Vsak začetek je težak: avtomatsko učenje prevajanja slovenščine v angleščino

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

Prispevek predstavlja poizkus avtomatskega prevajanja iz slovenskega jezika v angleški na osnovi statističnega strojnega prevajanja. Sistem temelji na zbirki orodij EGYPT, ki je namenjena obdelavi dvojezičnih vzporednih korpusov za strojno prevajanje. Osnova za učenje prevajanja je bil stavčno poravnani korpus IJS-ELAN, ki vsebuje milijon besed, prevodov iz slovenščine v angleščino in obratno; besede obeh jezikov so tudi oblikoslovno označene. V članku predstavimo statistične osnove sistema, zbirko orodij EGYPT in našo implementacijo prevajalnika. Sistem smo učili najprej neposredno na besedah (besednih oblikah) v korpusu, nato pa smo jih, za slovenski jezik, nadomestili z besednimi lemami, s čimer smo se želeli izogniti problemu redkih podatkov. Izvedeno je bilo osnovno vrednotenje sistema, tako za model z besednimi oblikami, kot za tistega z lemami. Vrednotenje smo je izvedli z dvema metodama: SA/TA, ki je različica urejevalne razdalje (edit distance), in omogoča avtomatsko vrednotenje; SSER (subjective sentence error rate), kjer prevode našega sistema ocenjujejo ljudje z razvrščanjem v kategorije. Prispevek zaključimo z načrti za nadaljnje delo.

1. Uvod

V prispevku predstavljamo prvi sistem za avtomatsko prevajanje iz slovenskega v angleški jezik na osnovi statističnega strojnega prevajanja (SMT - Statistical Machine Translation).

SMT je mlada veja računalništva, saj je bila do sedaj večina računalnikov prešibkih za zahtevne obdelave velikih količin podatkov, ki so osnova vseh statističnih prevajalnih sistemov. Zapletene matematične osnove, temelj SMT, so prav tako odvrnile marsikaterega raziskovalca.

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 pojavijo razni problemi: sistemi imajo majhno pokritje, majhno toleranco za napake in nove oblike, in so dostikrat počasni. Naravni jezik je kompleksna in živa tvorba in ustrezna pravila za opisovanje so temu primerno zapletena, če jih je sploh mogoče vsa zapisati. Že v začetku tega stoletja so prišli jezikoslovci do zaključka, da vse gramatike puščajo ("All grammars leak"), (Sapir, 1921).

Pri statističnem modeliranju jezika namesto razdeljevanja stavkov po slovničnih pravilih iščemo splošne vzorce, ki se porajajo pri uporabi jezika. Glavna metoda za iskanje takšnih vzorcev je štetje raznovrstnih objektov in odkrivanje statističnih zakonitosti domene, npr. prevajanja. 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.

Zbiranje dovolj velikih in pravilnih podatkovnih baz učnih primerov, v našem primeru postavljanje dovolj velikih in primernih dvojezičnih vzporednih korpusov je dolgotrajno opravilo. Za jezikovni par slovenščina-angleščina je edini prosto dostopen korpus IJS-ELAN (Erjavec, 2002), ki vsebuje milijon besed, prevodov iz slovenščine v angleščino in obratno. Besede v korpusu so tudi oblikoslovno označene: vsaki je bila avtomatsko pripisana oblikoslovna oznaka in njena lema, t.j. osnovna oblika. Korpus je zapisan po standardu XML (eXtenden Markup Language), skladno s priporočili TEI P4 (Sperberg-McQueen and Burnard, 2000). V ilustracijo podamo v Sliki 1 poravnana segmenta (povedi) iz korpusa.

Članek predstavi poskus implementacije in ovrednotenja sistema za SMT, ki je bil učen na korpusu IJS-ELAN. V razdelku 2. predstavimo matematično ozadje statističnega strojnega prevajanja, v razdelku 3. opišemo programsko opremo, ki implementira SMT, in kako smo naučili slovensko-angleškega modela; razdelek 4 opiše poskus poboljšanja sistema, z uporaba lem v dvojezičnem korpusu; razdelek 5 ovrednoti sistem po dveh različnih metodah; in razdelek 6 poda zaključke in načrte za nadaljnje delo.

2. Ozadje statističnega strojnega prevajanja

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 (Brown et al., 1993; Brown et al., 1994). Sistem temelji na osnovi parametričnih statističnih modelov; prikažemo jih v tem razdelku.

2.1. Poravnava

Vzemimo par slovenske in angleške povedi s, e, ki sta ena prevod druge. Čeprav 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 sliki 2.

```
<seg id="anx2.en.91" corresp="anx2.sl.91">
<w ana="Ncns" lemma="pig">Pig</w>
<w ana="Ncns" lemma="fat">fat</w>
<c type="open" ctag="(">(</c>
<w ana="Sp" lemma="include">including</w>
<w ana="Ncns" lemma="lard">lard</w>
<c type="close" ctag=")">)</c>
<w ana="Cc-n" lemma="and">and</w>
<w ana="Ncns" lemma="poultry">poultry</w>
<w ana="Ncns" lemma="fat">fat</w>
<c ctag=",">,</c>
<w ana="Dg--p" lemma="other">other</w>
<w ana="Cs" lemma="than">than</w>
<w ana="Cs" lemma="that">that</w>
<w ana="Sp" lemma="of">of</w>
<w ana="Ncns" lemma="heading">heading</w>
<w ana="Dg" lemma="no">No</w>
<w type="dig" ana="Mc" lemma="0209">0209</w>
<w ana="Cc-n" lemma="or">or</w>
<w type="dig" ana="Mc" lemma="1503">1503</w>
<c ctag=":">:</c>
</seg>

<seg id="anx2.sl.91" corresp="anx2.en.91">
<w ana="Aopfsn" lemma="prašičji">Prašičja</w>
<w ana="Ccs" lemma="in">in</w>
<w ana="Aopfsn" lemma="piščančji">piščančja</w>
<w ana="Ncfsn" lemma="mast">mast</w>
<c ctag=",">,</c>
<w ana="Spsg" lemma="razen">razen</w>
<w ana="Pd-fsg----a" lemma="tisti">tiste</w>
<w ana="Spsg" lemma="iz">iz</w>
<w ana="Y" type="abbr">tar.</w>
<w ana="Y" type="abbr">št.</w>
<w ana="Mc---d" type="dig">0209</w>
<w ana="Ccs" lemma="ali">ali</w>
<w ana="Mc---d" type="dig">1503</w>
</seg>
```

Slika 1: Jezikovne oznake v korpusu

Slika 2: primer povezav med besedami v angleški ter slovenski povedi

Takšen niz povezav imenujemo poravnava (alignment). Formalna definicija poravnave: Niz parov s, e, kjer vsak par predstavlja povezavo med j-to besedo s (slovenska beseda) ter i-to besedo e (angleška beseda). Povezati želimo s_j ter e_i, kjer e_i ustreza s_j v angleščini.

2.2. Sistem Candide

Statistično strojno prevajanje sicer še ni doseglo rezultatov, ki bi omogočali izdelavo komercialnega (uporabnega) prevajalnega sistema, pa vendar so v začetku devetdesetih 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*".

Za poved s 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, da 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 $P(e|s)$ zapišemo verjetnost, da je bila e izvorna angleška poved, ki je služila za sestavo slovenske povedi s. Pri dani slovenski povedi s postane naš problem, problem avtomatskega prevajanja, iskanje angleške povedi, ki maksimira $P(e|s)$. Torej iščemo:

$$\hat{e} = \arg\max P(e|s) \tag{1}$$

Z uporabo Bayesove formule dobimo:

$$\hat{e} = \arg\max P(e|s) = \arg\max P(s|e)P(e) \tag{2}$$

S $P(s|e)$ zapišemo verjetnost da dobimo s kot izhod, če je e 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 s, 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.

V nadaljevanju so obdelani štirje sklopi na katerih temelji Candide. 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.

2.3. 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 (vse dovoljene statistike). V takih primerih najprej postavimo parametrični model $P(c,h)$, nato postavimo vektor Q tako, da dobimo maksimalno podobnost:

$$\sum_c P(c) \tag{3}$$

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 $P(s|e)$ 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

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

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

2.4. 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 števila besed na še sprejemljivo mejo, ki bi sicer zmanjšala prostor, še vedno ne zadošča, besed je še vedno preveč. Uporabimo dekodiranje s skladom štack decoding"algoritem, ki se uporablja pri razpoznavi govora.

2.5. Modeliranje jezika

Naj bo e niz angleških besed $e_1 \dots e_l$ Model jezika ponuja verjetnost, da e predstavlja slovnično in semantično pravilno tvorjeno poved.

Z $|\xi|$ označimo velikost angleškega besednjaka. Izračunati želimo verjetnosti za vse fraze dolžine k, število izračunov 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 10 $(k = 10)$. Število vseh fraz naraste na 10000^{10}.

Sistem Candide uporablja model trigram, model trojk. Preiskati moramo učni korpus c, prešteti pojavitve 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).

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.

2.6. Modeliranje prevoda

Ta razdelek opisuje elemente prevajalnega modela (translation model) $P(s|e)$. 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. EM algoritem gotovo pripelje do lokalnega maksimuma, ne zagotavlja pa globalnega maksimuma. Formulacija modela 1 (prevajanje besed) usmerja EM algoritem k globalnemu maksimumu. Modeli so poimenovani z osnovnimi lastnostmi ter tudi s popularno različico imena, ki se je med uporabniki veliko bolje prijela (IBM-1, ... IBM-5).

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

Je najenostavnejši model, kaže verjetnosti posameznih besednih prevodov. Parameter tega modela je $t(s_i|e_i)$, verjetnost, da se določena slovenska beseda (s_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).

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

Ta model določa lego angleške besede v dvojezičnem korpusu, ki predstavlja prevod izbrane besede s 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:

$$P(a_j|j,m,l) \tag{5}$$

Formula 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.

2.9. 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.

2.10. 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 $< e, s >$ ne da bi se oziral na same

besede. Nov model odpravlja pomanjkljivost s pomočjo parametrov, ki temeljijo na razredu besede s. Vse besede s iz slovenskega besednjaka ter vse besede e iz angleškega razvrstimo v razrede (naša različica je omejena na 50 razredov).

2.11. Poravnava brez nesmislov, 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.

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

Definiramo poravnavo , ki določi besedo s_j na mestu j besedi e_i na mestu $i = a_j$. Verjetnost popravimo z uvajanjem "skritih" poravnav $a_J = a_1 \dots a_j \dots a_J$. za vsak par povedi (s_J, e_I). Za postavitev distribucije verjetnosti jo faktoriziramo prek celotne izvorne povedi ter se omejimo na odvisnosti prve stopnje.

HMM uspešno nadomesti model IBM-2, avtorji zatrjujejo, da so pri testiranju 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++.

3. Programska oprema in učenje modela

Nadaljevanje prinaša prikaz trenutno najbolj obetajoče in vsesplošno priznane zbirke orodij za rokovanje z dvojezičnimi, vzporednimi korpusi: Egypt. 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.

3.1. 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).

3.2. GIZA

Orodje za povzemanje jezikovnih informacij iz dvojezičnega korpusa se imenuje GIZA in je osnovano na algoritmih in modelih predstavljenih v (Brown et al., 1993). Napisano je v programskem jeziku C++ in omogoča kar najhitrejše učenje prevodnih pravil. 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.

3.3. Ostala enostavna orodja

Za enostavno povezovanje vseh sklopov sistema smo razvili skupek lastnih orodij. To so enostavni programčki za obdelavo besedil ter zbirka skript za lažjo uporabo in avtomatizacijo uporabe orodij. Poleg orodij za delo z besedili je tu 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 (Alshawi et al., 2000). 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.

3.4. Učenje prevajalskega modela

Celoten učni proces podpira več modulov, ki so delo različnih razvijalcev. Na začetku potrebujemo dvojezični vzporedni korpus, pri nas je to bil korpus IJS-ELAN (Erjavec, 2002). Enostaven programček RemoveSGMLMarks poskrbi za pretvorbo TEI oblike v enostavno zaporedje povedi ločenih v dve datoteki (eno za izvorni, drugo za ciljni jezik). Povedi so zapisane vsaka v svoji vrstici, istoležne povedi so prevod iz enega v drugi jezika 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. Prevajalni modeli nastajajo s pomočjo orodij zbirke Egypt.

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 kot 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 postavljajo 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 predhodnega modula je vhod naslednjega. Vhodni korpus Whittle predela v obliko primerno za obdelavo s programom GIZA++. Ta modul sestavlja parametrične modele po (Brown et al., 1994) ter (Vogel et al., 1996). Izhod predhodnega modela je vhod za naslednji model, izhod zadnjega modela je naučen sistem.

3.5. Prevajalski strežnik

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.

3.6. Programje za ovrednotenje kvalitete

V okviru našega sistema smo implementirali tudi avtomatsko testiranja po metodi SA/TA opisani v razdelku 5.. Testiranje je bilo izvedeno s pomočjo testih primerov, ki jih je izbral modul Whittle. Ti primeri vsebujejo tudi referenčne prevode, 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 prevodov za vsak par.

4. Uporaba lem v dvojezičnem korpusu

Osnovni model je bil razširjen z uporabo lem. Z redukcijo besednih oblik na leme smo želeli omiliti problem redkih podatkov (scarce data problem); z večanjem korpusa se veča število podatkov, veča pa se tudi osnovni prostor pregledovanja. Tako se gostota podatkov ne spreminja in iskanje pravil je prav tako nenatančno kot na izbranem manjšem korpusu.

Dvojezični vzporedni korpus ELAN, ki je osnova našega sistema, je že lematiziran. Slovenščina je visoko pregiben jezik 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.

Naša hipoteza je predvidevala, da bo sprememba korpusa ugodno vplivala na kakovost prevodov.

Pri učenju sistema smo uporabili lematizirane slovenske povedi ter osnovne angleške povedi. Vhod v novi sistem mora biti tako tudi spremenjen, vhodne slovenske povedi so najprej lematizirane, šele nato podane v prevajanje. Tako naučen sistem naj bi bil bolj odporen na sklanjanja in pregibanja v slovenskem jeziku, ki so v angleščini manj uporabljana. Nov izgled korpusa za prevajanje je tako:

angleški vhod: *manufacture in which all the material of chapter 1 and 2 use must be wholly obtained*

slovenski vhod: *izdelava pri kateri morati biti ves uporabljen material iz poglavje 1 in 2 v celota pridobiti*

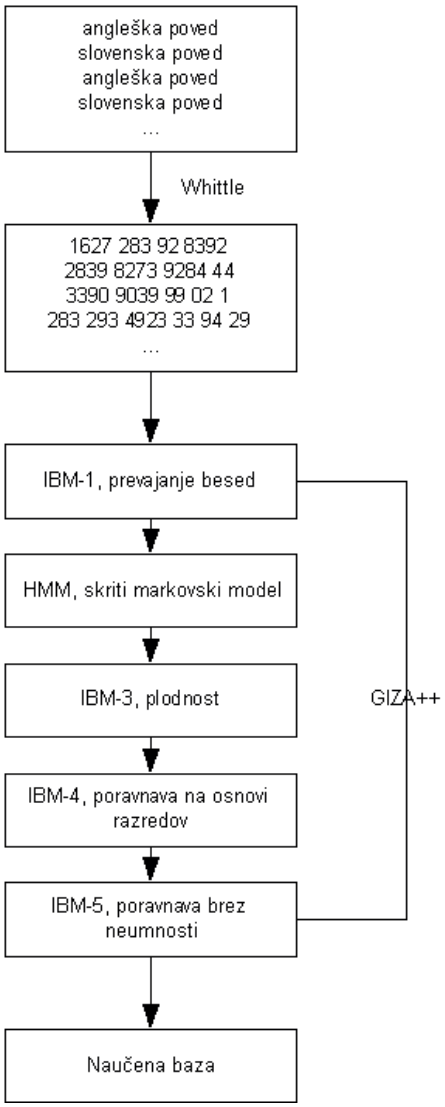
S konverzijo v leme smo zmanjšali število različnih besed v slovenskem delu korpusa s 39.221 na 27.211 oziroma na 69% v slovenskem delu korpusa.

Postavili smo nov sistem (ponovno učenje na novem, popravljenem korpusu) ter izvedli enaka testiranja kot pri osnovnem strežniku. Testiranja so bila izvedena na enakih testnih primerih, obširneje so predstavljena v nadaljevanju.

5. Testiranje

Pri testiranju osnovnega ter popravljenega sistema smo se odločili samo za testiranje kakovosti 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



Slika 3: prikaz prehoda učnih primerov pri učenju prevajalnih modelov

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 slabo raziskanostjo. 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.

SA/TA, enostavna natančnost, natančnost preslikav (Alshawi et al., 2000) in (Vogel et al., 1995): Za vsak prevod izračunamo urejevalno razdaljo (edit distance), tj. število vrinjenih, brisanih ali 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

$$SA = 1 - (I + D + S)/R \quad (6)$$

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

$$SA = 1 - (I' + D' + S + T)/R \quad (7)$$

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).

SSER, Subjective Sentence Error Rate (Vogel et al., 1996): Prevodi so rangirani v pet kakovostnih razredov:

- 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.

Metodi sta bili še dodatno testirani s pomočjo zbirke prevodov (ročni prevodi testne množice). Ta množica prevodov je bila postavljena kot ideal 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% predstavlja 100%)

Tabela 1: rezultati testne skupine z metodo 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

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

Rezultati, prikazani v tabeli 3, 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 in 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 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 ročna referenčnih prevodov.

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.

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 prevodov osnovnega sistema ter 100 prevodov sistema z uporabo lem.

5.1. 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, je pa 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.

5.2. Rezultati vrednotenja sistema z uporabo lem

Metoda uporabe lem za prirejeno opisovanje jezika se je izkazala kot primerna. Rezultati ocenjevanja prevodov so pri spremenjenem sistemu občutno boljši kot pri osnovnem. Zavedati se moramo, da je uporaba sistema z lemami veliko bolj zahtevna, saj sistem vhodne povedi najprej lematizira ter jih nato preda prevajalniku.

Sistem z uporabo lem ni brez pomanjkljivosti. Pri pretvarjanju osnovnih oblik v leme izgubimo veliko informacij, ki bi jih med prevajanjem potrebovali, kot so čas, spol...

Potrebujemo ločevanje pozitivnimi ter negativnimi lastnostmi nove metode. Popravljena metoda, ki bi zadržala izboljšane primere in ne bi uvajala novih napak v prevajalni sistem, bi temeljila na podmnožici lem.

Verjetno bi bili prevodi sistema, zgrajenega z osnovnim modelom ter nadgrajenega le z izbranimi vrstami lem, boljši.

Spodaj podamo par primerov bolj ali manj uspešnih prevodov:

- *razen izdelek iz navaden kovina*
  save the products from base metals
- *zmanjsanje emisija toplogrednih plin kioto*
  reduce emissions of greenhouse gases kyoto
- *pravilnik o gospodarski in sporten ribolov december*
  regulations on commercial and recreational fishing decembers
- *kava caj mat caj in zacimba*
  coffee or tea chicory tea and condiments
- *tapioka in njen nadomestek pripravljen iz skrob kot kosmiciti zrnce perle ali v podoben oblika*
  tapioka and her spongy prepared from the starch as kosmiciti zrnce perle or similar forms

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.

	SA/TA		SSER	
	povp.	odstotki	povp.	odstotki
Osnovni sistem	1,40	9,97	1,94	23,55
Lemma	1,99	14,19	2,42	35,50
testna skupina	2,47	36,84	4,48	87,00

- *trg delo*
  trg works

6. Zaključek in nadaljnje delo

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

Metoda razširitve osnovnih algoritmov z uporabo lem se je izkazala kot uspešna, saj se je natančnost prevodov občutno izboljšala. Zavedati pa se moramo, da sistem z uporabo lem ni brez pomanjkljivosti. Pri pretvarjanju osnovnih oblik v leme izgubimo veliko informacij, ki bi jih med prevajanjem potrebovali, kot so čas, spol... Ta problem bi lahko rešili z delno lematizacijo besedil, lematizirali bi le izbrane besedne vrste. Dodatna možnost je lematiziranje samo odprtih pregibnih besednih vrst, t.j. polnopomenski glagoli, samostalniki in pridevniki; IJS-ELAN oznake so Vm/N/A. Ostale besedne vrste, pomožni glagol ter zaimke (oznake iz IJS-ELAN Vc/P) pa izpustimo.

Poleg uporabe lem za dograditev sistema se poraja kot najočitnejša možnost uporaba slovensko – angleškega slovarja, s čimer bi zmanjšali število neznanih besed v prevodih.

Zahvala

Avtorja se zahvaljujeta vsem, ki so pomagali pri projektu, še posebej pa Janu Cuřínu s fakultete za matematiko in fiziko Karlove univerze v Pragi, ki je pomagal v začetnih fazah postavitve strežnika. Za koristne pripombe gre zahvala tudi anonimnima recenzentoma osnutka tega članka — za vse napake, ki so še ostale v članku, pa avtorja krivita drug drugega.

7. Literatura

Hijan Alshawi, Srinivas Bangalore, in Shona Douglas. 2000. Learning Dependency Translation Models as Collections of Finite State Head Transducers. *Computational Linguistics*, 26(1):45–60.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, in Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):163–311.

Peter Brown, Peter Cocke, Stephen Della Pietra, Vincent Della Pietra, Fredrik Jelinek, John Lafferty, Robert Mercer, in Paul S. Roossin. 1994. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

Tomaž Erjavec. 2002. The IJS-ELAN Slovene-English Parallel Corpus. *International Journal of Corpus Linguistics*, 7(1). http://nl.ijs.si/elan/.

Bernard Merialdo. 1992. Tagging Text with a Probabilistic Model. V: *1992 AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, str. 675–685.

Edward Sapir. 1921. *Language: an Introduction to the Study of Speech*. Harcourt, Brace and company, New York.

C. M. Sperberg-McQueen in Lou Burnard. 2000. Guidelines for Electronic Text Encoding and Interchange, The XML Version of the TEI Guidelines. *The TEI Consortium*, 26(1):349–357.

Stephan Vogel, Fraz Josef Och, Christof Tillmann, Sonja Niessen, Hassan Sawaf, in Hermann Ney. 1995. Statistical Methods for Machine Translation. *Lehrstuhl für informatik VI, Computer Science Department RWTH Aachen University of Technoogy*.

Stephan Vogel, Hermann Ney, in Christof Tillmann. 1996. HMM-based Word Alignment in Statistical Translation. V: *In COLING '96: The 16th Int. Conf. On Computational Linguistics*, str. 836–841.

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