Metode hitre izdelave gradiv za prevajalne sisteme plitkega prenosa za visoko pregibne jezike

Jernej Vičič
University of Primorska
Glagoljaška 8,
SI-6000 Koper
jernej.vicic@upr.si

Povzetek

Članek predstavlja pregled čez zbirko izbranih ter novo zasnovanih metod samodejne izdelave gradiv za postavitev prevajalnih sistemov na osnovi pravil, Rule Based Machine Translation (RBMT), še posebej RBMT plitkega prenosa, ki so primerne za postavitev prevajalnih sistemov jezikovnih parov sorodnih jezikov.

Metode so bile preizkušene na študiji primera: "postavitev popolnoma delujočega prevajalnega sistema za sorodne jezike za jezikovni par slovenščina - srbščina".
Posebne lastnosti tega jezikovnega para so visoka pregibnost in bogata morfologija, ki povzročata slabše rezultate nekaterih že razvitih metod.
Slovenščina in srbščina pripadata skupini južno slovanskih jezikov in sta bila uporabljana predvsem na območju nekdanje Jugoslavije. Ekonomije držav, kjer se govorita ta dva jezika so tesno povezane, mlajše generacije pa imajo težave v medsebojnem sporazumevanju, tako obstaja dovolj razlogov za postavitev takšnega prevajalnega sistema.

Sistem temelji na Apertiumu, ki je odrptokodno ogrodje za postavitev prevajalnih sistemov tipa RBMT plitkega prenosa. Prikazano je vrednotenje opisanih tehnik ter vrednotenje kakovosti prevodov celotnega sistema.

1 Uvod

Sorodnost naravnih jezikov neke tipološke skupine poenostavlja poenostavitev naloge prevajanja in tako omogoča uporabo enostavnejših tehnik, ki ne bi bile primerne za uporabo pri prevajanju nesorodnih naravnih jezikov. Uporaba enostavnih metod pa ne pomeni nujno slabše kakovosti prevodov, veliko napak pri prevajanju izvira ravno iz napak naprednejših tehnik kot je popolna analiza povedi. V sistemih klasične arhitekture strojnega prevajanja na osnovi prenosa, transfer based MT systems architecture, povzroča kopičenje napak analize, prenosa in sinteze zmanjšanje kakovosti prevajanja v primerjavi z izboljšavami, ki jih ta arhitektura ponuja.

Avtorji najbolj znanih prevajalnih sistemov sorodnih jezikov, [10], [23], [20], predlagajo plitko arhitekturo kot je prikazana na sliki 1.

Slovenščina in srbščina pripadata skupini južno slovanskih jezikov in sta bila uporabljana predvsem na območju nekdanje Jugoslavije. Slovenščina se danes uporablja največ v Sloveniji, srbščina pa v Srbiji ter Črni gori. Jezika si delita skupen izvor in še pomembneje delita si skupno nedavno zgodovinsko okolje, ta dva jezika smo uporabljali v isti državi, učili smo se ju v šolah kot jezika okolja.

Razlogov za postavitev prevajalnega sistema za opisan jezikovni par je dovolj, naštejmo

le najpomembnejše: ekonomije treh novih držav so še vedno povezane; mlajše generacije, generacije po razpadu Jugoslavije, imajo težave pri medsebojni komunikaciji.

Oba jezika sta visoko pregibna in morfološko ter derivacijsko bogata, razlikujeta se od najbolj uporabljanih jezikov v elektronskih medijih kot so angleščina, arabščina ali francoščina. To pomeni, da je potrebno večino podatkov ter prevajalnih metod vsaj pregledati, v najslabšem primeru pa tudi popraviti oziroma ustvariti nove. jezika tega jezikovnega para sta si sorodna tako na leksikalnem kot na sintaktičnem nivoju kar omogoča poenostavitev večine faz izdelave novega prevajalnega sistema.

Izdelava prevajalnega sistema za nov jezikovni par lahko v grobem opišemo na dva načina:

- Dolgotrajna in draga ročna izdelava slovarjev ter prevajalnih pravil v primeru klasičnega pristopa k gradnji prevajalnih sistemov na osnovi pravil, Rule-Based Machine Translation (RBMT) [22], oziroma vseh sorodnih paradigem.
- Samodejna nenadzorovana gradnja prevajalnega sistema na osnovi korpusnih podatkov. Najpomembnejši predstavniki te paradigme so statistično strojno prevajanje, (SMT) [7], [34], strojno prevajanje na osnovi primerov, Example-Based Machine Translation (EBMT) [30], [22]. Obstajajo še drugi pristopi vendar njihov opis presega namen članka .

Za hitro postavitev prevajalnega sistema, tudi za sorodne jezike, se zdi SMT kot idealna izbira saj nekateri prevajalni sistemi z največjo kakovostjo prevodov temeljijo na tehnologijah SMT [32], vendar ima nekaj slabih strani, ki jih ne moremo prezreti. Prevajalni sistemi, ki temeljijo izključno na tehnologijah SMT zahtevajo velike količine dvojezičnih poravnanih besedil [33], ki so dosegljiva samo za nekaj največjih svetovnih jezikov kot so angleščina, arabščina, kitajščina, španščina in francoščina. Morfološko bogat in visoko pregiben jezikovni par, kot je predstavljen v tem članku predstavlja še dodatne probleme, ki so predstavljeni v tabelah 1 ter 3 v razdelku 3.1.

Vse metode in materiali opisani v članku so bili preizkušeni na popolnoma delujočem prevajalnem sistemu GUAT [42], ki temelji na ogrodju Apertium [35], [10]. Apertium je odprtokodno ogrodje prevajalnih sistemov plitkega prenosa. Platforma vsebuje prevajalnik naravnih jezikov, ki ni odvisen od izbire jezikovnega para ter že pripravljene lingvistične podatke za široko paleto jezikovnih parov.

Oglejmo si razdelitev članka: pregled raziskovalnega področja je prikazan v razdelku 2, sledi prikaz uporabljenih metod v razdelku 3. metodologij evalvacije z rezultati je predstavljena v razdelku 4, članek se zaključi z razpravo.

2 Pregled področja

Sodeč po [32], trenutno najboljši samodejno zgrajeni sistemi za prevajanje večinoma sodijo v kategorijo sistemov SMT [7]. Primer takšnih sistemov je [16]. [33] trdi, da takšni sistemi zahtevajo ogromne količine vzporednih učnih podatkov (velikih vzporednih korpusov) za učenje prevajalnih modelov. Sistemi za jezikovne pare z velikimi vzporednimi korpusi dosegajo zavidljive rezultate, za nekatere jezikovne pare so takšni sistemi celo najboljši. Tako veliki vzporedni korpusi ne obstajajo za vse jezikovne pare, še več obstajajo le za majhno množico jezikov. Manjši korpusi, celo lingvistično označeni korpusi, so laže dosegljivi, vsaj za evropske jezike, glej [11], [29].

2.1 Pregled obstoječih sistemov za prevajanje sorodnih naravnih jezikov

Izvedenih je bilo že nekaj eksperimentov v domeni strojnega prevajanja za sorodne jezike, kot na primer Apertium [10] za romanske jezike, vsaj na začetku predvsem v povezavi s španščino ter katalonščino, [41] za keltske jezike, [12], [5], [1] za skandinavske jezike, [17], [18] za slovanske jezike, večinoma v povezavi s češčino ter [2] za turške jezike. Prvi prevajalni sistem, kjer je en od jezikov slovenščina je Guat [42].

2.2 Dostopne tehnologije in materiali

Pregled že obstoječih in dostopnih orodij in gradiv, večinoma korpusov, je pokazal, da je bilo opravljenega že mnogo dela za slovenski jezik, veliko manj za srbski jezik. Orodja za slovenski jezik (zadovoljive oziroma visoke kakovosti) so: morfosintaktični označevalnik, a part of speech tagger, [14], [6], lematizator, a lemmatizer, [14], [13], orodje za krnjenje, a stemmer, [38], [37]. Vseh teh orodij za srbski jezik še ni. Oba jezika imata dobre referenčne korpuse, velikosti več milijonov besed in majhen dvojezični poravnan korpus [11].
Ta raziskava se osredotoča predvsem na leksikalni nivo zaradi naslednjih razlogov:

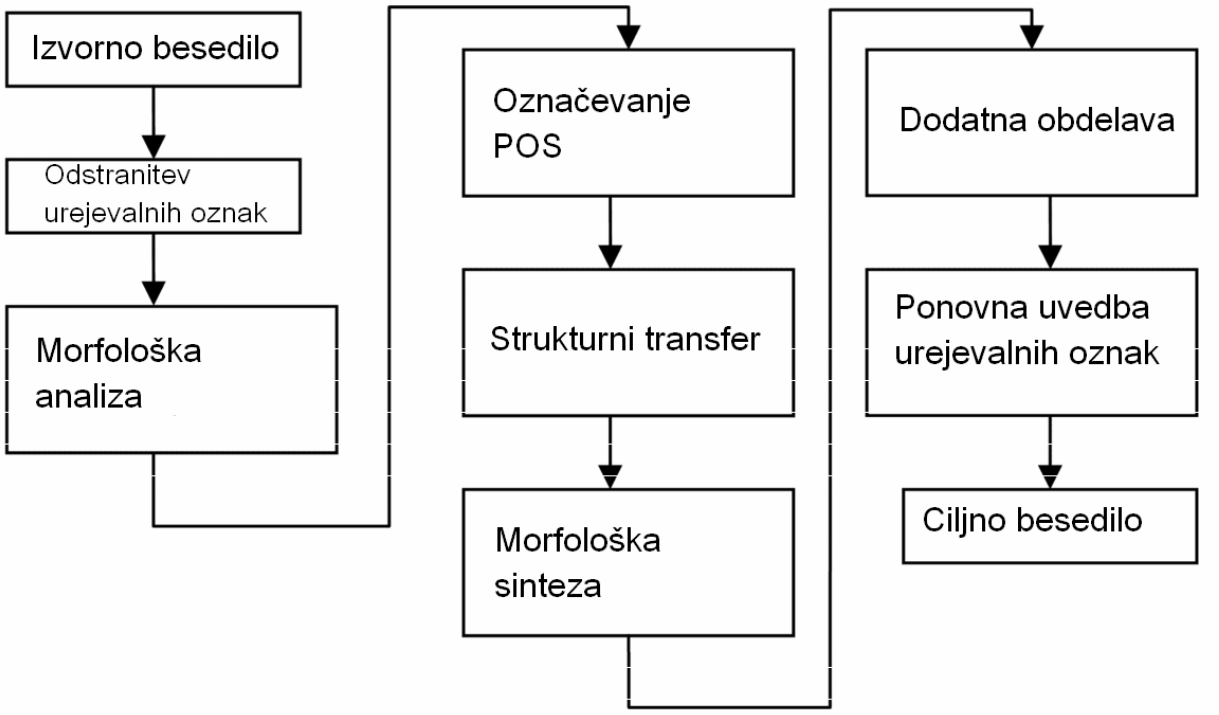
- Leksikalni nivo predstavlja osnovo za prevode pisane besede.
- Sorodni jeziki, posebej jezikovni par, ki smo ga vzeli za osnovo te študije, ponavadi delijo podobno stavčno strukturo. Večina prevoda se dogaja na leksikalnem nivoju.
- Južno-slovanski jeziki izražajo večino pomena s pregibanjem besed in veliko manj z organizacijo besed v povedih. Slednje je veliko bolj izraženo v jezikih kot je angleščina.

Vsi prevajalni moduli so bili upoštevani pri tej raziskavi.

3 Namen

Kar nekaj metod, ki avtomatizirajo postavitev določenih delov sistemov RBMT, je že bilo predstavljenih in so celo uporabljene v orodjih za postavitev prevajalnih sistemov, kot na primer [40]. Ta članek predstavlja poskus avtomatizacije vseh procesov izdelave podatkov za izdelavo sistemov za prevajanje na osnovi plitvega prenosa. Pri poskusih smo uporabljali sistem Apertium [35], sistem za postavitev prevajalnih sistemov na osnovi plitvega prenosa, vendar bi večino metod uporabili pri snovanju večine sistemov te paradigme kot na primer [23], [20]. Podatki predstavljajo osnovo modelov prevajalnega cevovoda kot je prikazan na sliki 1:

1. Enojezični izvorni slovar z morfološkimi opisi, uporabljen za razpoznavo izvornega jezika.
2. Enojezični izvorni slovar z morfološkimi opisi, uporabljen za razpoznavo ciljnega jezika.
3. Dvojezični slovar (prevajanje).
4. Pravila za plitvi prenos.
5. Naučeni modeli označevalnika POS, uporabljeni pri razdvoumljanju.



Slika 1: Moduli tipičnega prevajalnega sistema plitkega prenosa, arhitektura kaže sistem Apertium, sistemi avtorjev [10], [19] sledijo temu vzorcu.

Enojezični slovarji so uporabljeni pri analizi izvornega besedila ter pri sintezi besedila v ciljni jezik. Dvojezični slovar se uporablja pri dejanskem prenosu besed iz izvornega v ciljni jezik, ponavadi prevajamo vsako besedo ločeno, v določenih primerih pa tudi fraze, Ponavadi so dvojezični prevajalni slovarji organizirani s pomočjo osnovnih besednih oblik (lem), ki omogočajo kompaktnejši zapis. Pravila za plitki prenos opisujejo sintaktična ter morfološka pravila kot je lokalno ujemanje besed ter lokalno premeščanje besednega vrstnega reda. Morfološko razdvoumljanje izvornega jezika temelji na implicitnih pravilih, v našem primeru na parametrih Skritega Markovega Modela (HMM) [4], [44] stohastičnega označevalca POS, čeprav so možne tudi alternativne metode, kot je uporaba izbire najboljšega kandidata za končni prevod na osnovi jezikovnega modela ciljnega jezika [20].

Za vsak modul prevajalnega sistema, predstavljen kot posamezna alineja gornjega spiska, smo poskusili poiskati že preizkušeno metodo, ki omogoča samodejno izdelavo potrebnih gradiv oziroma smo takšno metodo sami zasnovali. Metodo so natančneje predstavljene v posameznih razdelkih. Kot končni rezultat smo postavili popolnoma delujoč prevajalni sistem s pomočjo opisanih metod.

3.1 Izdelava enojezičnega slovarjev (izvornega in ciljnega jezika)

Oglejmo si primer prevoda iz angleškega jezika: predelavo besede *walk* (hoditi/hodi/...) v *walked* (hodil/hodila/hodili ...) lahko dosežemo z enostavnim morfološkim pravilom za preteklik (past tense). Posebna različica tega pravila bi podobno spremenila nepravilno besedo *sleep* (spati) v *slept*.

Jezike, ki uporabljajo konkatenativno morfologijo[1], kot je večina evropskih jezikov, tvorijo različne besedne oblike s pomočjo sprememb predpon ter pripon osnovnih besednih oblik. Tako dobimo obliko *slept* iz osnovne oblike (leme) s pomočjo spremembe končnice *-ep* v končnico *-pt*. Enako načelo velja tudi pri visoko pregibih jezikih, le v veliko večjem obsegu, kot kaže tabela 1.

Tabela 1: Vse besedne oblike slovenske leme mesto

besedna oblika	število	sklon
mest-o	ednina	imenovalnik
mest-a	ednina	rodilnik
mest-u	ednina	dajalnik
mest-o	ednina	tožilnik
mest-u	ednina	mestnik
mest-om	ednina	orodnik
mest-a	množina	imenovalnik
mest-∅	množina	rodilnik
mest-om	množina	dajalnik
mest-a	množina	tožilnik
mest-ih	množina	mestnik
mest-i	množina	orodnik
mest-i	dvojina	imenovalnik
mest-∅	dvojina	rodilnik
mest-oma	dvojina	dajalnik
mest-i	dvojina	tožilnik
mest-ih	dvojina	mestnik
mest-oma	dvojina	orodnik

3.1.1 Izdelava paradigem

Besede korpusa družimo v paradigme zaradi lažjega rokovanja z mnogimi besednimi oblikami obeh jezikov, saj sta tako slovenski kot srbski jezik visoko pregibna kot kaže primer iz prejšnjega poglavja.

Vsaka paradigma je predstavljena z:

- tipično lemo, iz te leme je bila sestavljena osnova paradigme;
- krnom, najdaljšim začetkom besede, ki je lasten vsem besednim oblikam v paradigmi;
- množico vseh besednih oblik razdeljenih na krne in končnice ter morfosintaktične deskriptorje (MSD) [13]

Z uporabo algoritma na sliki 2 je iz korpusa sestavljen leksikon za izvirni in ciljni jezik. Leksikon je sestavljen iz seznama enoličnih besed predstavljenih z lemami ter MSDji [13].

Vse besedne oblike določene leme so združene v razred, ki ga predstavlja lema. Iz vsakega razreda sestavimo paradigmo, torej za vsako lemo sestavimo ločeno paradigmo. Z

[1]besede so sestavljene iz množice zlepljenih morfemov; morfemi sestavljajo krn ter obrazila

združevanjem posameznih paradigem sestavljamo večje paradigme, ki predstavljajo vse leme iz katerih so nastale.

```
//par – paradigme
for(i = 0; i < par.size; i++){
  for(j = i; j < par.size; j++){
    if(par[i].POS == par[j].POS){
      if(all entries agree){
        join(par[i], par[j])
      }
    }
  }
}
```

Slika 2: Algoritem za izdelavo paradigem

Dve paradigmi združimo, če se leme obeh paradigem ujemajo, v prvem koraku sta le dve lemi, sčasoma pa število narašča. Lemi se ujemata, če spadata v isto besedno vrsto, imata isto POS - part of speech oznako, in imajo vse besedne oblike z enakimi MSD oznakami obeh lem enake pripone. Zapise obeh paradigem združimo v novo množico (brez duplikatov). Vsaka paradigma hrani tudi podatke o vseh lemah, ki jo sestavljajo.

Enojezični izvorni in cilji slovarji so bili sestavljeni s pomočjo paradigem, tako smo izdelali slovarje s približno 20-krat več besednimi oblikami kot smo jih dobili iz samega korpusa. Rezultati testiranja spremembe velikosti enojezičnih slovarjev z uporabo paradigem so prikazani v razdelku 4.1.

3.2 Izdelava dvojezičnega prevajalnega slovarja

V primerjavi z jeziki, kjer pregibnost ni posebej izražena, je za visoko-pregibne jezike, kot so slovanski jeziki, značilno veliko število enoličnih besednih oblik v enakih količinah besedila. Slika 5 kaže razlike v številu besednih oblik v istem korpusu [11] za štiri jezike; tri visoko pregibne slovanske jezike: slovenščino, srbščino in češčino ter za angleščino kot primerjavo.

Tabela 2: število lem v korpusu MULTEXT-EAST [11]

jezik	število besed	leme
slovenščina	22134	6512
srbščina	21435	6832
češčina	23654	7263
angleščina	11293	8182

Zmanjšanje iskalnega prostora intuitivno omogoča izboljšavo kakovosti modela, prevajalnega modela temelječega na posameznih besedah (the word-by-word translation model). Rezultat smo tudi preverili in je zapisan ter natančneje razložen v poglavju z rezultati. Veliko informacij pa smo pri lematizaciji izgubili, vsaj del te informacije smo ohranili z vpeljavo oznak.

Oglejmo si ta pojav na primeru: model za poravnavo besed je naučen na učni množici sestavljeni iz parov lema+besedna vrsta dejanskih besednih oblik iz vzporednega korpusa.

Nekaj enostavnih definicij, ki omogočajo formulacijo enačbe (1)

L - jezik, vse besede

E_L - leme jezika L

$E_{L(i)}$ - i^{th} lema z vsemi pripadajočimi besednimi oblikami

$$|L| = \sum_{i=0}^{|E_L|} E_{L(i)} \tag{1}$$

Iskalni prostor je zmanjšan iz $|L|$ na $|E_L|$.

Oglejmo si primer:

Poglejmo si vrednosti v tabeli 2, ob predvidevanju, da je novela Georga Orwella "1984", ki predstavlja večinski del večjezičnega povedno poravnanega korpusa [11], dober vzorec jezika, v našem primeru slovenskega jezika. Iskalni prostor se je zmanjšal iz 22134 besednih oblik na 6512 lem.

Izvorni jezik $|L| = 22134$

Lematiziran jezik $|E_L| = 6512$

Slika 3: zmanjšanje iskalnega prostora za slovenski jezik (relativno majhen korpus MULTEXT-EAST [11])

Dvojezični, vzporedni, označeni korpus [11] vsebuje izvorno besedilo z dodatnimi opisi v obliki XML oznak sledeč smernicam TEI-P4 [9] ter smernicam EAGLES [27]. Primer korpusa je predstavljen na sliki 4.

```
<s id="Osl.2.3.5.11">
  <w lemma="priti" ana="Vmps-dma">Prisla</w>
  <w lemma="biti" ana="Vcip3d--n">sta</w>
  <w lemma="do" ana="Spsg">do</w>
  <w lemma="podrt" ana="Afpnsg">podrtega</w>
  <w lemma="drevo" ana="Ncnsg">drevesa</w>
  <c>,</c>
  <w lemma="o" ana="Spsl">o</w>
  <w lemma="kateri" ana="Pr-nsl----a"> katerem</w>
  <w lemma="on" ana="Pp3msd--y-n">mu</w>
  <w lemma="biti" ana="Vcip3s--n">je</w>
  <w lemma="praviti" ana="Vmps-sfa"> pravila</w>
  <c>.</c>
</s>
```

Slika 4: poved iz korpusa

```
priti_V biti_V do_S podrt_A drevo_N ,
o_S kateri_P on_P biti_V praviti_V .
```

Slika 5: pripravljeni podatki: leme in besedne vrste vsake besede iz korpusa

Prevajalni model na osnovi besed paradigme statističnega strojnega prevajanja, SMT word-to-word alignment model, [7], [34] je bil naučen na vzporednem, povedno poravnanem seznamu lematiziranih besed. Seznam je samodejno izdelan na osnovi korpusa, izsek je prikazan na sliki 5. Poravnava lem omogoča veliko boljšo kakovost poravnav kot poravnave vseh besednih zvez zaradi zmanjšanega prostora iskanja, kot je opisano v enačbi (1) in na sliki 3. Besede enojezičnih slovarjev povežemo s prevodnimi zapisi (dvojezičnimi pari izvornih in ciljnih lem) s pomočjo paradigem, ki vsebujejo informacijo o lemah, natančneje je proces prikazan v razdelku 3.1.1.

3.3 Samodejna izdelava prevajalnih pravil

Pri snovanju preizkusa smo se omejili na samodejno izdelavo pravil plitkega prenosa. Takšna pravila so najprimernejša za prevajalne sisteme sorodnih jezikov. Samodejne metode popolne slovnične analize povedi ne dosegajo zadovoljive kakovosti in njihova vpeljava v sisteme za prevajanje sorodnih jezikov bi zmanjšala kakovost prevodov v primerjavi s prevodi z uporabo pravil plitkega prenosa (shallow transfer rules). To dejstvo predstavljajo avtorji [15] in [20]. Samodejna izdelava pravil plitvega prenosa je bila izvedena s pomočjo metode in orodja predstavljenega v [40]. To orodje izdeluje pravila, ki jih lahko nepredelana uporabimo v ogrodju prevajalnih sistemov temelječih na Apertiumu [10], za ostale sisteme moramo format pravil spremeniti, vendar je tudi ta postopek lahko popolnoma samodejen.

Samodejna, nenadzorovana izdelava pravil veliko množico pravil, za izbiro najboljših pravil smo uporabili metodo predstavljeno v [43]. Evalvacija kakovosti izdelanih pravil ter osnovni podatki o številu ter velikosti pravil so podani v razdelku 4.2.

3.4 Učenje modela za razdvoumljanje

Večina avtorjev sistemov za prevajanje sorodnih jezikov [10], [19] priporoča zgradbo prevajalnih sistemov kot je predstavljena na sliki 1 apertium sistema. Takšna zgradba prevajalnega sistema uporablja morfološki označevalnik, označevalnik POS - POS tagger kot orodje za radvoumljanje analiziranega izvornega besedila. Pri razdvoumljanju označevalnik na podlagi vnaprej naučenega modela ter že obdelanega besedila izbere najverjetnejšo morfološko obliko besed. [21] predlaga zgradbo, ki ne uporablja označevalnika in razdvoumljanje prepusti iskanju najprimernejše končne povedi s pomočjo rangiranja na osnovi jezikovnega modela ciljnega jezika. Ta metoda ni bila uporabljena zaradi možnosti kombinatorične eksplozije števila možnih kandidatov za končne prevode.

V začetni fazi postavitve sistema smo preizkusili dva označevalca POS in sicer: TnT [6], del orodja TOTALE [14] ter označevalec POS, ki je priložen orodju Apertium [39]. Prvi je že bil samodejno in nenadzorovano naučen na označenem korpusu [11], drugi pa na večjem a neoznačenem korpusu, ki smo ga pripravili za ta preizkus z zbiranjem gradiv iz interneta.

Izkazalo se je, da je kakovost prvega označevalnika [14] veliko boljša, vgradili smo ga v prevajalni sistem. Sami primerjavi ne smemo posvečati preveč pozornosti, saj je označevalnik [39] naučen na veliko slabših učnih podatkih.

4 Evalvacija, metodologija in rezultati

Testiranje je bilo razdeljeno na tri dele, ki pa niso imeli enake uteži. Najprej smo si ogledali spremembo velikosti enojezičnih slovarjev z uporabo paradigem, rezultati so prikazani v razdelku 4.1, sledi primerjava testiranje izboljšave kakovosti prevodov sistema s samodejno zgrajenimi pravili v razdelku 4.2. Največjo pozornost smo posvetili kakovosti končnih prevodov prevajalnega sistema, ta testiranja so bila izvedena v štirih delih, vsak del je predstavljen v posebnem razdelku tega poglavja:

1 Samodejna objektivna evalvacija z uporabo metrike BLEU [36].
2 Samodejna objektivna evalvacija z uporabo metrike METEOR [3], [25].
3 Ročna metoda z uporabo Levenshteinove razdalje (Levenshtein edit distance) [28], prevedene povedi do najbližje pravilne povedi. Pri tej metodi ročno popravimo prevedene besedilo ter preštejemo število sprememb, ki smo jih naredili.
4 Ročna subjektivna metoda z uporabo smernic [26].

Dodatne metode preverjanja kakovosti prevodov smo izvedli po prvih poskusih uporabe metode BLEU, ki je pokazala presenetljivo slabe rezultate. Veliko avtorjev se strinja [8], da metrika BLEU, ki je sicer najbolj razširjena samodejna metrika in zelo čaščena v krogih SMT, ni primerna za evalvacijo prevajalnih sistemov tipa RBMT [22], še posebej pa ni primerna za visoko pregibne jezike kot so slovanski jeziki.

Avtorji metode ter tudi orodja METEOR [3], [25] trdijo, da njihova metrika rešuje večino problemov, ki so se pokazali pri uporabi metrike BLEU, vendar ta metrika še ni razširjena in tako so primerjave z drugimi sistemi težavne. Na žalost METEOR ni podpiral jezikovnega para, ki smo ga uporabili v naših poizkusih. Pred izvedbo evalvacije smo sistemu dodali poseben modul za krnjenje besedila za slovenski jezik.

Dvojezični vzporedni korpus [11] je bil uporabljen v evalvaciji prevodov.

K-kratno prečno preverjanje [24] je bilo uporabljeno kot metoda generalizacije napake, saj je najprimernejše za majhne učne ter testne množice. Izbrali smo petkratno-prečno preverjanje namesto pogosteje uporabljenega desetkratnega prečnega preverjanja, saj izdelava končnega popolnoma delujočega sistema ni bila popolnoma avtomatizirana. Korpus smo razdelili na pet delov, vsak del je obsegal približno 1700 povedi. Evalvacija je obsegala izbiro enega dela tako razdeljene množice za testiranje, ostali štirje deli so bili uporabljeni kot učna množica. Prevajalni sistem je bil postavljen po predstavljeni metodologiji iz razdelka 3 z uporabo izbrane učne množice. Vrednosti evalvacije v vsakem koraku preverjanja so zbrane v tabelah 5 in 6.

4.1 Evalvacija spremembe velikosti enojezičnih slovarjev

Iz korpusa [11] smo izluščili besednjak, sestavljen je bil iz vseh besednih oblik prisotnih v korpusu ter njihovih lem. S pomočjo metode predstavljene v razdelku 3.1 smo zgradili paradigme ter izdelali seznam vseh lem razširjenih na vse možne oblike v paradigmah. V tabeli 3 so prikazane velikosti osnovnih besednjakov ter razširjenih besednjakov.

Tabela 3: rezultati metrike BLEU, vsak korak je prikazan v posebni vrstici, zadnji dve vrstici kažeta povprečja in standardno deviacijo.

jezik	Osnovni besednjak	Razširjeni besednjak
testiran sistem	84,8	77,8
Česílko	96,4	88,3

4.2 Evalvacija samodejne izdelave pravil za plitki transfer

Izdelali smo ločen prevajalni sistem, ki se je od osnovnega razlikoval le po modulu, ki je vseboval pravila za plitki transfer.

Preverjali smo spremembo kakovosti končnih prevodov osnovnega prevajalnega sistema brez uporabe pravil za plitki transfer ter sistema, ki je takšna pravila vseboval. V tabeli 4 so prikazani rezultati. Uporabljena je bila metrika utežene Levenshteinove razdalje, ki je predstavljena v razdelku 4.5. Tabela 4 kaže kakovost prevodov osnovnega sistema ter izboljšanega sistema z uporabo pravil plitkega prenosa.

Tabela 4: rezultati kakovosti prevodov osnovnega sistema ter sistema z uporabo pravil plitkega transfera.

Sistem	E.D.[2] znaki	E.D.[3] besede
Osnovni sistem	69,7	59,3
Sistem s pravili	84,8	77,8

4.3 Samodejna objektivna evalvacija z uporabo metrike BLEU [36]

Pri testiranju je bila uporabljena javno dosegljiva implementacija metrike BLEU [31], različica v11b. Rezultati so prikazani v tabeli 5.

Tabela 5: rezultati metrike BLEU, vsak korak je prikazan v posebni vrstici, zadnji dve vrstici kažeta povprečja in standardno deviacijo.

korak	vrednost BLEU
1	0.1167
2	0.1211
3	0.1206
4	0.1198
5	0.1201

[2]Utežena Levenshteinova razdalja [28] na znakih med osnovnim ter popravljenim prevodom

[3]Utežena Levenshteinova razdalja [28] na besedah

povprečje	0.1196
STDEV	0.0017

Vrednosti so relativno nizke, še posebej, če upoštevamo bližino jezikov jezikovnega para. Nizke vrednosti lahko deloma pripišemo visoki pregibnosti jezikov, deloma pa sami metriki BLEU, ki pripisuje sistemov RBMT nižje vrednosti [8].

4.4 Samodejna objektivna evalvacija z uporabo metrike METEOR

Pri testiranju je bila uporabljena javno dosegljiva implementacija metrike METEOR [25] verzija v0.6. METEOR uporablja krnjenje besed kot enega od mehanizmov, ki povečujejo korelacijo med metriko METEOR ter človeško, ročno evalvacijo. Uporabljen je bil algoritem za krnjenje, ki je stranski izdelek metod opisanih v razdelku 3. rezultati so prikazani v tabeli 6.

Tabela 6: rezultati metrike METEOR, vsak korak je prikazan v posebni vrstici, zadnji dve vrstici kažeta povprečja in standardno deviacijo.

korak	vrednost METEOR
1	0.6344
2	0.6296
3	0.6316
4	0.6297
5	0.6352
Povprečje	0.6321
STDEV	0.0026

4.5 Ročna metoda z uporabo Levenshteinove razdalje

Levenshteinova razdalja [28], ki šteje število sprememb, zamenjav, izbrisov, vrivov ter premikov besed, ki jih opravimo pri predelavi prevoda testiranega sistema v slovnično popolnoma pravilno poved v ciljnem jeziku. Ta procedura ponazarja koliko dela moramo vložiti v predelavo našega izdelka v dober izdelek. Metrika grobo ponazarja kompleksnost opravila končnega čiščenja prevodov, post-editing task.

Uporabili smo dve različici edit-distance [28], različico, ki računa uteženo razdaljo v znakih ter različico, ki izračuna uteženo razdaljo v besedah

Pri evalvaciji je bilo upoštevanih 100 naključno izbranih povedi iz testnega dela korpusa. Povedi so bile prevedene s testiranim prevajalnim sistemom, rezultati so bili ročno popravljeni v pravilne povedi ciljnega jezika, v našem primeru srbskega jezika. Kot pravilne povedi v tem primeru pojmujemo povedi, ki so sintaktično pravilne in izražajo isti pomen kot izvorne povedi. Izračunana je bila Levenshteinova razdalja med osnovnimi ter popravljenimi prevodi.

Končni rezultat je bil 84,8%, kar pomeni, da je bilo potrebno popraviti ali premakniti 100% - 84,8% vseh znakov. Te rezultate lahko primerjamo s podobnim sistemom, ki je bil ročno zgrajen in je nastajal več let [20]. Sistem [20] prevaja iz češkega v slovaški jezik. Lastnosti obeh

jezikovnih parov so si podobne in tudi razlike med jeziki so primerljive, tako je primerjava rezultatov obeh sistemov opravičljiva. Sistem [20] predstavlja enega najboljših prevajalnih sistemov sorodnih jezikov za visoko pregibne jezike, vrednost metrike Levenshteinove razdalje je 96,45 % in predstavlja dobro točko končnega cilja predstavljenega sistema. Rezultati so predstavljeni v tabeli

Tabela 7: rezultati metrike Levenshteinove razdalje, primerjava testiranega sistema s sistemom Česílko [20].

sistem	E.D.[4] znaki	E.D.[5] besede
testiran sistem	84,8	77,8
Česílko	96,4	88,3

4.6 Ročna subjektivna metoda z uporabo smernic [26]

Subjektivna ročna evalvacija kakovosti prevodov je bila izvedena po smernicah letne delavnice NIST Machine Translation Evaluation Workshop [26], ki jo organizira LDC, Linguistic Data Consortium. Ta metodologija je najširše uporabljana pri evalvaciji kakovosti prevodov sistemov za strojno prevajanje. Sestavljena je iz dveh lestvic s po petimi vrednostmi, ki prikazujeta pravilnost prevodov ter slovnično pravilnost ciljnega besedila.

Prva lestvica kaže kakovost prevodov, koliko izvornega pomena se je pri prevodu ohranilo:

- 5 = Vse
- 4 = Večino
- 3 = Precej
- 2 = Malo
- 1 = Nič

Druga lestvica kaže kako slovnično pravilne so povedi v ciljnem jeziku. Pri prevodu v srbski jezik velja:

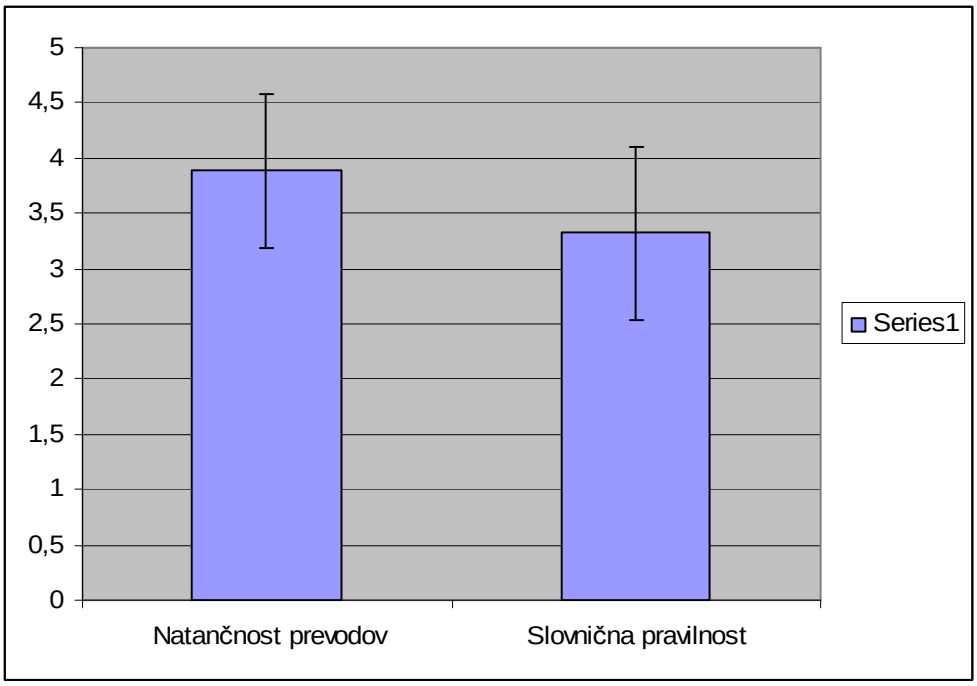
- 5 = Prevod brez napak
- 4 = Dobra srbščina
- 3 = Srbščina, kot ne-materni jezik (non-native)
- 2 = Srbščina z veliko napakami
- 1 = Nesmiselno besedilo

Ločeni lestvici za kakovost prevodov ter slovnično pravilnost sta bili izdelani ob predpostavki, da lahko tudi prevod z veliko slovničnimi napakami prikaže vso informacijo, ki je zapisana v originalu. Štirje samostojni ocenjevalci, dvema med njimi je bil ciljni jezik materni jezik, so pregledali in ocenili po 100 s pomočjo opisane metodologije. Rezultati so prikazani na

[4]Utežena Levenshteinova razdalja [28] na znakih med osnovnim ter popravljenim prevodom

[5]Utežena Levenshteinova razdalja [28] na besedah

sliki 6.



Slika 6: rezultati evalvacije na podlagi smernic [26]. Povprečne vrednosti štirih neodvisnih ocenjevalcev kažejo visoke vrednosti za kakovost prevodov, nekoliko nižje vrednosti za slovnično pravilnost prevodov v ciljnem jeziku.

5 Razprava

Članek predstavlja poskus združevanja večih metod za hitro postavitev prevajalnih sistemov za sorodne visoko pregibne jezike. Sistem temelji na osnovi pravil plitkega prenosa. Metode so bile preizkušene na primeru samodejne izdelave prevajalnega sistema. Evalvacija kaže perspektivne rezultate, čeprav je možnost napredovanja še vedno dovolj velika. Zadnja različica prevajalnega sistema zgrajenega po opisanih metodah je postavljena v obliki spletnega vmesnika na naslovu: http://jt.upr.si/guat/

Literatura

[1] Ahrenberg, L. and M. Holmqvist. Back to the Future? The Case for English-Swedish Direct Machine Translation. *Proceedings of Recent Advances in Scandinavian Machine Translation*, 2005.

[2] Altintas, K. and I. Cicekli. A Machine Translation System between a Pair of Closely Related Languages. *Proceedings of the 17th International Symposium on Computer and Information Sciences (ISCIS 2002)*, 2002.

[3] Banerjee, S. and A. Lavie. METEOR: An Automatic Metric for MT Evaluation with

Improved Correlation with Human Judgments. *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*, 2005. Ann Arbor, Michigan.

[4] Baum, L. and T. Petrie. Statistical inferrence for probabilistic functions in finite state Markov chains. *Ann. Math. Stat.*, 37:1554-1563, 1966.

[5] Bick, E. and L. Nygaard. Using Danish as a CG Interlingua: A Wide-Coverage Norwegian-English Machine Translation System. *Proceedings of NODALIDA, Tartu*, 2007.

[6] Brants, Thorsten. TnT -- a statistical part-of-speech tagger. *Proceedings of the 6th Applied NLP Conference*, 2000. Seattle, WA.

[7] Brown, Peter F. and Stephen A. Della Pietra and Vincent J. Della Pietra and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational linguistics*, 19(2):163-311, 1993.

[8] Callison-Burch, Chris and Miles Osborne and Philipp Koehn. Re-evaluating the role of Bleu in machine translation research. *Proceedings of EACL*, 2006.

[9] TEI Consortium. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Technical report, TEI consortium, 2007.

[10] Corbi-Bellot, Antonio M. and Mikel L. Forcada and Sergio Ortiz-Rojas and Juan Antonio Prez-Ortiz and Gemma Ramirez-Sanchez and Felipe Sanchez-Martinez and Inaki Alegria and Aingeru Mayor and Kepa Sarasola. An open-source shallow-transfer machine translation engine for the Romance languages of Spain. *Proceedings of the Tenth Conference of the European Association for Machine Translation*, pages 79--86, 2005.

[11] Dimitrova,Ludmila and Nancy Ide and Vladimir Petkevic and Tomaz Erjavec and Heiki Jaan Kaalep and Dan Tufis. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. *COLING-ACL*, pages 315-319, 1998.

[12] Dyvik, H. Exploiting Structural Similarities in Machine Translation. *Computers and Humanities*, 28:225-245, 1995.

[13] Erjavec, Tomaz. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04*, 2004.

[14] Erjavec, Tomaz. Multilingual tokenisation, tagging, and lemmatisation with totale. *Proceedings of the 9th INTEX/NOOJ Conference*, 2006.

[15] Mikel L. Forcada. Open-source machine translation: an opportunity for minor languages. *Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages)*, 2006.

[16] Google. Google translator. 2008.

[17] Hajic, J. An MT System Between Closely Related Languages. *Proceedings of the third conference of the European Chapter of the Association for Computational Linguistics*, 1987.

[18] Hajic, J. and J. Hric and V. Kubon. Machine translation of very close languages. *Proceedings of the 6th Applied Natural Language Processing Conference*, 2000.

[19] Haji č, Jan and Petr Homola and Vladislav Kubon. A simple multilingual machine translation system. *Proceedings of the MT Summit IX*, New Orleans, 2003.

[20] Homola, Petr and Vladislav Kubon. A method of hybrid MT for related languages. *Proceedings of IIS*, 2008.

[21] Homola, Petr and Vladislav Kubon. Improving machine translation between closely related Romance languages. *Proceedings of EAMT*, pages 72 - 77, 2008.

[22] Hutchins, John. Towards a definition of example-based machine translation. *MT Summit X, Proceedings of Workshop on Example-Based Machine Translation*, pages 63-70, 2005.

[23] J. Haji č and J. Hric and V. Kubo ň. Machine translation of very close languages. *Proceedings of the 6th Applied Natural Language Processing Conference*, pages 7--12, Seattle, Washington, USA, 2000.

[24] Kohavi, Ron. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1137-1143, 1995.

[25] Lavie, A. and A. Agarwal. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. *Proceedings of Workshop on Statistical Machine Translation at the 45th Annual Meeting of the Association of Computational Linguistics (ACL-2007)*, 2007.

[26] LDC. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Technical report, LDC, 2005.

[27] Leech, GN and A. Wilson. EAGLES Recommendations for the Morphosyntactic Annotation of Corpora. Technical report, ILC-CNR, Pisa, 1996.

[28] Levenshtein, V. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk*, :845-848, 1965.

[29] Multext. The Multext Project. 2007.

[30] Nagao, Makoto. A framework of a mechanical translation between Japanese and

English by analogy principle. *Artificial and Human Intelligence*, 1984.

[31] NIST. Evaluation Software. 2008.

[32] NIST. NIST 2006 Machine Translation Evaluation Official Results. 2006.

[33] Och, Franz Josef. Challenges in Machine Translation. *Proceedings of ISCSLP*, 2006.

[34] Och, Franz Josef and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational linguistics*, 29:19-51, 2003.

[35] Carme Armentano Oller and Mikel L. Forcada. Open-source machine translation between small languages: Catalan and Aranese Occitan. *Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages)*, pages 51-54, 2006. (organized in conjunction with LREC 2006 (22-28.05.2006)).

[36] Papineni, Kishore and Salim Roukos and Todd Ward and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. Technical report, IBM, 2001.

[37] Popovic, M. and P. Willett. Krnjenje kot osnova nekaterih nekonvencionalnih metod poizvedovanja. *Knjinica, Ljubljana*, 44, 2000.

[38] Popovic, M. and P. Willett. The effectiveness of stemming for natural language access to Slovene textual data. *Journal of the American Society for Information Science*, 43(5), 1992.

[39] Sanchez-Martinez, Felipe and Carme Armentano-Oller and Juan Antonio Perez-Ortiz and Mikel L. Forcada. Training part-of-speech taggers to build machine translation systems for less-resourced language pairs. In Daz Madrigal, Vctor J. and Enrquez de Salamanca Ros, Fernando, editors, *Procesamiento del Lenguaje Natural (XXIII Congreso de la Sociedad Espanola de Procesamiento del Lenguaje Natural)*, pages 257--264, 2007.

[40] Sanchez-Martinez, Felipe and Mikel L. Forcada. Automatic induction of shallow-transfer rules for open-source machine translation. In Andy Way and Barbara Gawronska, editors, *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, pages 181--190, 2007. Skovde University Studies in Informatics.

[41] Scannell, K.P. Machine translation for closely related language pairs. *Proceedings of the Workshop Strategies for developing machine translation for minority languages*, 2006.

[42] Vicic, Jernej. Rapid development of RBMT systems for related languages. *Translating and the computer 29 : proceedings of the twenty-ninth international conference on translating and the computer*, pages 162-1733, 2007.

[43] Vicic, Jernej and Mikel L. Forcada. Comparing greedy and optimal coverage

strategies for shallow-transfer machine translation. *Intelligent information systems XVI : proceedings of the International IIS `08 conference*, pages 307-316, 2008.

[44] Welch, Lloyd R. Hidden Markov Models and the Baum-Welch Algorithm. *IEEE Information Theory Society Newsletter*, 53(4):1-14, 2003.