Statistično strojno prevajanje na osnovi vzporednih korpusov

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

Statistično strojno prevajanje na osnovi vzporednih korpusov

The paper presents an experiment in automatic translation from Slovenian to English language based on SMT, Statistical Machine Translation.
EGYPT is the result of a summer workshop at John Hopkins University, and is currently most widely used toolbox for processing bilingual parallel corpora for translation system production.
The IJS-ELAN corpus contains 1 million words of annotated parallel and sentence aligned Slovene-English texts, with both languages word-tagged with context disambiguated morphosyntactic descriptions and lemmas. The corpus is encoded in XML, according to the TEI Guidelines P4.
A Slovene to English translation system was produced using the EGYPT toolbox and the IJS-ELAN corpus. We discuss the motives for source/target language selection, i.e. why we chose to train the system for Slovenian to English translation rather than vice-versa.

1. Uvod

Sistem omogoča avtomatsko prevajanje iz slovenskega jezika v angleški. Postavljen je na osnovi statističnih parametričnih prevajalnih ter jezikovnih modelov s pomočjo zbirke EGYPT ter ostalih pomožnih orodij.

Pri postavitvi statističnih modelov smo črpali znanje o obeh jezikih ter njunih odvisnostih v dvojezičnem vzporednem korpusu IJS-ELAN.
EGYPT je zbirka orodij za obdelavo dvojezičnih vzporednih korpusov za strojno prevajanje. Je končni izdelek poletne delavnice na univerzi JHU - John Hopkins University in je trenutno najširše uporabljana zbirka v sistemih strojnega prevajanja.

IJS-ELAN korpus je največji slovenski dvojezični korpus, dvojezični korpus, ki ima polovico besedila zapisanega v slovenskem jeziku, drugo polovico pa v izbranem tujem jeziku. 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.

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.

Izvedeno je bilo tudi osnovno vrednotenje obeh sistemov. Vrednotenje osnovnega sistema nam je pokazalo primernost uporabe izbranih tehnologij za prevajanje slovenskega jezika. S tem vrednotenjem smo postavili tudi referenčno osnovo za vrednotenja novih sustemov . Nove prevedbe so bile ovrednotene ter rezultati primerjani z osnovnim sistemom, pri vrednotenju so bili uporabljeni isti testni primeri.

1.1. 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 kupec 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) (Sapir, 1921).

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.

1.2. The *Candide* system

Razdelek prikazuje sistem *Candide*, ki so ga v začetku devedesetih let razvili pri *IBM*.

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 . Poimenovali so ga »The Candide system for machine translation«.

Oglejmo si primer prevoda besedila v nekem jeziku, izberemo slovenski jezik, v besedilo v angleškem jeziku. Za poved $\boldsymbol{f}$ v slovenskem jeziku si zamislimo, da je bila zgrajena iz pripadajoče povedi $\boldsymbol{e}$ 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 *f*. Pri dani slovenski povedi *f* postane naš problem, problem avtomatskega prevajanja, iskanje angleške povedi, ki maksimira $P(e|f)$. Torej iščemo:

$$\hat{e} = \arg\max_{e} P(e \mid f) \quad \textbf{(0)}$$

Z uporabo Bayesove formule dobimo:

$$\hat{e} = \arg\max_{e} P(e \mid f) = \arg\max_{e} P(f \mid e)P(e) \quad \textbf{(0)}$$

S $P(f|e)$ zapišemo verjetnost da dobimo *f* 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 *f*, jezikovni model zagotavlja, da je *e* res poved. Candid izbere takšno poved *e*, ki maksimizira produkt prej opisanih funkcij.

2. Egypt

Na poletni delavnici, leta 1999, na JHU (John Hopkins University) so po vzoru izdelali zbirko orodij, ki omogočajo postavitev popolnega SMT sistema osnovanega na dvojezičnih vzporednih korpusih. Zbirko so poimenovali Egypt. Osnovni cilji delavnice so bili: postavitev zbirke orodij za statistično strojno prevajanje, zbirka naj bo splošno dosegljiva raziskovalni srenji. Postavitev češko-angleškega sistema za prevajanje besedil na osnovi izdelanih orodij. Osnovno testiranje sistema na osnovi objektivnih mer (statistično modeliranje težavnosti).

3. Slovenščina → Angleščina

Razdelek predstavlja izbiro izvornega in ciljnega jezika.

Oglejmo si najprej izvorni jezik, Slovenščino. 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 zaimki (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% manj besed v korpusu kot angleščina.

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.

4. 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 nadaljevanju.

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.

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.

```
all the animal of chapter 1 use must be …
ves zival iz poglavje 1 morati biti …
```

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 smo povečali iz 27421 na 58803. Več kot dvakratna količina informacij naj bi povečala natančnost prevajanja, saj je bil osnovni korpus relativno majhen. Število 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.

5. 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, dobrih rezultatov 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 je 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 : Prevodi so rangirani v pet kakovostnih razredov od »popoln prevod«, ki je ocenjen s 100% do »popolna bedarija«, vredna 0%.

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 : Za vsak prevod je izračunana 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 .

- *SA, simple accuracy, enostavna natančnost*

$$SA = I - (I + D + S)/R \qquad \textbf{(0)}$$

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*

$$TA = I - (I' + D' + S + T)/R \qquad \textbf{(0)}$$

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 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.

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

	SA/TA		
	Povprečje	V %	st. dev.
testna skupina	2,47	36,84	1,38

Testna množica je bila ocenjena z metodo SSER s koeficientom 4,48 oziroma s 87 odstotki. (87% predstavlja 100%).

Tabela 2: rezultati testne skupine z metodo SSER

	SSER		
	Povprečje	V %	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).

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).

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.

6. 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.

6.1. 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.

6.2. Rezultati vrednotenja osnovnega sistema

Prevodi osnovnega sistema so kar vzpodbudni. Veliko prevodov je popolnoma uporabnih, te so eksperti pri metodi SSER ocenili s 75% ali celo 100%, seveda pa je kar veliko tudi popolnoma zgrešenih prevodov (ocenjeni z 0% 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.

6.3. 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 mrd 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.

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	V %	stdev	Povp.	V %	stdev
osn. Sistem	1,40	9,97	0,71	1,94	23,55	1,25
Lema	1,40	9,96	0,70	1,91	22,77	1,31
Test	2,47	36,84	1,38	4,48	87,00	0,65

Tabela 4: del rezultatov ocenjevanja testnih primerov z metodo SSER

	E1 – lema	E1 - norm	E2 - lema	E2 – norm	E3 – lema	…
povpr.	1,86	1,79	2,02	2,01	1,85	…
stdev	1,30	1,29	1,35	1,25	1,30	…
V %	21,51	19,77	25,58	25,29	21,22	…
0%	56	61	49	44	59	…
25%	24	21	25	33	21	…
50%	5	5	8	8	3	…
75%	5	2	5	5	8	…
100%	10	10	11	9	8	…

8. Literatura

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 delno predstavljena v 7. poglavju.

7. 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. Že 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.

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

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

[3] 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, Germany.

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

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