Vsak začetek je težak

Jernej Vičič*, Tomaž Erjavec†

*Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Tržaška 2, 1000 Ljubljana
jernej.vicic@guest.arnes.si

†Odsek za inteligentne sisteme
Institut "Jožef Stefan"
Jamova 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

Povzetek

Delo predstavlja poizkus avtomatskega prevajanja iz slovenskega jezika v angleški na osnovi statističnega strojnega prevajanja (Statistical Machine Translation, SMT). EGYPT je zbirka orodij za obdelavo dvojezičnih vzporednih korpusov 1.2 za strojno prevajanje. Narejen je bil na poletni delavnici na univerzi JHU - John Hopkins University in je trenutno najširše uporabljana zbirka v sistemih strojnega prevajanja. IJS-ELAN korpus vsebuje milijon besed, prevodov iz slovenščine v angleščino in obratno, ki so komentirane in poravnane po povedih. Oba jezika sta označena morfosintanktičnimi opisi ter kontekstno neodvisnimi lemami. Korpus je kodiran v XML zapis po navodilih TEI Guidelines P4. Sistem za prevajanje iz slovenščine v angleščino z uporabo zbirke EGYPT in korpusa IJS-ELAN predstavlja osnovo tega dela, osnovni sistem je bil še razširjen. Delo predstavlja še motive za izbiro izvornega ter ciljnega jezika, to je zakaj smo izbrali prevajanje iz slovenščine v angleščino in ne obratno. Izvedeno je bilo tudi osnovno vrednotenje sistema. Osnovni model je bil razširjen z uporabo beležk v korpusu, posebno z uporabo kontekstno neodvisnih lem, ki niso odvisne od bogate uporabe sklanjatev ter spreganja v slovenskem jeziku. Z uvajanjem lem smo želeli obiti največjo pomanjkljivost našega sistema: osnovan je na relativno majhnem korpusu. Nove prevedbe so bile ovrednotene ter rezultati primerjani z osnovnim sistemom, pri vrednotenju so bili uporabljeni isti testni primeri. Vrednotenje je bilo izvedeno z dvema metodama:

- SA/TA, je različica urejevalne razdalje (edit distance), omogoča avtomatsko vrednotenje. Prevodi, narejeni z našim sistemom so primerjani z referenčnimi prevodi, računa se razdalja med obema prevodoma. Razdalja pomeni normalizirano najmanjše število vstavitev, brisanj, zamenjav ter premestitev besed v ocenjevani povedi v primerjavi z referenčno. Korpus je razdeljen na testne ter učne pare, učni pari so uporabljeni pri gradnji sistema, testni pa pri testiranju urejevalne razdalje kot referenčni prevodi.
- SSER, subjective sentence error rate. Prevode našega sistema so pri tej metodi ocenjevali strokovnjaki, razvrščali so jih v pet razredov od "popolne nesmiselnosti" do "popolnega prevoda". Rezultati so predstavljeni ter utemeljeni.

1. Uvod

To delo predstavlja prvi sistem za avtomatsko prevajanje iz slovenskega jezika v angleški na osnovi statističnega strojnega prevajanja (SMT - Statistical Machine Translation).

SMT je mlada veja računalništva saj so bili do sedaj večini dostopni računalniki za resnejše premlevanje ogromnih količin podatkov, ki so osnova vseh statističnih prevajalnih sistemov, premalo močni. Zapletene matematične osnove, temelj SMT, so prav tako odvrnile marsikaterega raziskovalca.

V uvodnih delih so predstavljeni osnovni pojmi SMT ter bolj splošno strojnega prevajanja ter teoretične matematične osnove statističnih algoritmov. Na začetku devetdesetih let prejšnjega stoletja so pri IBM osnovali prvi sistem za statistično strojno prevajanje in postavili temelje za nadaljnja raziskovanja in izboljšave predlaganih metod. Sistem temelji na osnovi parametričnih statističnih modelov, ki so natančno prikazani.

Nadaljevanje prinaša prikaz trenutno najbolj obetajoče in vsesplošno priznane zbirke orodij za rokovanje z dvojezičnimi, vzporednimi korpusi Egipt. Zbirka je nastala kot rezultat poletne delavnice na John Hopkins University. Po pričevanjih mnogih avtorjev člankov s področja SMT, po pregledu referenc v literaturi ter po pregledu povezav na internetu ostaja Egypt daleč najbolj uporabljana ter opisovana zbirka orodij za SMT. Narejena je bila z namenom zapolniti vrzel na tem področju ter kot enostavna osnova za nadaljnje raziskave ter izboljšave osnovnih algoritmov. Nekatera orodja so bila kasneje še dopolnjena in popravljena, uporabljajo tudi dodatne prevajalne modele, ki so opisani.

S pomočjo predstavljene zbirke orodij je bil postavljen sistem za prevajanje besedil iz Slovenščine v Angleščino. Predstavljen je sam sistem ter najpomembnejše faze učnega in prevajalnega dela sistema. Opisani so tudi glavni deli testiranja.

Naslednje poglavje predstavlja osnovne motivacije za izbiro izvornega jezika - Slovenščine. Kot materin jezik avtorja tega dela je bil prvi na lestvici izbire, predstavljene so osnovne značilnosti jezika in težave, ki jih prinašajo. Drugi razlog je bil, da v Sloveniji takšen sistem še ni bil postavljen in preizkušan, tretji razlog pa leži v sami zapletenosti slovenskega jezika, ki prinaša dodatne izzive.

Razlogi za izbiro Angleščine kot ciljnega jezika so predstavljeni v razdelku 6. Angleščina je trenutno najširše uporabljan jezik v računalniškem svetu, prikazane so smernice širitve jezika. Pri izbiri je pomembno vlogo igrala

pogostost uporabe tega jezika in pomembne izkušnje ostalih raziskovalcev tega področja.

Osnovni algoritmi po [1] so bili razširjeni z uporabo lem v korpusu, prikazani so rezultati izboljšav ter motivacije za izbiro te nove metode.

Testiranje, predstavitev rezultatov in njihovo komentiranje zaključujejo to delo. Rezultati, prevodi, so bili ocenjeni po dveh metodah ocenjevanja. Metodi, ureditvena razdalja (edit-distance) in SSER - Subjective Sentence Error Rate, sta predstavljeni v 8 ter povzeti (z logičnimi popravki) po [7] in [8].

2. Namen članka

To delo predstavlja prvi sistem za avtomatsko prevajanje iz slovenskega jezika v angleški na osnovi statističnega strojnega prevajanja (SMT - Statistical Machine Translation). Za postavitev statističnih modelov je bila uporabljena zbirka orodij Egypt, kot osnova za prevajanje prevajanje pa je služil dvojezični vzporedni korpus IJS-ELAN.

2.1. Statistično strojno prevajanje

Razdelek predstavlja osnovne pojme statističnega strojnega prevajanja - SMT ter osnovne motivacije za razvoj te mlade in do sedaj zapostavljane veje računalništva.

že od nekdaj je poskušal človek opisati jezik s pomočjo pravil, prvi primeri segajo vsaj 2000 let nazaj. Pri opisovanju večine naravnih jezikov s strogimi pravili pa se pojavi kup problemov. Naravni jezik je preveč kompleksna ter živa tvorba in pravila za opisovanje so preveč kompleksna, če jih je sploh mogoče vsa zapisati. že v začetku tega stoletja so prišli strokovnjaki do tega zaključka, "All grammars leak", (vse gramatike puščajo) [11].

Natančno določanje pravil jezika, ukleščanje v stroge okvire pravil, ni obrodilo sadov, potrebujemo bolj ohlapne omejitve, ki upoštevajo tudi kreativnost pri uporabi jezika.

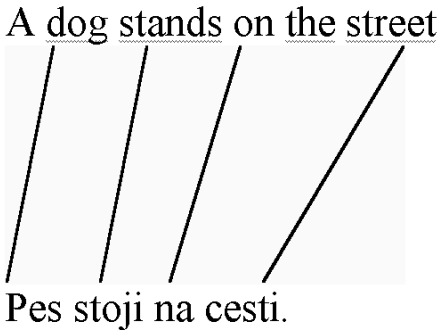
Namesto razdeljevanja stavkov na po gramatičnih pravilih iščemo splošne vzorce, ki se porajajo pri uporabi jezika. Glavno orodje za iskanje takšnih vzorcev je štetje raznovrstnih objektov, bolj strokovno izraženo statistika. Od tod izvira tudi ime statistično strojno prevajanje.

SMT je osnovan na parametričnih statističnih modelih, podmnožica teh modelov je bila uporabljena tudi pri snovanju našega sistema. Modeli so natančneje predstavljeni v 2.6.

2.2. Problem nepopolnih podatkov (Scarce data problem)

Razdelek opisuje najpomembnejši problem statističnega strojnega prevajanja, učenje na nepopolnih podatkih.

Zbiranje dovolj velikih in pravilnih podatkovnih baz učnih primerov, v našem primeru postavljanje dovolj velikih in primernih dvojezičnih vzporednih korpusov je dolgotrajno opravilo. Med zbiranjem podatkov se pogosto prikradejo napake, ki jih v veliki količini podatkov le s težavo odkrivamo. Poleg dolgotrajnega dela pa postavljajo še dodatno prepreko avtorske pravice večine del, ki jih želimo uporabljati.



Slika 1: primer povezav med besedami v angleški ter slovenski povedi

Problem pomanjkljivih podatkov rešujemo s pomočjo naprednih algoritmov, ki poskušajo zakriti pomanjkljive oziroma manjkajoče podatke. Algoritmi upoštevajo predhodno znane o problemu, izkušnje iz sorodnih področij ali pa celo povsem tujih področij. (c)umne podatke izločamo s pomočjo zakonitosti v podatkih, z izločanjem ekstremov. Paziti moramo, da pri izločanju napačnih podatkov ne pretiravamo in korpus preveč "porežemo", poenostavimo.

Napake v učni bazi poskušamo izločiti z avtomatskimi orodji ter ročnim pregledovanjem podatkov. Prednost avtomatskih metod je v enostavnosti in hitrosti uporabe, nimamo pa popolnega nadzora nad delovanjem. Pri ročnem pregledovanju podatkov nam pomagajo eksperti področja, v našem primeru jezikoslovci, ki sodelujejo že pri sestavljanju gradiv za korpus. S pomočjo njihovih izkušenj lahko že v začetni fazi definiramo manjši a bolj informativen korpus, ki je lažji za obdelavo, skriva manj napak a vsaj enako dobro opisuje jezik.

2.2.1. Poravnava, alignment

Predstavljajmo si par slovenske in angleške povedi f,e , ki sta ena prevod druge. Čeprav smo si v prejšnjih razdelkih ogledali, da direktno prevajanje besed ne prinaša dovolj dobrih prevodov, pa vseeno obstaja določena povezava med posameznimi besedami v obeh povedih. Primer povezav si lahko ogledamo na prikazu Slika 1.

Takšen niz povezav imenujemo poravnava (alignment). Formalna definicija poravnave: Niz parov f,e , kjer vsak par predstavlja povezavo med j-to besedo f (slovenska beseda)ter i-to besedo e (angleška beseda).Povezati želimo fj ter ei, kjer ei izraža vsebino fj v angleščini. Za lažjo predstavo se bomo omejili na bijektivne povezave, čeprav bi lahko v splošnem bile povezave poljubne (poljubne slovenske besede in besedne zveze povezane s poljubnimi besedami ter besednimi zvezami v angleščini). Vseh povezav ne moremo odkriti z gotovostjo, sistem postavi parametrični model

$$P(f, a|e) \tag{1}$$

;kjer je poravnava a skrita. želeno verjetnost

$$P(f|e) \tag{2}$$

lahko dobimo kot

$$\sum P(f, a|e) \tag{3}$$

vsota vseh poravnav e in f.

V našem sistemu uporabimo ta postopek samo za prva dva modela (prevedba besed ter lokalna poravnava), pri ostalih modelih pa aproksimiramo

$$P(f|e) \tag{4}$$

po naslednjem postopku: med učenjem najdemo najverjetnejšo poravnavo â in izračunamo vsoto

$$\sum P(f, a|e) \tag{5}$$

v ozki okolici â. Ob dekodiranju uporabimo:

$$P(f|e) \tag{6}$$

2.3. The Candide system

Razdelek prikazuje sistem *Candide*, ki so ga v začetku devedesetih let razvili pri IBM. Razdeljen je na štiri sklope, prvi uvaja v teorijo verjetnostnih modelov, drugi prikazuje shemo dekodiranja, tretji načrta osnove modeliranja jezika ter zadnji, najpomembnejši, prikazuje prevajalne modele razvite v okviru projekta Candide. Opisan je še dodatni prevajalni model HMM, ki ni del osnovnega sistema, razvitega pri IBM. Ta model se je pri testiranjih veliko boljše obnesel in nove različice orodij uporabljajo ta model namesto IBM-2. Statistično strojno prevajanje do sedaj še ni doseglo rezultatov, ki bi omogočali izdelavo komercialnega prevajalnega sistema oziroma izdelavo uporabnega prevajalnega sistema. V začetku devetdesetih so pri IBM zaključili s projektom, ki je obrodil kar nepričakovano dobre rezultate. Temeljil je na avtomatični statistični analizi dvojezičnih besedil, rezultati in zaključki so opisani v (Brown et al., 1993). Poimenovali so ga "*The Candide system for machine translation*".

Oglejmo si primer prevoda besedila v nekem jeziku, ne bodimo skromni, vzemimo slovenski jezik, v besedilo v angleškem jeziku. Za poved *f* v slovenskem jeziku si zamislimo, da je bila zgrajena iz pripadajoče povedi *e* v angleškem jeziku. Angleška poved je prepotovala šumni komunikacijski kanal z zanimivo lastnostjo, vsako angleško poved prevede v slovensko. Osnovna ideja sistema Candide je, da lahko eksperimentalno določimo lastnosti našega "kanala"in jih lahko zapišemo s pomočjo matematičnih pravil. S $P(e|f)$ zapišemo verjetnost, da je bila *e* izvorna angleška poved, ki je služila za sestavo slovenske povedi *f*. Pri dani slovenski povedi *f* postane naš problem, problem avtomatskega prevajanja, iskanje angleške povedi, ki maksimira $P(e|f)$. Torej iščemo:

$$\hat{e} = \arg\max P(e|f) \tag{7}$$

Z uporabo Bayesove formule dobimo:

$$\hat{e} = \arg\max P(e|f) = \arg\max P(f|e)P(e) \tag{8}$$

S $P(f|e)$ zapišemo verjetnost da dobimo *f* kot izhod, če je *f* vhod našega prevajalnega kanala. Funkcijo bomo poimenovali prevajalni model (translation model). $P(e)$ predstavlja apriorno verjetnost, da se je poved *e* pojavila na vhodu prevajalnega kanala, to funkcijo poimenujemo jezikovni model (language model). Obe funkciji neodvisno porajata rezultata za kandidata za angleški prevod *e*. Prevajalni model zagotavlja, da besede povedi *e* izražajo vsebino zapisano v *f*, jezikovni model zagotavlja, da je *e* res poved. Candid izbere takšno poved *e*, ki maksimizira produkt prej opisanih funkcij. V nadaljevanju si bomo podrobneje ogledali odgovora na vprašanji kako sestavimo opisana modela ter kako naj pregledamo vsa angleške besede pri postavljanju rešitve - *e*.

2.3.1. Verjetnostni modeli

Verjetnostni model je matematična formula, ki dovolj verno opisuje neko zapažanje. Pojem razširimo z uvedbo parametrov v parametrizirani verjetnostni model, parametri omogočajo prireditev modela določeni podatkovni domeni. S *c* označimo telo podatkov, ki jih modeliramo, *Q* pa vektor parametrov. Verjetnost $P(c)$, ki jo izračunamo po neki vnaprej definirani formuli, ki je odvisna od *c* in *Q*, imenujemo maksimalna podobnost (maximum likelihood) *c*. Predstavlja verjetnost, ki jo določa naš model na opazovanih podatkih *c* ter parametri *Q*. Problem učenja parametričnega modela na podatkih *c* je enostavno iskanje maksimuma $P(c)$. Iskanje *Q* je primer optimizacije z omejitvami, omejitve so definirane z modelom, iščemo *Q*, ki določa maksimum funkcije. Pogosto iščemo več kot le verjetnostni model opazovanih podatkov *c*. Iščemo možno skrito statistiko *h*, ki je odvisna od *c*, in je ne moremo direktno določiti. *h* je v splošnem podmnožica *H*, vseh dovoljenih vrednosti. V takih primerih najprej postavimo parametrični model $P(c, h)$, nato postavimo vektor Q tako, da dobimo maksimalno podobnost:

$$\sum_c P(c) \tag{9}$$

Na žalost pa rešitev, vektor *Q*, ni vedno enolično določljiv. Ponavadi dobimo seznam odvisnosti med parametri ter podatki. Ponavadi uporabimo iterativni postopek Expectation Maximization (EM algoritem), ki je recept za izračun sekvence vektorjev parametrov Qi. Za to sekvenco lahko dokažemo, da ob izbranih pogojih, ponuja izboljšavo rešitve v vsakem koraku:

$$\sum P_{\theta_{i+1}}(c) \geq P_{\theta_i}(c) \tag{10}$$

Takšno nastavitev parametrov imenujemo EM-učenje. (c)e nekaj besed o uporabi EM metode v našem primeru. Izdelali bomo model prevajanja

$$P(f|e) \tag{11}$$

ter model jezika

$$P(e) \tag{12}$$

Najenostavnejši prevajalni model bi sestavili kot enostavno prevajanje slovenskih besed v angleške. Takšno prevajanje le slabo odraža prevode iz realnega sveta, besedni red se spreminja, med prevodom se porajajo besede in besedne zveze ter nekatere tudi izginjajo. Tako bomo postavili ogromno parametrično enačbo $P(f|e)$ za model prevajanja. Enačbo bomo postavili s pomočjo EM-učenja na dvojezičnem, vzporednem, slovensko/angleškem korpusu.

Parametre enačbe si bomo podrobneje ogledali v nadaljevanju. Parametrično enačbo bomo postavili tudi za model jezika

$$P(e) \tag{13}$$

EM postopek bomo opravili na angleškem besedilu (postavili bomo verjetnosti pojavljanja vseh besed).

2.3.2. Dekodiranje

Iskanje angleških besed, ki maksimirajo enačbo (9), brez omejitev, torej preiskovanje celotnega prostora angleški besed je prehud problem za še tako dobre računalnike. Tudi omejitev dolžine besed na še sprejemljivo mejo, ki bi sicer zmanjšala prostor, še vedno ne zadošča, besed je še vedno preveč. Uporabmo stack decoding algoritem (Jelinek, 1969), ki se uporablja pri razpoznavi govora.

2.3.3. Modeliranje jezika

Naj bo *e* niz angleških besed $e_1...e_l$ Model jezika ponuja verjetnost, da *e* predstavlja gramatično in semantično pravilno tvorjeno poved. Verjetnost zapišemo:

$$P(e) = P(e_1...e_l) =$$

$$P(e_1)P(e_2|e_l)P(e_3|e_le_2)...P(e_l|e_1..e_{l-1}) \tag{14}$$

Izračunati moramo vseh *l* verjetnosti.

Z $|\xi|$ označimo velikost angleškega besednjaka. Izračunati želimo verjetnosti za vse fraze dolžine *k* naraste na nepreglednih $|\xi|^{k-1}$. Kot primer vzemimo velikost besednjaka 10000 besed, kar je zelo majhna številka za vsak naravni jezik ter velikost fraz na 10 (k=10). (c)tevilo vseh fraz naraste na 1000010 = 10000000000000000000000000000000000000000000000000 (preveč ničel).

Sistem Candide uporablja model trigram, ki uporablja aproksimacijo:

$$P(e_k|e_1..e_{k-1}) \approx P(e_k|e_{1-2}..e_{k-1}) \tag{15}$$

Omejili smo zgodovino na dve besedi, trojko $e_{k-1}e_{k-1}e_k$ imenujemo trigram. Ostaja izračun $P(e_k|e_{1-2}..e_{k-1})$. Postaviti moramo tabelo verjetnosti za vsako trojico. Preiskati moramo učni korpus *c*, prešteti pojavitev vseh trigramov ter izračunati verjetnost pojavitve za vsak trigram. Za že tako kratko zgodovino pa pogosto naletimo na trigrame, ki se ne pojavljajo v učnem korpusu. Največje število trigramov v učnem korpusu je le $|c|$ (če bi se vsak trigram pojavil le enkrat), število vseh možnih trigramov pa je $r - 1$, ki je še vedno veliko večje število za vsak primerno velik besednjak.

Uporabimo tehniko deleted interpolation (Merialdo, 1992).

Izrazimo $P(e_k|e_{1-2}..e_{k-1})$ kot linearno kombinacijo:

- verjetnosti trojk $T(e_k|e_{1-2}..e_{k-1})$ (trigram probability)
- verjetnosti parov $D(e_k|e_{k-1})$(bigram probability)
- verjetnosti samih besed $D(e_k)$ (unigram probability)
- uniforme verjetnosti $1/|\xi|$

Distribuciji *B* in *U* lahko dobimo s preštevanjem pojavitev parov ter posameznih besed v učnem korpusu. (c)tevilo možnih parov nad nekim besednjakom je veliko manjše kot število trojk, torej je tudi možnost, da je nek par prisoten v našem korpusu večja kot za trojko, seveda je verjetnost pojavitve posamezne besede v učnem korpusu primerno večja. Model trojk ne omogoča upoštevanja semantičnih in sintaktičnih odvisnosti med besedami, ki so oddaljene za več kot dve mesti (ne sodijo v isto trojko). Pomagali si bomo z link grammar modelom (White et al., 1993). Ta model poskuša poiskati oddaljene povezave med besedami.

2.3.4. Modeliranje prevoda

Ta razdelek opisuje elemente prevajalnega modela (translation model) $P(f|e)$. Sistem Candide uporablja dva prevajalna modela in sicer že opisani model, ki temelji na EM-učenju ter model največje entropije (maximum-entropy model). Uporabljena različica EM modela temelji na petih začasnih modelih, rezultati učenja predhodnega modela se uporabljajo kot vhod v naslednji model, z rahlimi odstopanji. Zadnji model predstavlja že naučeni prevajalni model. EM algoritem gotovo pripelje do lokalnega maksimuma, ne zagotavlja pa globalnega maksimuma. Formulacija modela 1 (prevajanje besed) usmerja EM algoritem k globalnemu maksimumu. Izhod predhodnega modela predstavlja vhod naslednjega modela. Modeli so poimenovani z osnovnimi lastnostmi ter tudi s popularno različico imena, ki se je med uporabniki veliko bolje prijela (IBM-1, .. IBM-5).

2.3.5. Prevajanje besed, word translation (IBM-1)

Je najenostavnejši model, kaže verjetnosti posameznih besednih prevodov. Parameter tega modela je $t(f_i|e_i)$, verjetnost, da se določena slovenska beseda (f_i) prevede v angleško (e_i) . Vrednost za vsako besedo je na začetku postavljena na $1/|S|$, kjer z $|S|$ označimo velikost slovenskega besednjaka. Vse besede imajo na začetku enako verjetnost za prevod v določeno angleško besedo. Z iterativnim izvajanjem algoritma spreminjamo verjetnosti za posamezne besede (večamo pogojno verjetnost, če zasledimo obe besedi v dveh vzporednih povedih).

2.3.6. Lokalna poravnava, local alignment (IBM-2)

Ta model določa lego angleške besede v dvojezičnem korpusu, ki predstavlja prevod izbrane besede *f* iz slovenske vzporedne povedi. Vse besede, ki se porajajo brez osnov v drugem jeziku, označimo z vrednostjo lokalne poravnave null, nastajajo iz "ničte"besede v izvornem jeziku. Formalno zapišemo to verjetnost s tremi spremenljivkami:

$$P(a_j|j, m, l) \tag{16}$$

predstavlja verjetnost, da je mesto *j* v slovenski povedi dolžine *m* poravnano z lego a_j v neki angleški povedi dolžine *l*, ki je prevod prej opisane slovenske povedi.

2.3.7. Plodnost, fertility (IBM-3)

Ena sama angleška beseda lahko med prevajanjem "rodi"nič, eno ali celo več slovenskih besed. Vzemimo

primer ”It is correct.”ter slovenski prevod ”Pravilno je.”. Implicitno smo to dejstvo zajeli že s prejšnjim modelom, nov model nam eksplicitno določa verjetnost, da se neka angleška beseda prevede v določeno število slovenskih. Plodnost je število slovenskih besed, ki jih proizvede angleška beseda e_i ob prevodu, samo verjetnost pa zapišemo kot:

$$\Phi(n, e_i) \tag{17}$$

, ki podaja verjetnost, da je vrednost $\Phi(n, e_i)$enaka n.

2.3.8. Poravnava na osnovi razredov, class-based alignment (IBM-4)

V prejšnjem modelu lahko opazimo, da so ”plodnosti”besed odvisne od samih besed, poravnave pa ne. Model poravnava besede iz para $< e, f >$ne da bi se oziral na same besede. Nov model odpravlja pomanjkljivost s pomočjo parametrov, ki temeljijo na razredu besede *f*. Vse besede *f* iz slovenskega besednjaka ter vse besede *e* iz angleškega razvrstimo v razrede (naša različica je omejena na 50 razredov).

2.3.9. Poravnava brez neumnosti, non-deficient alignment (IBM-5)

Dva predhodna modela imata hudo pomanjkljivost, pripisujeta več kot ničelno verjetnost poravnavam, ki sploh ne ustrezajo slovenskim besedam. Na primer dve besedi lahko z enako verjetnostjo ležita na istem mestu v prevodu, besede ležijo pred začetkom in po koncu povedi. Zadnji model takšne nesmisle odkrije in odstrani.

2.3.10. Skriti Markovski model, Hidden Markov Model (HMM)

Definiramo poravnavo , ki določi besedo f_j na mestu *j* besedi e_i na mestu $i = a_j$. Verjetnost popravimo z uvajanjem škritih”poravnav $a_J = a_1 \dots a_j \dots a_J$. za vsak par povedi (f_J, e_I) . Za postavitev distribucije verjetnosti jo faktoriziramo prek celotne izvorne povedi ter se omejimo na odvisnosti prve stopnje. Dobimo naslednji model:

$$P(f_J|e_J) =$$

$$P(J|I) \sum_{a^i} \prod_{j=1}^{J} P(a_j, a_{j-1}, I, J) P(f_j|a_j) \tag{18}$$

kjer je:

- $P(J|I)$, verjetnost dolžine povedi
- $P(f|e)$, verjetnost pojavitve povedi *f*, če je izvorna poved *e* (izvorna poved šumnega kanala)
- $P(a_j, a_{j-1}, I, J)$, verjetnost poravnav

Z razširitvijo odvisnosti verjetnosti poravnav na razdalj $a_j - a_{j-1}$ z obsolutnih pozicij a_j dobimo homogeni, skriti Markovski model, krajše HMM (Vogel et al., 1996). HMM uspešno nadomesti model IBM-2, avtorji zatrjujejo, da so pri tetiranju novega modela dobili najboljše rezultate, če so izpustili še model IBM-3, vendar so bila ta testiranja nekoliko prirejena in vseeno svetujejo uporabo vseh IBM modelov razen IBM-2, ki ga nadomestimo z HMM. Nov model je, poleg še dodatnih izboljšav implementiran v novejši različici programa za gradnjo parametrični prevajalnih modelov GIZA++.

2.4. Egypt

Na poletni delavnici, leta 1999, na JHU (John Hopkins University) so po vzoru (Brown et al., 1993) izdelali zbirko orodij, ki omogočajo postavitev popolnega SMT sistema osnovanega na dvojezičnih vzporednih korpusih. Zbirko so poimenovali Egypt. Pri snovanju delavnice so si zadali pet osnovnih ciljev (vse cilje so izpolnili): Postavitev zbirke orodij za statistično strojno prevajanje, zbirka naj bo splošno dosegljiva raziskovalni srenji. Sestavljena naj bo iz orodij za pripravo korpusov, orodij za dvojezično učenje (postavitev parametričnih modelov) ter orodij za takojšnje dekodiranje tekstov. Postavitev češko-angleškega sistema za prevajanje besedil na osnovi izdelanih orodij. Osnovno testiranje sistema na snovi objektivnih mer (statistično modeliranje težavnosti). Izboljšanje osnovnih rezultatov z uporabo morfoloških in sintaktičnih prevajalnikov. V zadnjih dneh delavnice naj bi postavili prevajalni sistem za nek nov jezik v enem samem dnevu (potrditev enostavnosti uporabe orodij.). Vse zadane cilje so dosegli, še več, izdelali so še orodje za grafično pregledovanje poravnav (Cairo).

2.4.1. GIZA, GIZA++

Razdelek predstavlja orodje za povzemanje jezikovnih informacij iz dvojezičnega korpusa. Modul se imenuje GIZA in je osnovan na algoritmih in modelih predstavljenih v (Brown et al., 1993). Napisan je v programskem jeziku C++ in omogoča kar najhitrejše učenje prevajalnih. Osnovna različica, napisana na delavnici uporablja samo modele IBM-1,2,3, poznejše različice pa prinašajo implementacijo modelov IBM-4, IBM-5 ter z novim imenom GIZA++ še dodatnega modela, ki nadomešča osnovni IBM-2, HMM - skriti Markovski model. 4. Slovenščina Razdelek predstavlja uporabo motivacije pri izbiri izvornega jezika. Opisane so značilnosti slovenskega jezika in težave, ki izhajajo iz teh posebnosti. Te posebnosti so privedle do določenih omejitev v delovanju strežnika.

Slovenščina je slovanski jezik, je visoko pregiben in s skoraj prostim besednim redom. Večina funkcij, ki jih v Slovenščini izražamo s končnicami besed (pregibanje), se v Angleščini izraža z besednim redom in dodatnimi funkcijskimi besedami.

Kot primer navedimo osnovne značilnosti jezika. Kot glavno značilnost omenimo dvojino, ki nas loči tudi od večine slovanskih jezikov. Dvojina pri samem prevajanju ni problematična, če je le v korpusu dovolj povedi, ki jo uporabljajo. Večina samostalnikov lahko tvori edninsko, dvojinsko ter množinsko obliko v šestih sklonih. Večina pridevnikov lahko tvori 3 spole, vsa tri števila, 6 sklonov, 3 osnovne ravni primerjanja.

Slovenščina je jezik z mnogimi izpuščanji. To pomeni, da imajo osebni zaimek (jaz, on, oni) ponavadi nično obliko, so izpuščeni. V Slovenščini ni določnih in nedoločnih členov. V dokaz h kompleksnosti jezika je tudi velikost korpusa, Slovenščina ima kar 12 odstotkov manj besed v korpusu kot Angleščina.

2.4.2. Ostala enostavna orodja

Razdelek predstavlja skupek lastnih orodij, ki omogočajo enostavno povezovanje vseh sklopov sistema. Orodja so nastajala med snovanjem ter implementacijo sistema, predstavljajo zbirko enostavnih programčkov za obdelavo besedil ter zbirko skript za lažjo uporabo in avtomatizacijo uporabe orodij. Poleg orodij za delo z besedili je še orodje za testiranje kakovosti prevodov (evaluation tool), ki po metodi SA/TA (urejevalna razdalja, avtomatsko oceni kakovost prevodov.

- RemoveSGMLMarks prevede korpus iz TEI oblike s SGML zapisi v prosto nanizane povedi ločene z novimi vrsticami. Ta zapis bere orodje za pripravo korpusa Whittle. Kot parametre sprejme ime vhodne ter ime izhodne datoteke. Sestavi tudi posebno listo povedi sestavljenih iz lem IJS-ELAN korpusa.

- TestEditDistance izračuna urejevalno razdaljo po [9]. Med parametri navedemo metodo (simple ter translation) ter ime vhodne datoteke. Kot vhod sprejme tudi prevode s standardnega vhoda (STDIN). Vhodna datoteka je zgrajena iz referenčnih ter ocenjevanih prevodov. Rezultat je niz ocen za vsak par referenčni/ocenjevani prevod.

- MakeTranslations skripta prevede niz slovenskih povedi s pomočjo prevajalnega strežnika. Omogoča avtomatično prevajanje večjega števila povedi.

- EvaluateTranslations skripta sestavi referenčne prevode testnih primerov ter prevode namenjene ocenjevanju (nove prevode). Sestavi datoteko, ki je primerna kot vhod za TestEditDistance.

2.5. Izgled sistema

Razdelek predstavlja uporabo opisanih orodij pri snovanju novega strežnika. Strežnik omogoča prevajanje iz slovenskega v angleški jezik. Predstavljene bodo osnovne faze postavitve ter problemi, ki so se porajali med delom, ki izvirajo iz posebnih lastnosti slovenskega jezika. Te posebnosti so privedle do določenih omejitev v delovanju strežnika. Orisan je načrt postavitve sistema, faze priprave korpusa ter samega učenja. Prikazana je tudi uporaba samega sistema, dejansko prevajanje ter testiranje sistema, izdelava množice testnih prevodov ter njihovo ocenjevanje.

Slika 2 prikazuje celoten učni proces. Sestavlja ga več ločenih modulov, moduli so celo delo različnih razvijalcev. Na začetku potrebujemo dvojezični vzporedni korpus, naš sistem je osnovan na IJS-ELAN korpusu. Enostaven programček RemoveSGMLMarks poskrbi za pretvorbo TEI oblike v enostavno zaporedje povedi ločenih v dve datoteki (eno za izvorni jezik, drugo za ciljni). Povedi so zapisane vsaka v svoji vrstici, istoležne povedi so prevod iz enega jezika v drugi in obratno. Tako predelan vhodni korpus prevzameta dva sklopa, programi za postavitev jezikovnih modelov ter programi za postavitev prevajalnih modelov.

Jezikovni modeli nastajajo s pomočjo zbirke orodij za obdelavo besedil ter postavljanje jezikovnih modelov CMU Cambridge language modelling toolkit version 2, ki zgradijo jezikovni model angleškega jezika na osnovi angleškega dela korpusa. ˇ Prevajalni modeli nastajajo s pomočjo orodij zbirke Egypt. Ta del si bomo podrobneje ogledali.

Predelan korpus podamo kot vhod orodju whittle, ki omogoča razdelitev korpusa na učno ter testno množico ter predelavo v zapis, ki ga sprejme naslednji program v verigi GIZA++. Ta zapis je podan v obliki dveh osnovnih tipov datotek, prva vsebuje vse besede razporejene po frekvenci pojavljanj v korpusu. Besede so zapisane ko naravna števila, najpogosteje uporabljane besede imajo manjše število, ki jih opisuje. Druga datoteka predstavlja prepis osnovnega korpusa v obliko besed zapisanih s števili. Datoteke so ločene za testne ter učne primere.

GIZA++ je glavni modul sistema, omogoča izdelavo prevajalnih modulov.

Prevajalni moduli ter jezikovni modul predstavljajo zbirko tabel za preslikavo med slovensko ter angleško povedjo, predstavljajo tabele, ki jih uporablja dekoder pri sestavi prevodov.

Slika 3 predstavlja prehod učnega dela korpusa prek vseh modulov učnega sistema, izhod predhodnjega modula je vhod naslednjega. Vhodni korpus Whittle predela v obliko primerno za obdelavo s programom GIZA++. Ta modul sestavlja parametrične modele po [1] ter [10]. Izhod predhodnjega modela je vhod za naslednji model, izhod zadnjega modela je naučen sistem.

Slika 4 kaže potek posameznega prevajanja. Naučeni modeli so osnova za iskanje pravilnih prevodov vhodnih slovenskih povedi. ISI Rewrite Decoder preiskuje postavljene parametrične modele in sestavlja poved, ki jo ti modeli ocenjujejo kot najbolj verjetno. Vhodne povedi sprejema prek nastavljivih TCP/IP vrat, na istem naslovu se po poizvedbi nahaja tudi odgovor (angleški prevod vhodne povedi).

Do prevajalnega sistema lahko dostopamo prek skripte napisane v programskem jeziku perl ali prek spletnega vmesnika. Skripta omogoča avtomatizacijo prevajanja, primerna je za testiranje sistema ter modularno uporabo strežnika. Spletni vmesnik omogoča enostavno uporabo prevajalnega sistema praktično vsem uporabnikom, saj je dostop do strežnika s praktično vsakim spletnim brskalnikom.

Slika 5 kaže način izvedbe avtomatskega testiranja po metodi SA/TA opisani v 8. Testiranje je bilo izvedeno s pomočjo testih primerov, ki jih je izbral modul Whittle. Ti primeri vsebujejo tudi referenčne prevedbe, saj so sel dvojezičnega vzporednega korpusa.

Prevajalni modul se sprehodi prek vseh primerov ter prevede slovenske povedi na novem sistemu. Skupaj z izvornimi angleškimi povedmi iz korpusa sestavi spisek parov referenčna poved/prevedena poved ter ta spisek ponudi kot vhod modulu ta računanje urejevalne razdalje TestEditDistance.

Slednji izpiše ocene kakovosti prevedb za vsak par

2.6. Slovenščina → Angleščina

Razdelek predstavlja izbiro izvornega in ciljnega para jezikov. Oglejmo si najprej izvorni jezik, Slovenščino. Kot materin jezik avtorja tega dela je bil prvi na lestvici izbire, dodatni razlog pa je, da trenutno še ne obstaja strojni preva-

jalnik za ta jezik v poljubni tuj jezik. Z izbiro Slovenščine kot izvornega jezika želimo utreti novo pot k izdelavi prevajalnika komercialne kakovosti za pomembnejše tuje jezike. Razlogi za izbiro Angleščine kot ciljnega jezika so dovolj enostavni in nazorni. Angleščina je trenutno najširše uporabljan jezik v računalniškem svetu, smernice širitve jezika pa kažejo še večjo razširjenost. Pri izbiri je pomembno vlogo igrala pogostost uporabe tega jezika v SMT in pomembne izkušnje ostalih raziskovalcev tega področja ter vsa že opravljena testiranja programskih orodij.

2.7. Uporaba lem v dvojezičnem korpusu

Razdelek predstavlja poskus izboljšave osnovnega delovanja prevajalnega sistema s pomočjo uvajanja informacije o jezikih v obliki lem. Uspešnost novega postopka je natančneje prikazana v Napaka! Vira sklicevanja ni bilo mogoče najti..

Osnovni model je bil razširjen z uporabo beležk v korpusu, posebno z uporabo kontekstno neodvisnih lem, ki niso odvisne od bogate uporabe sklanjatev ter spreganja v slovenskem jeziku. Z uvajanjem lem smo želeli obiti največjo pomanjkljivost našega sistema, osnovane je na relativno majhnem korpusu.

Dvojezični vzporedni korpus ELAN, ki je osnova našega sistema, je že lematiziran. Naša hipoteza je predvidevala, da bo ta dodatna informacija pozitivno vplivala na kakovost prevodov.

Poleg osnovnih poravnanih, vzporednih povedi v obeh jezikih (Slovenščini in Angleščini) smo dodali še vzporedne ter poravnane povedi sestavljene iz lem osnovnih povedi. Oglejmo si primer uporabe:

live animal

ziv zival

all the animal of chapter 1 use must be wholly obtain

ves zival iz poglavje 1 morati biti v celota pridobiti

meat and edible meat offal

meso in uziten mesen klavnicen izdelek

manufacture in which all the material of chapter 1 and 2 use must be wholly obtain

izdelava pri kateri morati biti ves uporabljen material iz in poglavje v celota pridobiti

Tako smo dobili iz vsakega para slovenska poved/angleški prevod nov dodaten par slovenska poved zapisana s pomočjo kontekstno neodvisnih lem/angleški prevod zapisan s pomočjo kontekstno neodvisnih lem. Sistemu smo tako pomagali pri določanju nedoločnikov ter neprepognjenih in nesklanjanih besed.

Osnovni korpus smo razširili število povedi iz 27421 na 58803. Več kot dvakratna količina informacij naj bi povečala natančnost prevajanja, saj je bil osnovni korpus relativno majhen. (c)tevilo lem je večje od osnovnega korpusa, ker so tukaj upoštevane še leme testnih povedi, ki niso vključene v osnovnem korpusu.

Postavili smo nov sistem (ponovno učenje na novem, razširjenem korpusu) ter izvedli enaka testiranja kot pri osnovnem strežniku. Testiranja so obširneje predstavljena v

2.8. Testiranje

Poglavje predstavlja motivacijo za izvedbo testiranja sistema ter omejitve in sredstva, ki smo jih uporabili. Predstavljeni so tudi osnovni teoretični temelji.

Pri testiranju osnovnega ter popravljenega sistema smo se odločili samo za testiranje kvalitete prevodov, hitrost prevajanja oziroma odzivnost celotnega sistema pa prepustili poznejšim raziskavam in možnim izboljšavam.

V strojnem učenju se pojavlja ravno toliko metod za vrednotenje sistemov prevajanja kot je samih metod učenja (učenja strojnega prevajanja). Tolikšno število metod izvira iz dejstva, da se strokovnjaki le slabo strinjajo kaj sploh je dober prevod, kaj šele kaj je dobra mera za ocenitev prevoda. Vrednotenje MT sistemov je postalo samo zase dovolj močno področje razvoja MT, zaenkrat pa rezultatov, ki ne bi zbujali valov polemik, še ni.

Naša naloga pri izbiri metod je bila zapletena, še posebej z osnovo našega sistema, ki je vezan na slovenski jezik s svojimi posebnostmi in le slabo raziskan. Uporabili smo dve osnovni metodi preverjanja kakovosti prevoda, avtomatsko ter ”ročno”metodo, ocenjevanja prevodov s pomočjo strokovnjaka. Avtomatska metoda se še nadalje loči na dve podrazličici, ki pa se razlikujeta le v upoštevanju ene količine.

SSER, Subjective Sentence Error Rate (Vogel et al., 1996): Prevodi so rangirani v pet kakovostnih razredov od ”popoln prevod”, ki je ocenjen s 100 odstotki do ”popolna bedarija”, vredna 0 odstotkov.

- popoln prevod, 100 odstotkov
- dober prevod, 75 odstotkov
- prevod, 50 odstotkov
- zanič prevod, 25 odstotkov
- popolna bedarija, 0 odstotkov

Porazdeljevanje prevodov v razrede opravlja človek, po možnosti strokovnjak. Za preverjanje kakovosti ocenjevanja je uvedena še posebna skupina referenčnih prevodov, ki se prav tako razvrščajo v razrede.

SA/TA, enostavna/preslikav natančnost (Alshawi et al., 2000) in (Vogel et al.,): Za vsak prevod je fizračunana urejevalna razdalja (edit distance), število vrinjenih, brisanih ter zamenjanih besed, med vrednotenim ter referenčnim prevodom. Ta razdalja je še utežena z dolžino povedi. Uporabili smo dve različni izvedenki te osnovne metode po (Alshawi et al., 2000):

SA, simple accuracy, enostavna natančnost

$$SA = 1 - (I + D + S)/R \tag{19}$$

Kjer je *I* število vrinjenih besed (Inserted), *D* število brisanih (Deleted), *S* število zamenjanih besed (Substituted) in R dolžina referenčne povedi (Reference length).

TA, translation accuracy, natančnost preslikav

SA = 1 - (I’ + D’ + S + T)/R

Kjer je *I’* število vrinjenih besed (Inserted), *D’* število brisanih (Deleted), *S* število zamenjanih besed (Substituted), *R* dolžina referenčne povedi (Reference length), če upoštevamo še število premeščanj *T* (Transposition). Po (Alshawi et al., 2000) je natančnost preslikav bolj primerna mera za opisovanje preslikav, saj pravilne besede

na napačnih mestih štejejo kot ena sama napaka in ne kot dve (brisanje besede ter vrivanje na pravo mesto).

Metodi sta bili še dodatno testirani s pomočjo zbirke prevodov (ročni prevodi testne množice). Ta množica prevodov je bila postavljena kot idealni prevodi in njena ocena predstavlja oceno h kateri stremimo. Tako smo vse rezultate normalizirali s pomočjo ocene testne množice. Testna množica je bila ocenjena z metodo SA/TA s koeficientom 2,82 oziroma s 36,84 odstotki. (36,84

Tabela 1: rezultati testne skupine z metodo SA/TA

	SA/TA		
	povprečje	odstotki	st. dev.
testna skupina	2,47	36,84	1,38

Dodatno normaliziranje smo uporabili s pomočjo ročnega ocenjevanja testne množice prevodov. Testna množica je bila ocenjena z metodo SSER s koeficientom 4,48 oziroma s 87 odstotki. (87 odstotkov predstavlja 100 odstotkov)

Tabela 2: rezultati testne skupine z metodo SSER

	SA/TA		
	povprečje	odstotki	st. dev.
testna skupina	4,48	87,00	0,65

Rezultati, prikazani v tabelah, ki so normalizirani z opisanim postopkom so posebej označeni in komentirani.

Pri testiranju je bila pri avtomatski metodi uporabljena metoda desetkratnega prečnega preverjanja (tenfold cross validation). Desetina korpusa se odredi za testne namene, devet desetin pa za učenje modelov. Testiranje s pomočjo opisanih metod se izvaja s testnimi primeri na naučenih modelih. Postopek se ponovi še devetkrat, tako so vsi primeri korpusa prisotni tako v testni kot v učni množici. Razdeljevanje na testne ter učne primere (pare slovenska/angleška poved) omogoča Whittle, orodje za razdelitev korpusa na učne ter testne primere ter za pripravo korpusa za program GIZA. Metoda SSER, ki zahteva prisotnost eksperta, bi bila za celotno desetkratno prečno preverjanje preveč zamudna. Opravili smo le nekaj deset ocenjevanj obeh sistemov (na manjši množici testnih primerov, okrog 100 primerov).

V pomoč testiranju, predvsem ročnemu delu, je bil izdelan dodatek k osnovnemu spletnemu vmesniku sistema za prevajanje, ki je omogočal izbiro povedi ter zapis ocen v podatkovno bazo.

Rezultati so razdeljeni so na preverjanje kakovosti prevajanja osnovnih algoritmov ter izboljšane različice.

Testiranje je bilo izvajano s pomočjo testnih primerov, ki niso del učnega korpusa. Isti testni primeri so bili uporabljeni za oba sistema, navadnega ter sistema, ki uporablja leme. Za preverjanje metod je bila izdelana še dodatna množica umetnih testnih primerov ter zbirka umetnih referenčnih prevodov.

2.9. Rezultati

Poglavje predstavlja rezultate testiranja po posameznih kategorijah in kriterijih. Zapisane so tudi razlage rezultatov in možne izboljšave.

Testiranje je potekalo v dveh stopnjah, najprej testiranje kakovosti prevodov osnovnega sistema, v nadaljevanju pa še testiranje novega sistema. Rezultati so med seboj primerljivi, same metode pa niso primerljive z metodami ostalih avtorjev.

2.9.1. Predstavitev rezultatov

SA/TA avtomatska metoda je bila izvedena na 519 primerih, preverjanje s testno množico pa na 100 primerih.

SSER metodo je izvajalo deset izvedencev različnih izobrazb (vsi vsaj z univerzitetno izobrazbo). Vsak izvedenec je ovrednotil 100 prevedb osnovnega sistema ter 100 prevedb sistema z uporabo lem. Rezultati posameznih ovrednotenj so podani v Tabela 2.

2.9.2. Rezultati vrednotenja osnovnega sistema

Prevodi osnovnega sistema so kar vzpodbudni. Veliko prevodov je popolnoma uporabnih, te so eksperti pri metodi SSER ocenili s 4 ali celo 5, seveda pa je kar veliko tudi popolnoma zgrešenih prevodov (ocenjeni z 1 po SSER metodi).

Torej je osnovna metoda kar obetajoča in primerna tudi za slovenski jezik. Do sedaj je bila preizkušena že na drugih jezikih in rezultati so bili prav tako obetavni.

Opazili smo vseeno še velik prostor za izboljšanje, saj rezultati še niso na ravni komercialnih izdelkov (Systran in podobni), primerjava s takšnimi izdelki pa tudi ni popolnoma vmesna, saj komercialnega prevajalnika iz slovenščine v angleščino sploh še ni.

2.9.3. Rezultati vrednotenja sistema z uporabo kontekstno neodvisnih lem

Na žalost se je metoda uporabljanja kontekstno neodvisnih lem za dodatno opisovanje jezika izkazala kot neuporabno. Rezultati ocenjevanja prevodov so pri obeh sistemih približno enaki. Zavedati se moramo, da je učenje sistema z lemami veliko bolj kompleksno, saj je število učnih primerov v korpusu več kot podvojeno, večje število učnih primerov pa lahko skriva tudi veliko več napak.

Rezultati podani v tabelah kažejo le povprečne vrednosti mnogih prevedb. Pri natančnejšem pregledu posameznih prevedb opazimo, da se prevedbe obeh sistemov precej razlikujejo. Mnoge prevedbe so v novem sistemu izboljšane, na žalost pa so se poslabšale tudi mnoge dobre prevedbe starega sistema.

Ta zapažanja nam vseeno vzbujajo določeno mero zaupanja v novo metodo. Potrebujemo le ločevanje pozitivnimi ter negativnimo lastnostmi nove metode. Popravljena metoda, ki bi zadržala izboljšane primere in ne bi uvajala novih napak v prevajalni sistem, bi temeljila na podmnožici kontekstno neodvisnih lem.

Verjetno bi bile prevedbe sistema, zgrajenega z osnovnim modelom ter nadgrajenega le z izbranimi vrstami lem, boljše. Izbira te podmnožice pa presega okvire tega dela in je natančneje predstavljena v 9. poglavju.

2.10. Zaključek in nadaljnje delo

Poglavje predstavlja avtorjeva razmišljanja ob rezultatih. Nadaljuje pa z načrti za nadaljnje delo.

Osnovni strežnik se je pokazal kot zelo dobra osnova za testiranje novih idej. Osnovni algoritmi omogočajo dobre prevode. Prenos orodij na slovensko-angleški prevajalnik je bil uspešen, algoritmi se obnašajo v okviru pričakovanj.

Metoda razširitve osnovnih algoritmov z uporabo kontekstno neodvisnih lem se je izkazala v osnovi za neuporabno. Rezultati pa vseeno kažejo na možnosti izboljšave tudi v tej smeri, saj so bile mnoge prevedbe z novim sistemom izboljšane. Nov sistem pa je pokvaril nekatere že dobre prevode.

Prva možnost za izbiro je lematiziranje samo odprtih besednih vrst (oznake iz IJS-ELAN Vm/N/A): glavni glagoli, vsi samostalniki ter pridevniki. Ostale besedne vrste, predvsem veznike ter zaimke (oznake iz IJS-ELAN Vc/P) pa izpustimo.

Poleg uporabe lem za dograditev sistema se poraja kot najočitnejša možnost uporabe dvojezičnega enosmernega slovarja (slovensko - angleški slovar). Tako bi se na eleganten in enostaven način izognili vsem nepoznanim besedam v prevodih .

3. Literatura

Brown, Peter, Peter Cocke, Stephen Della Pietra, Vincent Della Pietra, Fredrik Jelinek, John Lafferty, Robert Mercer, Paul S. Roossin 1994. : A statistical approach to machine translation. *omputational linguistics*, C 16(2).

Fredrik Jelinek 1969. A fast sequential decoding algorithm using a stack. *IMB Journal of research and development*, pp 675-685.

Edward Sapir 1921. Language: an introduction to the study of speech. *IHartcourt Brace*, pp 675-685.

Bernard Merialdo 1992. Tagging text with a probabilistic model. *Proceedings of the 1992 AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pp 675-685.

John White, Theresa A. O'Connell, Lynn M. Carlson 1993. Evaluation of machine translation. Human Language Technology. *Morgan Kaufman Publishers*, pp 206-210

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

Bernard Merialdo 1990. Tagging text with a probabilistic model. *Proceedings of the IBM Natural Language ITL*, pp161-172.

Stephan Vogel, Fraz Josef Och, Christof Tillmann, Sonja Niessen, Hassan Sawaf, Hermann Ney . Statistical methods for machine translation. *Lehrstuhl fur informatik VI, Computer Science Department RWTH Aachen university of technoogy*

Paul C. Davis, Chris Brew 2002. Stone Soup Translation. *The 9th Conference on Theoretical and Methodological Issues in Machine Translation, Keihanna, Japan*

Hijan Alshawi, Srinivas Bangalore, Shona Douglas 2000. Learning dependency translation models as collections of finite state head transducers. *Computational linguistics*, 26(1):45-60.

S. Vogel, H. Ney, C. Tillmann 1996. HMM-based word alignment in statistical translation. *In COLING '96: The 16th Int. Conf. On Computational Linguistics*, p. 836-841.

Tabela 3: primerjava ocen osnovnega sistema, sistema z lemami ter testne množice prevodov; primerjava je bila izvedena z obema metodama dodane so popravljene vrednosti z popravkom glede na rezultate ocen testnih prevodov.

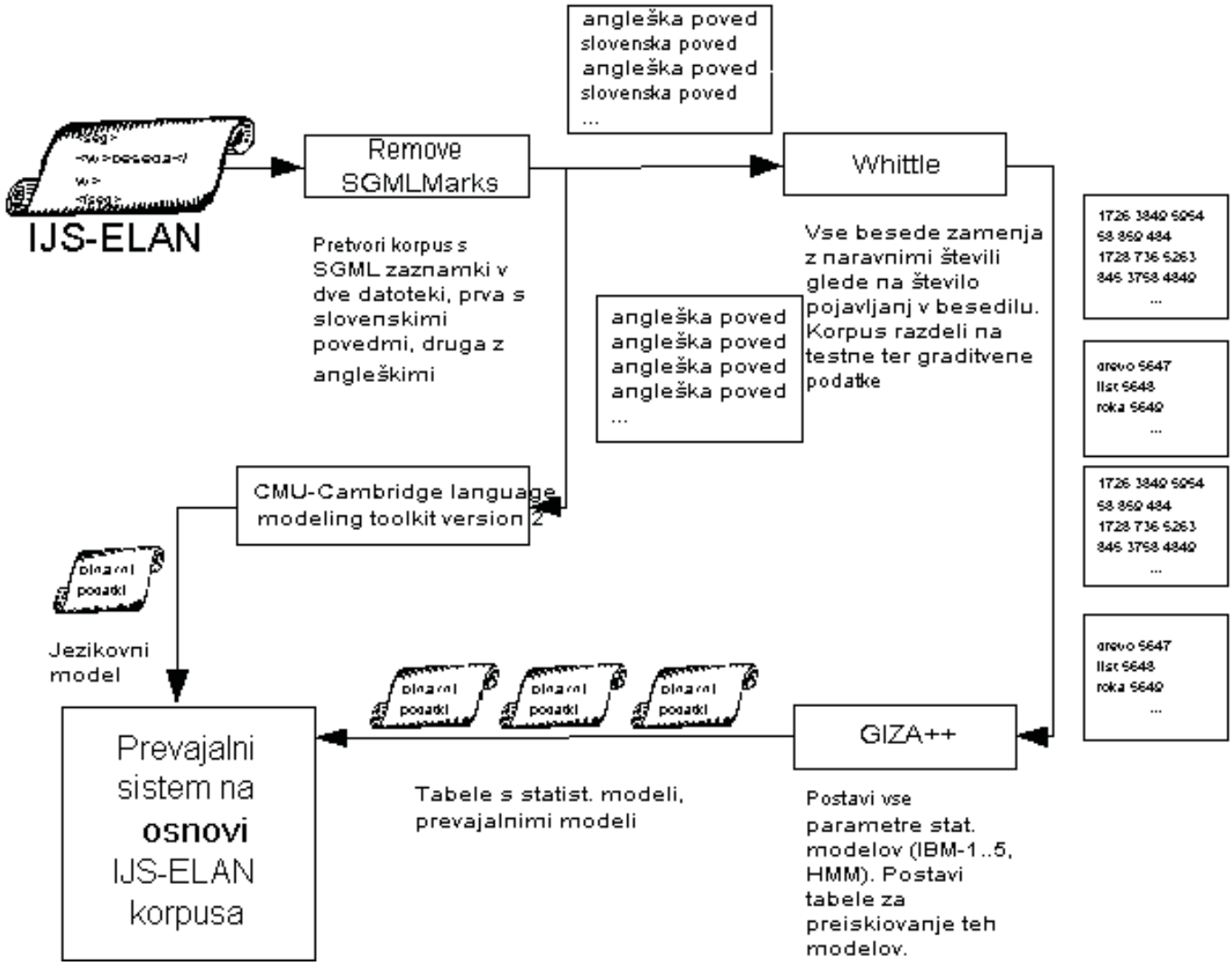
- *ocena osnovnega sistema je normalizirana z rezultatom testne skupine pri metodi SA/TA, testna skupina je upoštevana kot najboljši možen prevod
- **ocena sistema z lemami je normalizirana z rezultatom testne skupine pri metodi SA/TA, testna skupina je upoštevana kot najboljši možen prevod
- ***ocena osnovnega sistema je normalizirana z rezultatom testne skupine, testna skupina je upoštevana kot najboljši možen prevod, upoštevan je še popravek SSER ocene testne skupine
- ****ocena osnovnega sistema je normalizirana z rezultatom testne skupine, testna skupina je upoštevana kot najboljši možen prevod, upoštevan je še popravek SSER ocene testne skupine

	SA/TA			SSER		
	povp.	odstotki	st.dev.	povp.	odstotki	st.dev.
Osnovni sistem	1,40	9,97	0,71	1,94	23,55	1,25
Lemma	1,40	9,96	0,70	1,91	22,77	1,31
testna skupina	2,47	36,84	1,38	4,48	87,00	0,65
Osnovni sistem*	2,82	27,07	na	2,17	27,06	na
Lemma**	2,82	27,04	na	2,13	26,17	Na
Osnovni sistem***	3,15	31,12	na	2,42	31,11	Na
Lemma****	3,15	31,08	na	2,38	30,08	Na

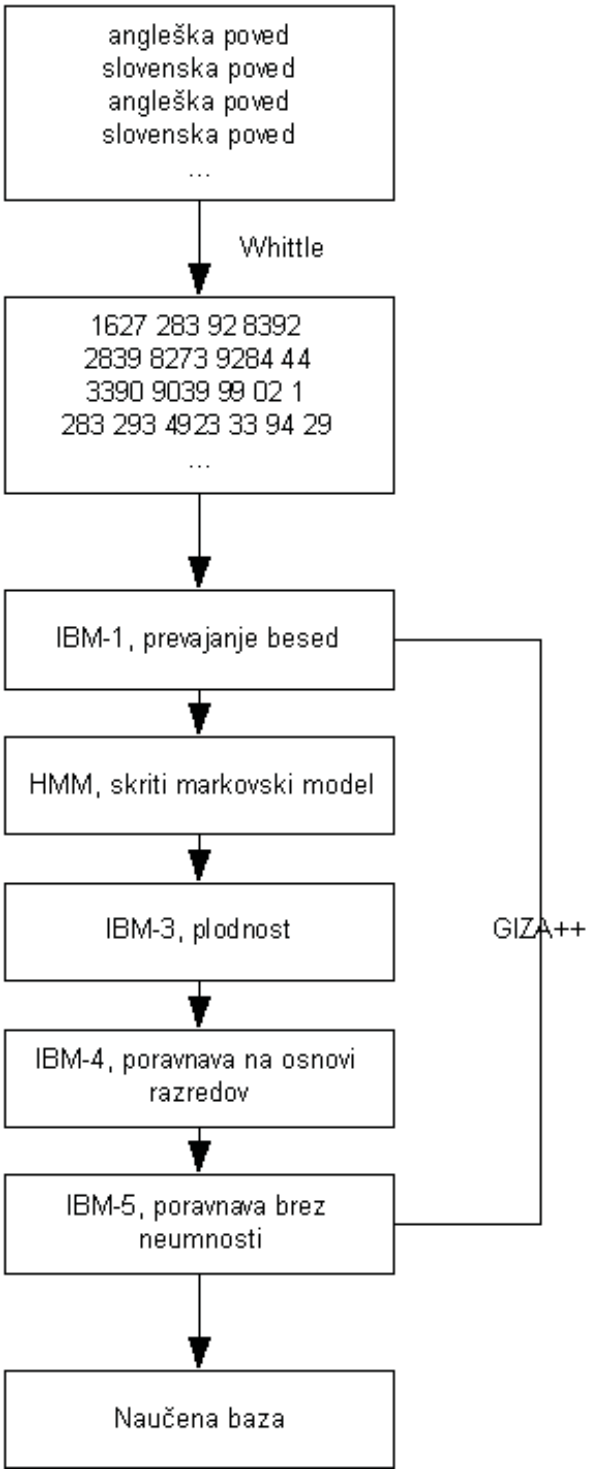
Tabela 4: del rezultatov ocenjevanja testnih primerov z metodo SSER

- *ocena osnovnega sistema je normalizirana z rezultatom testne skupine pri metodi SSER, testna skupina je upoštevana kot najboljši možen prevod
- **ocena sistema z lemami je normalizirana z rezultatom testne skupine pri metodi SSER, testna skupina je upoštevana kot najboljši možen prevod

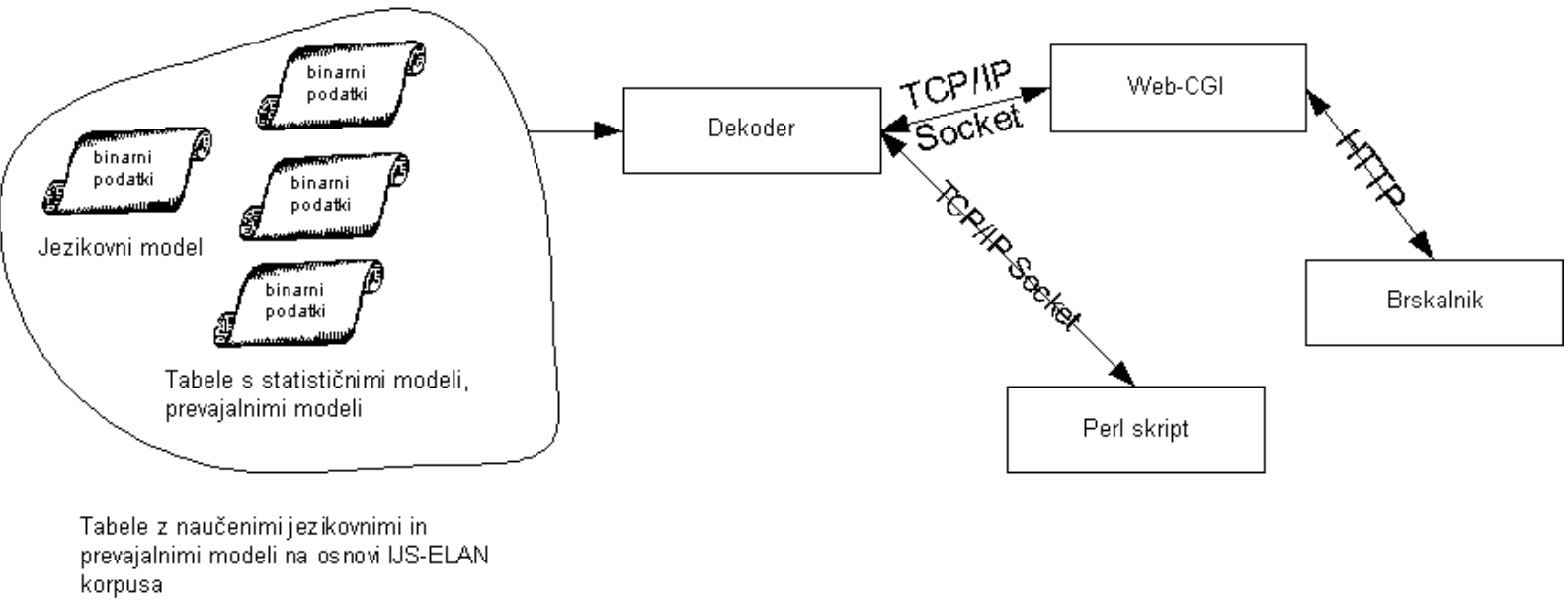
	E1 - lemma	E1 - normal	E2 - lemma	E2 - normal	E3 - lemma	E3 - normal
Povprečje	1,86	1,79	2,02	2,01	1,85	2,02
std. Deviacija	1,30	1,29	1,35	1,25	1,30	1,26
Lemma	1,40	9,96	0,70	1,91	22,77	1,31
odstotki	21,51	19,77	25,58	25,29	21,22	25,58
Povprečje*	2,08	2,00	2,26	2,25	2,06	2,26
V odstotkih**	24,01	22,06	28,55	28,23	23,68	28,55
odstotek ocen 1	56	61	49	44	59	48
odstotek ocen 2	24	21	25	33	21	21
odstotek ocen 3	5	5	8	8	3	17
odstotek ocen 4	5	2	5	5	8	5
odstotek ocen 5	10	10	11	9	8	8



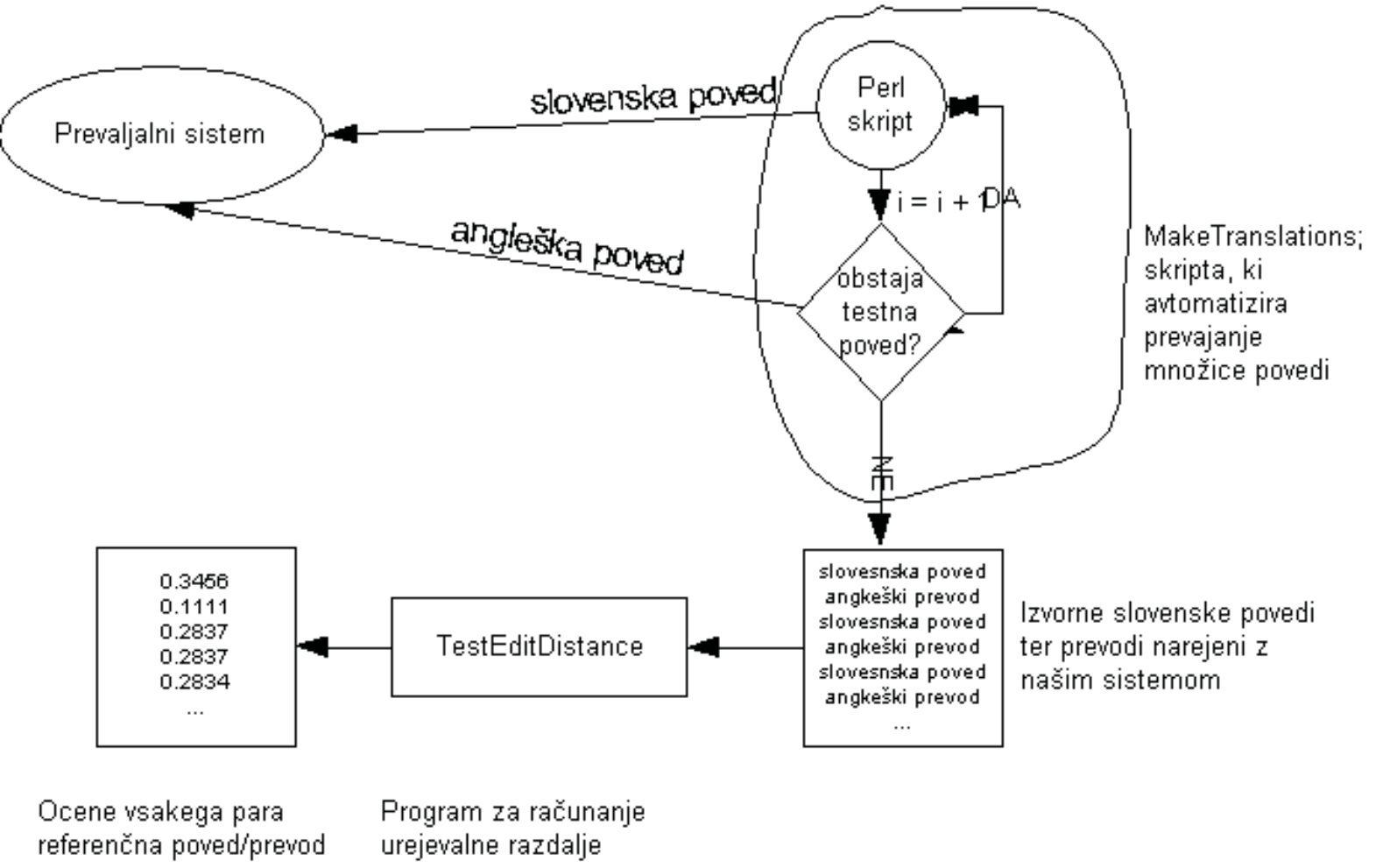
Slika 2: prikaz celotnega učnega procesa



Slika 3: prikaz prehoda učnih primerov pri učenju prevajalnih modelov



Slika 4: dekoder s pomočjo statističnih modelov prevaja dostavljene vhodne povedi



Slika 5: skripta dostavlja testne povedi naučenemu sistemu, zbrani prevodi so ocenjeni z metodo ureditvene razdalje