočili na splošno oceno celotnega prevoda, s človeškim pregledom pa bomo podrobneje pogledali vsak stavek posebej.

5 OCENJEVANJE IN ANALIZA PREVODOV

Najprej predstavimo besedili, ki smo ju prevajali, in podamo rezultate ocenjevanje kakovosti prevodov z mero BLEU in s človeškimi ocenjevalci. Razdelek zaključimo z analizo ujemanja med ocenjevalci.

5.1 Analiza besedil

Za prevod smo izbrali dve precej različni besedili, ki pa imata skupno lastnost –

vsako se na svoj način dotika področja informacijskih tehnologij. Prvo besedilo svetuje glede dobrih tehnik fotografiranja, v drugem delu pa predstavi priznanega fotografa, ki opisuje, kako je fotografiral Amsterdam. Besedilo je napisano v tržnem slogu, kar pomeni, da so stavki povezani, tekoči in da so v njih preneseni pomeni, na primer *There is something truly magical about getting up close and personal with a subject*, kar na drugačen način pove, da se fotograf subjektu približa. S takimi strukturami imajo strojni prevajalniki običajno težave, zato nas je zanimalo, kako jih bo prevedel naš model. V besedilu je tudi nekaj za nas nenavadnih nizozemskih imen, na primer *The iconic 17th century canals of Herengracht, Prinsengracht, and Keizersgracht stand out from the Centraal station in an almost semi-hexagonal shape*, ki lahko prav tako predstavljajo težavo za prevajalnik. Zaradi tega pričakujemo, da bo naš model za prevod tega besedila prejel nižjo oceno, da pa bo opazno boljši prehod na faktorski model, saj je pri tekočem besedilu vrstni red besed še opaznejši kot pri tehničnih besedilih z več naštevanja.

Drugo besedilo je tipičen primer besedila, ki ga proizvajalci elektronskih naprav objavljajo na spletnih straneh. Sestavljeno je iz treh kratkih opisov večnamenskih naprav za optično branje, kopiranje, tiskanje in pošiljanje faksov. Poleg tega so v besedilu tudi tehnične lastnosti ene od naprav. Slog takega besedila je nekoliko bolj tog, stavki so krajši in bolj preprosti, v besedilu pa je več tehničnih izrazov, imen tehnologij, številk in podobno. Dobra primera stavkov v takih besedilih sta *Wi-Fi and Wi-Fi Direct® connectivity ensure the flexibility every small business needs* in *20 Pages/min Colour (plain paper 75 g/m²), 33 Pages/min Monochrome (plain paper 75 g/m²)* (slednji je element seznama tehničnih lastnosti ene od naprav). Pričakujemo, da bo prevod takega besedila ocenjen bolje, vendar pa prehod na faktorsko prevajanje zaradi relativno preprostih stavčnih struktur ne bo tako občuten.

Odločili smo se za dolžino besedila, ki jo prevajalec v povprečju prevede v enem dnevu. Prvo besedilo je dolgo 2402 besedi, drugo pa 1437 besed.

5.2 Ocenjevanje besedil

Moses ima za ocenjevanje po metriki BLEU namensko skripto multi-bleu, ki referenčno besedilo primerja z izbranim. Vse strojne prevode bomo primerjali z dvema neodvisnima človeškima prevodoma, zato nas je najprej zanimalo, kakšna je podobnost med obema prevodoma. Pri primerjavi obeh različic prvega besedila smo dobili rezultat 37,00, pri primerjavi drugega pa 46,90. Kot smo pričakovali, je drugi rezultat višji, saj so v prvem besedilu stavki kompleksnejši in jih je mogoče tolmačiti na več načinov, to pa pomeni, da je večja verjetnost, da ga bodo različni prevajalci prevedli drugače. Tehnična besedila morajo biti bolj natančna in pri njih je manj prevajalske svobode, zato je večja verjetnost, da bodo prevedena podobno. Kljub temu rezultata nazorno kažeta, kako različni so si človeški prevodi. Tudi pri precej nedvoumnem drugem besedilu obseg podobnosti ni presegel 50 % ujemanja, zato se je odločitev, da vključimo tudi človeški pregled strojnih prevodov izkazala za smiselno.

Pri drugem načinu ocenjevanja smo del besedila dali v pregled petim neodvisnim človeškim pregledovalcem. Prosili smo jih, da z oceno od 1 do 5 ocenijo vsak stavek posebej in napišejo komentar za posamezen model. Vsaka oseba je pregledala 317 besed prvega in 301 besedo drugega besedila. Za zanesljivejše rezultate bi morali vzeti večje vzorce, vendar nam časovni in finančni obseg raziskave tega ni dovoljeval. Za vse modele smo izračunali povprečno oceno in jo primerjali z rezultati BLEU, ki smo jih dobili v prejšnjem poglavju. Da bi zagotovili čim bolj enoten kriterij, smo vsem pregledovalcem posredovali enake kriterije za ocenjevanje. Odločili smo se za sistem ocen, ki je opisan spodaj in se zgleduje po sistemu agencije ARPA (White in O'Connell 1994). Poleg takega ocenjevanja je mogoče strojni prevod oceniti tudi glede na ustreznost, berljivost, primerjavo z drugimi prevodi istega besedila in z analizo napak (www.ebaytechblog.com, dostop januar 2017). Po končanem pregledu smo s povprečjem vseh parov Cohenovega koeficienta kapa in s Fleissovim koeficientom kapa preverili ujemanje ocen.

Pregledovalci so prevode ocenili tako:

5 - Stavek je popolnoma jasen in razumljiv. Prevod ni nujno popoln, vendar je slovnično pravilen, vse informacije pa so posredovane natančno.

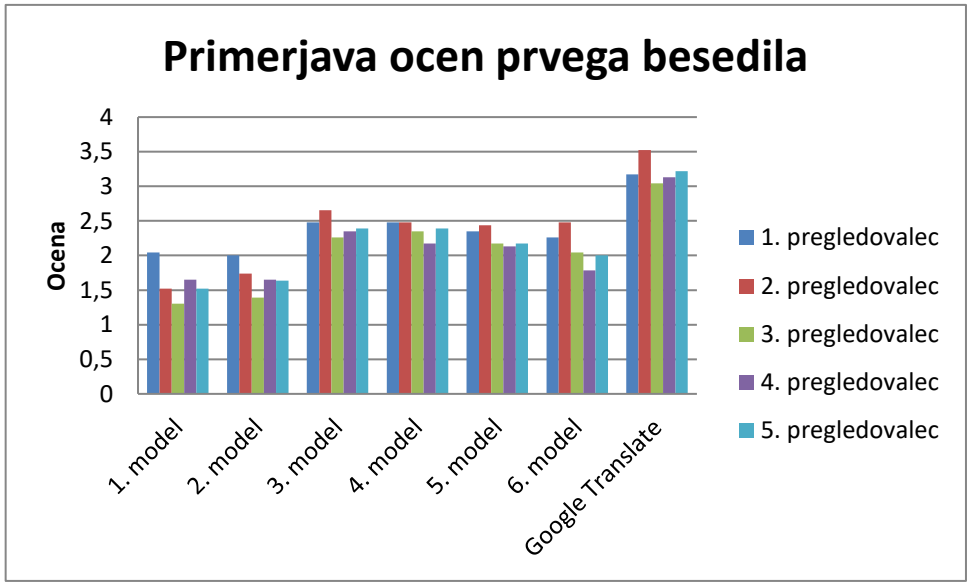
4 - Stavek je v splošnem jasen in razumljiv. Ni popoln, je pa sprejemljiv, mogoče ga je razumeti in zajame večino pomena izvirnika.

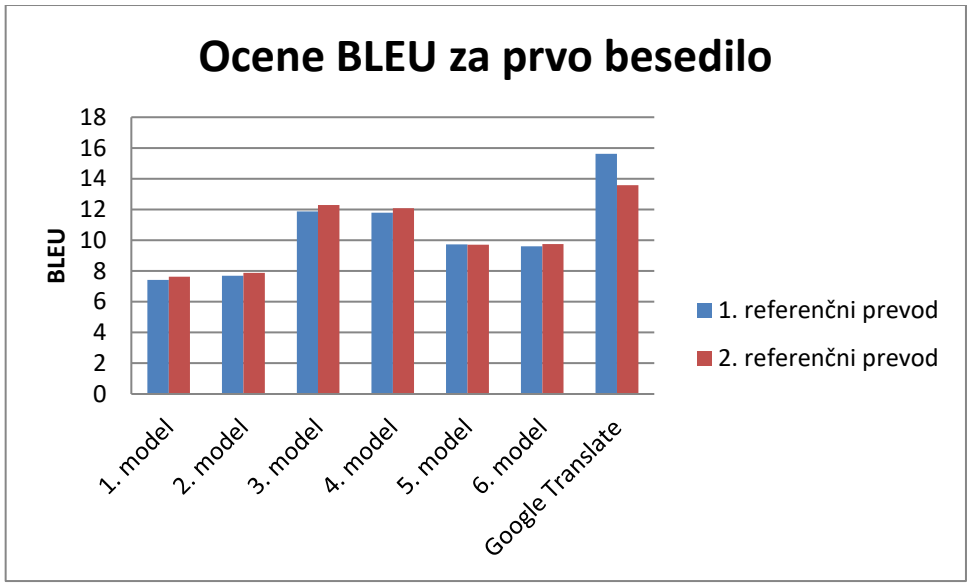
3 - V stavku so slovnične napake in/ali napačno izbrane besede. Z nekaj truda je mogoče razbrati del, ne pa vsega, pomena izvirnika.

2 - V stavku je nekaj ključnih besed, vendar je v prevodu izraženega malo pomena izvirnika.

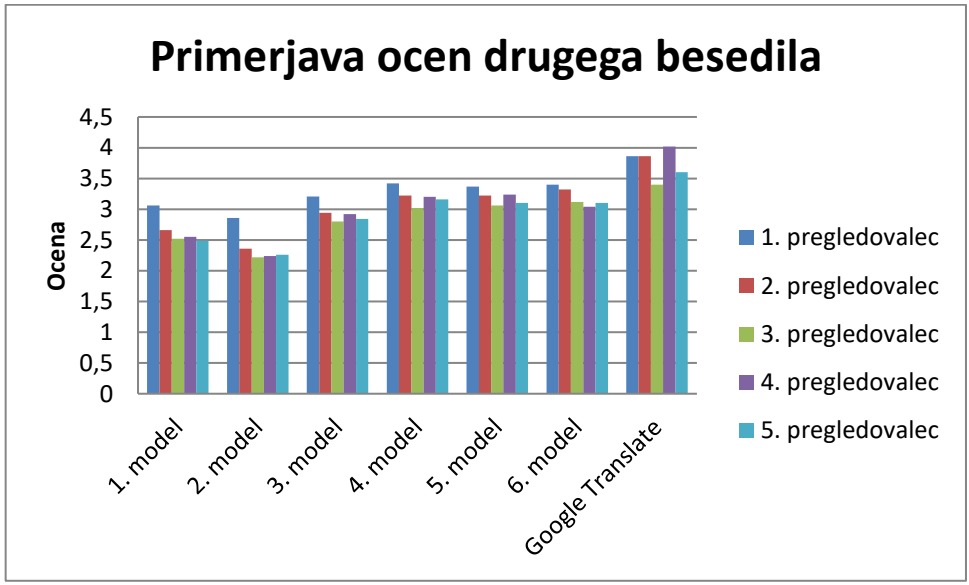
1 - Prevod je nesprejemljiv. Stavek je nerazumljiv, natančno je posredovano malo ali nič informacij. Prevod ne odraža pomena izvirnika.

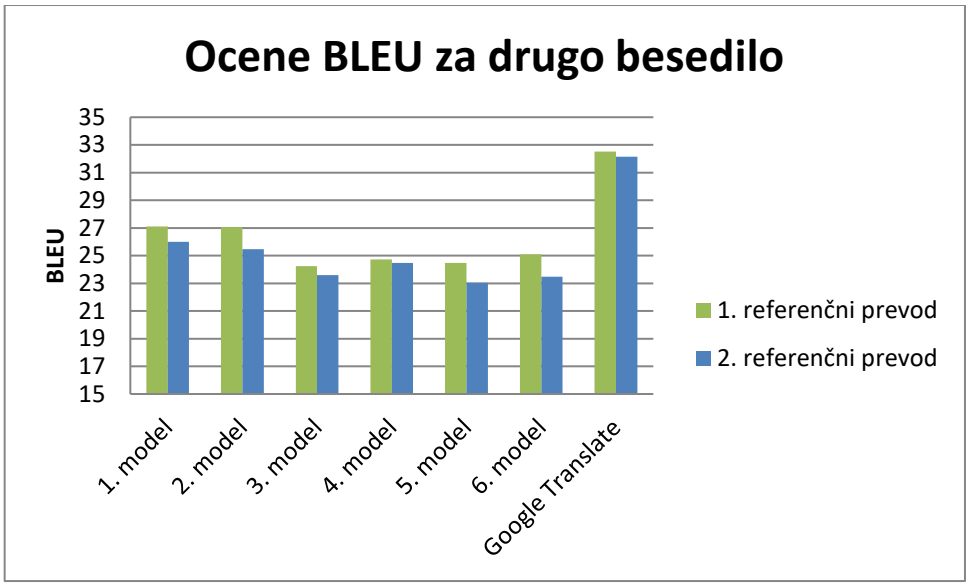
Če grafa z ocenami, pridobljenimi z algoritmom BLEU, primerjamo z ocenami pregledovalcev, dobimo spodnji sliki:





Graf 1: Primerjava ocen človeških pregledovalcev in ocen BLEU za prvo besedilo.





Graf 2: Primerjava ocen človeških pregledovalcev in ocen BLEU za drugo besedilo.

Čeprav rezultatov ne moremo primerjati neposredno zaradi različnih metrik, na grafih 1 in 2 jasno vidimo, da se gibanje ocen kakovosti pri obeh besedilih dobro ujema. Pri prvem besedilu so rezultati z nefaktorskimi modeli nizki, pri prehodu na faktorske modele se občutno izboljšajo, pri različnih faktorskih modelih pa nekoliko nihajo. Pri drugem besedilu so rezultati že z nefaktorskimi modeli boljši, vendar pa pri prehodu na faktorske modele pride do edine občutne razlike med računalniško in človeško oceno – ocena BLEU je za faktorske modele nižja, človeški pregledovalci pa so jih ocenili bolje kot nefaktorske.

5.3 Ujemanje ocenjevalcev

Po pregledu vseh rezultatov smo preverili še stopnjo strinjanja med ocenjevalci. Ta ocenjuje, koliko homogenosti ali strinjanja je v ocenah ocenjevalcev. Uporabna je pri izboljševanju orodij, danih človeškim ocenjevalcem. Če se različni ocenjevalci ne strinjajo, je metrika slaba ali pa je treba ocenjevalce bolje podučiti. Za ugotavljanje stopnje strinjanja je na voljo več statistik, ki so odvisne

od vrste meritev. V našem primeru smo uporabili Cohenov koeficient kapa (Cohen 1968) in Fleissov koeficient kapa (Fleiss 1971).

Cohenov koeficient kapa je statistična mera, ki določa strinjanje med ocenjevalci za kvalitativne elemente. Meri strinjanje med dvema ocenjevalcema, ki N elementov razvrstita v C medsebojno izključujočih se kategorij. Če se ocenjevalca popolnoma strinjata, potem velja $\kappa = 1$. Če med ocenjevalcema ni drugega strinjanja kot naključno, velja $\kappa \leq 0$. Kapo smo najprej izračunali za vsak par pregledovalcev vseh modelov, nato pa smo določili povprečni koeficient najprej za vsak model, nato pa še za vse modele skupaj. Rezultate smo zbrali v Tabeli 2:

	1. model	**2. model**	**3. model**	**4. model**	**5. model**	**6. model**	**7. model**
Povprečni Cohenov koeficient kapa	0,372	0,3544	0,4127	0,3893	0,4343	0,3707	0,4122
			Povprečni Cohenov koeficient kapa vseh modelov:				**0,3922**

Tabela 2: Povprečni Cohenov koeficient kapa za vse modele.

V Tabeli 2 lahko vidimo, da so si bili pregledovalci najbolj enotni pri 5. modelu, najmanj pa pri 2. modelu. Po lestvici, ki sta jo predlagala Landis in Koch (Landis in Koch 1977), je naš povprečni Cohenov koeficient kapa ravno na meji med dokajšnjim in zmernim strinjanjem, kar je glede na subjektivno naravo človeškega pregledovanja sprejemljiv rezultat.

Cohenov koeficient kapa meri le strinjanje med dvema ocenjevalcema. Za več ocenjevalcev uporabimo Fleissov koeficient kapa, ki je statistična mera za določanje zanesljivosti strinjanja med nespremenljivim številom ocenjevalcev pri dodeljevanju kategoričnih ocen množici elementov ali elementov za razvrščanje. Rezultate prikazuje Tabela 3.

	1. model	2. model	3. model	4. model	5. model	6. model	7. model
Povprečni Fleissov koeficient kapa	0,3613	0,3598	0,4222	0,391	0,427	0,3682	0,4199
	Povprečni Fleissov koeficient kapa vseh modelov:						**0,3928**

Tabela 3: Povprečni Fleissov koeficient kapa za vse modele.

S Tabele 3 lahko razberemo, da sta oba koeficienta ujemanja podobna. Tudi glede na Fleissov koeficient kapa so se pregledovalci najbolj strinjali pri 5. modelu, najmanj pa pri 2. modelu. Fleissov povprečni koeficient je zelo podoben Cohenovemu in je na meji med dokajšnjim ter zmernim strinjanjem.

6 SKLEPNE UGOTOVITVE IN PREDLOGI ZA NADALJNJE DELO

V raziskavi smo poskušali odgovoriti na dve glavni vprašanji: je strojno prevajanje že primerno za splošno rabo in ali faktorski modeli izboljšajo kakovost strojnega prevajanja v slovenščino? Strojni prevajalni modeli se dobro obnesejo pri prevodih tehnične narave, to so besedila, kjer so stavki relativno preprosti, poleg tega pa niso dvoumni in vsebujejo malo prenesenih pomenov. Pregledovalci so prevode takih stavkov pri večini modelov ocenili s 4 ali 5, kar pomeni, da pomena takega stavka ni težko razumeti. Drugačno sliko dobimo pri besedilih, ki so bolj zapletena. Strojni prevajalniki imajo z daljšimi in kompleksnimi stavki še vedno veliko težav. To ne velja samo za modele, ki smo jih ustvarili sami, pač pa tudi za Google Translate, ki se ponaša z veliko večjim korpusom kot naši modeli. To je dobro vidno ne samo iz subjektivnega pregleda, pač pa tudi iz rezultatov BLEU, ki so v našem primeru pri bolj zapletenem besedilu nižji tudi za več kot polovico. Trdimo lahko, da strojno prevajanje dobro služi namenu prenosa približnega pomena v drug jezik, ni pa še primerno za visokokakovostne prevode brez človeškega pregleda (izjema so seveda močno specializirana besedila, kjer je besedišče veliko ožje).

Modeli, ki smo jih ustvarili, za splošno uporabo niso primerni, saj je korpus, ki smo ga ustvarili, premajhen, poleg tega pa je usmerjen preozko, saj je večinoma sestavljen iz besedil, ki se dotikajo področja informacijskih tehnologij. Koristen bi bil pri prevajanju besedil iz omenjenega področja, torej z informacijsko tehnologijo povezanih spletnih mest, navodil za uporabo in programske opreme.

Kako na strojne prevode vplivajo faktorski modeli? Pri tehničnem besedilu uvedba faktorskega modela ni imela občutnega vpliva. Rezultat BLEU je bil celo nekoliko nižji kot tisti z nefaktorskimi modeli, človeška ocena faktorskih strojnih prevodov je sicer nekoliko višja, vendar ne dovolj, da bi upravičila čas in sredstva, ki jih je treba vložiti v ustvarjanje faktorskih korpusov ter modelov. Drugačne rezultate smo dobili pri prevajanju kompleksnejšega besedila. Tam sta računalniška in človeška ocena pri uvedbi faktorskega modela močno poskočili in se povečali skoraj za polovico. V najboljšem primeru se je storitvi Google Translate ocena BLEU približala na 12 odstotkov, kar je glede na neprimerno manjši korpus odličen rezultat in velika spodbuda za nadaljnje delo s faktorskimi modeli. Tako je na primer ocena prevoda niza *Street in Amsterdam* pri enem od pregledovalcev z ocene 1 zrasla na 5 (iz *Street v Amsterdam* v *Ulica v Amsterdamu*) in ocena prevoda *They say nothing lasts forever* z 1 na 3 (iz *They izgovorite nič zdrži forever* v *Te izgovorite nič ne traja za vedno*). Tudi v svojih komentarjih so pregledovalci zapisali, da je prevod prvega besedila z uvedbo faktorskih modelov postal precej bolj uporaben. Vseeno moramo biti pri uporabi faktorskih modelov nekoliko previdni. Najprej seveda glede besedila, ki ga želimo prevajati, pa tudi glede na vrsto modela, ki ga izberemo. Kot smo pokazali, bolj kompleksen model s prerazporejanjem in več koraki prevajanja in ustvarjanja ne pomeni boljših prevodov. Izbrati moramo torej pravi model za svoj namen, kar pa ni vedno preprosto.

Zelo zanimiva je tudi primerjava med ocenami človeških pregledovalcev in avtomatično pridobljenimi ocenami po metriki BLEU. Čeprav rezultatov ne moremo primerjati neposredno zaradi različnih metrik, se gibanje ocen

kakovosti pri obeh besedilih dobro ujema. Edina občutna razlika med računalniško in človeško oceno pride pri prehodu na faktorske modele pri drugem besedilu, kar nam potrjuje, da je metrika BLEU primerna za ocenjevanje strojnih prevodov.

Kako bi lahko naše modele izboljšali? Prvi in najbolj očiten način je uporaba večjega korpusa. Statistično strojno prevajanje je odvisno od podatkov, ki jih ima na voljo, zato bi z obsežnejšimi korpusi modeli delovali bolje. Korpus, s katerim smo delali, je imel 2,4 milijona vrstic, kar je za to vrsto prevajanja relativno malo. Prevodi opisov tiskalnikov so bili kljub temu dobri, ker je bilo besedilo podobno tistemu iz korpusa. Če bi želeli prevajati splošna besedila, bi potrebovali bistveno večji korpus. Drugi način bi bil uporaba drugačnega nabora za prilagajanje. Naš je imel 200 vrstic, če pa bi uporabili takega z nekaj tisoč vrsticami, bi bili prevodi nekoliko boljši, vendar bi to imelo manjši vpliv kot večji korpus ustreznih prevodov. Že v našem primeru smo za prilagajanje kompleksnih faktorskih modelov potrebovali 10-12 ur na zmogljivem strežniku, v primeru večjega nabora za prilagajanje bi bil ta čas temu ustrezno daljši. Poleg velikosti nabora je pomembna tudi njegova vsebina, ki mora biti čim bolj podobna besedilu, ki ga prevajamo. Naš nabor je bil zelo podoben prevajanemu tehničnemu, manj pa tržnemu besedilu. Če bi za vsako besedilo uporabili drugačen nabor, bi bili rezultati nekoliko boljši, ker pa je bil naš glavni namen primerjati modele med seboj, nam je omenjeni nabor za prilagajanje zadostoval.

Primerjali smo le majhno število modelov, ki jih omogoča Moses. Želeli smo namreč narediti prerez možnosti in postopno dodajati funkcije ter tako ugotavljati, kako vplivajo na kakovost prevodov. Na ta način smo lahko preizkusili le omejen nabor možnosti in parametrov, ki jih omogoča Moses – kombinacij jezikovnih modelov, prerazporejanja, izbranih faktorjev, korakov prevajanja in ustvarjanja ter prevajalnih modelov, ki bi jih lahko preizkusili, je namreč veliko. Izkazalo se je, da vsi modeli niso primerni za prevajanje, saj je treba paziti, da so koraki prevajanja in ustvarjanja med seboj neodvisni ter da

uporabljajo ustrezne jezikovne modele. Tako smo pri nekaterih modelih dobili neprevedeno izhodno besedilo ali pa je bilo v prevodu nenavadno veliko število ločil, na primer v stavku *, ko je executed with care abstract photography) , , , , ki jih je mogoče hugely creative . . . . . .*

Videti je tudi, da bi bila dobra rešitev uporabiti hibridni model, kjer bi združili faktorsko statistično strojno prevajanje in strojno prevajanje, ki temelji na pravilih. Če pogledamo prevod stavka *To be able to really give an honest account of a city you need to fully understand its make up*, vidimo, da je preveden kot *, da boste lahko zares zagotavlja pristno račun mesta, kar potrebujete, če želite povsem razumemo, da je njegova navzgor*. V dvojezičnem korpusu je tak primer razbit na dva dela, zato se prevod začne z vejico. Če bi imeli v modelu pravilo, ki bi določalo, da je to začetek stavka, takih težav ne bi imeli. Z dodajanjem pravil bi odpravili tudi težave z nizi, v katerih se pojavljajo oklepaji (besedilo v oklepajih v našem primeru ni prevedeno), in s števili, kjer moramo vejice za tisočicami v angleščini zamenjati s pikami v slovenščini (naši modeli so pustili vejice).

Pregled naših prevodov je opravilo 5 jezikoslovcev, saj smo želeli tudi mnenje o slovnični ustreznosti. Vseeno bi lahko pregled dopolnili tako, da bi ga opravile tudi osebe brez jezikoslovnega ozadja, ki jim je strojni prevod pogosteje namenjen, saj velja, da manj jezikovnega znanja, kot imajo uporabniki, bolj uporabno je strojno prevajanje (Hutchins 2009).

LITERATURA

A. L. F. Han, D. F. Wong, L. S. Chao (2012): LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors, *Proceedings of COLING 2012: Posters*.

A. Ratnaparkhi (1996): A maximum entropy model for part-of-speech tagging, Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Human Evaluation of Machine Translation. Dostopno prek: http://www.ebaytechblog.com/2016/06/26/human-evaluation-of-machine-translation (januar 2017)

J. L. Fleiss (1971): Measuring nominal scale agreement among many raters, *Psychological Bulletin, 76*, 378–387.

J. Cohen (1968): Nominal scale agreement provision for scaled disagreement or partial credit, *Psychological Bulletin*, *70*, 213–220.

J. R. Landis, G. G. Koch (1977): The Measurement of Observer Agreement for Categorical Data, *Biometrics, 33*, 159–174.

J.S. White in T. O'Connell (1994). The ARPA MT evaluation methodologies: evolution, lessons, and future approaches, *Proceedings of the First Conference of the Association for Machine Translation in the Americas,* Columbia, Maryland.

K. Papineni, S. Roukos, T. Ward in W. Zhu (2002): BLEU: a Method for Automatic Evaluation of Machine Translation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics.

M. Arčan, M. Popović in P. Buitelaar (2016): Asistent – a machine translation system for Slovene, Serbian and Croatian, *Proceedings of the 10th Conference on Language Technologies and Digital Humanities*.

M. Grčar, S. Krek in K. Dobrovoljc (2012): Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik, *Zbornik osme konference Jezikovne tehnologije*.

M. Popović in M. Arčan (2015): Identifying main obstacles for statistical machine translation of morphologically rich south slavic languages, *18th Annual Conference of the European Association for Machine Translation (EAMT)*.

M. Popović in Nikola Ljubešić (2014): Exploring cross-language statistical machine translation for closely related South Slavic languages, *Proceedings of the EMNLP14 Workshop on Language Technology for Closely Related Languages and Language Variants*.

Moses – Moses/FactoredModels. Dostopno prek: http://www.statmt.org/moses/?n=Moses.FactoredModels (maj 2016).

O. Bojar, B. Jawaid in A. Kamran (2012): Probes in a taxonomy of factored phrase-based models, *Proceedings of the Seventh Workshop on Statistical Machine Translation*.

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

P. Koehn, F. J. Och, D. Marcu (2003): Statistical Phrase-Based Translation, Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, University of Southern California.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst (2007): Moses: open source toolkit for statistical machine translation. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*.

R. R. Antonio Toral in G. Ramirez Sanchez (2016): Re-assessing the Impact of SMT Techniques with Human Evaluation: a Case Study on English-Croatian. *In Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT)*.

Satanjeev Banerjee, Alon Lavie (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaulation Measures for MT and/or Summarization*, Association for

Computational Linguistics.

Š. Vintar, Računalniške tehnologije za prevajanje, *Uporabna Informatika*, *VII/1*, 17–24.

T. Erjavec, S. Krek, Š. Arhar, D. Fišer, N. Ledinek, A. Saksida, B. Sivec, B. Trebar (2008): Oblikoskladenjske specifikacije, *Zbornik Šeste konference Jezikovne tehnologije*, Ljubljana.

V. M. Sánchez-Cartagena, N. Ljubešić in F. Klubička (2016): Dealing with data sparseness in SMT with factored models and morphological expansion: a Case Study on Croatian, *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT)*.

W. Hutchins (2009): Uses and Applications of Machine Translation, *Presentation at Westminster University*.

COMPARING STANDARD AND FACTORED MODELS IN STATISTICAL MACHINE TRANSLATION FROM ENGLISH TO SLOVENE USING THE MOSES SYSTEM

Machine translation is a field in computational linguistics that explores the use of software to translate text from one language to another. Factored statistical translation is an extension of statistical machine translation, where linguistic annotation is added on the word level. Words are turned into vectors in an attempt to improve translation quality. We describe the use of the open-source Moses system for factored statistical machine translation from English to Slovenian. We created several factored and non-factored language and translation models from a text corpus, containing IT-related texts. We translated two different IT-related documents. The first one was marketing-orientated with a complex structure, while the second one was technical with a simpler structure. We used two methods to compare the generated translations with two independent human translations and a translation, created by the Google Translate service. The first comparison method was the BLEU metrics and the second one were evaluations of human reviewers. The latter method expressed a subjective score, which is still very important in the machine translation field. Even though the results can't be compared directly due to different metrics, the movement of the grades is well correlated for both texts. The only bigger difference can be seen while implementing factored models for translating the second text. In the conclusion we analysed the inter-evaluator coherence and the obtained results. We discovered that our models are more suitable for technical texts, and that factored models improve the translation of complex texts more.

Keywords: statistical machine translation, factored machine translation, Moses system, BLEU, human evaluation