Univerza v Mariboru

Fakulteta za elektrotehniko,
računalništvo in informatiko

Jani Dugonik

UGLAŠEVANJE PARAMETROV PRI STATISTIČNEM STROJNEM PREVAJANJU

Magistrsko delo

Maribor, julij 2013

UGLAŠEVANJE PARAMETROV PRI STATISTIČNEM STROJNEM PREVAJANJU

Magistrsko delo

Študent(ka):	Jani Dugonik
Študijski program:	Računalništvo in informacijske tehnologije (MAG)
Smer:	Inteligentne informacijske tehnologije
Mentor(ica):	red. prof. dr. Janez Brest
Somentor(ica):	doc. dr. Borko Bošković
Lektor(ica):	Teja Pušaver

ZAHVALA

Zahvaljujem se mentorju red. prof. dr. Janezu Brestu, somentorju doc. dr. Borku Boškoviću in izr. prof. dr. Mirjam Sepesy Maučec za pomoč in vodenje pri opravljanju magistrskega dela. Zahvaljujem se tudi Teji Pušaver za lektoriranje magistrskega dela.

Posebna zahvala velja staršem, ki so mi omogočili študij.

Univerza v Mariboru
Fakulteta za elektrotehniko
računalništvo in informatiko

Številka: E5000445
Datum in kraj: 26. 03. 2013, Maribor

Na osnovi 330. člena Statuta Univerze v Mariboru (Ur. l. RS, št. 01/2010)
izdajam

SKLEP O MAGISTRSKEM DELU

1. **Jani Dugonik**, študent študijskega programa 2. stopnje RAČUNALNIŠTVO IN INFORMACIJSKE TEHNOLOGIJE, izpolnjuje pogoje, zato se mu dovoljuje izdelati magistrsko delo.

2. Tema magistrskega dela je s področja Inštituta za računalništvo, pri predmetu Povezljivi sistemi in inteligentne storitve.

MENTOR: red. prof. dr. Janez Brest
SOMENTOR: doc. dr. Borko Bošković

3. Naslov magistrskega dela:
UGLAŠEVANJE PARAMETROV PRI STATISTIČNEM STROJNEM PREVAJANJU

4. Naslov magistrskega dela v angleškem jeziku:
TUNING PARAMETERS IN STATISTICAL MACHINE TRANSLATION

5. Magistrsko delo je potrebno izdelati skladno z »Navodili za izdelavo magistrskega dela« in ga do 26. 03. 2014 v 2 vezanih in 1 v spiralo vezanem izvodu oddati v pristojni referat ter elektronski izvod v Digitalno knjižnico Univerze v Mariboru.

Obvestiti:
1. kandidata
2. mentorja
3. somentorja
4. arhiv

Dekan:
red. prof. dr. Borut Žalik

Naslov: Uglaševanje parametrov pri statističnem strojnem prevajanju
Ključne besede: strojno prevajanje, statistično strojno prevajanje, jezikovni model, evolucijski algoritem, uteži, uglaševanje, optimizacija
UDK: 004.5:004.93(043.2)

Povzetek

V magistrskem delu se bomo osredotočili na uglaševanje sistema za strojno prevajanje. V okviru tega bomo vzpostavili sistem za strojno prevajanje, ki temelji na statističnih modelih. Omejili se bomo na prevajanje iz slovenščine v angleščino. Rezultate bomo ocenili z metriko BLEU. Pričakujemo, da bo naš pristop primerljiv z ostalimi metodami.

Title: Tuning parameters in statistical machine translation
Keywords: machine translation, statistical machine translation, language model, evolutionary algorithm, weights, tuning, optimization
UDK: 004.5:004.93(043.2)

Abstract

In this master's thesis, we will focus on tuning the machine translation system. In this context, we will establish the machine translation system based on statistical models. We will confine ourselves to translating from Slovene to English. The results will be evaluated by the BLEU metric. We expect that approach is comparable to other tuning methods.

KAZALO VSEBINE

1 UVOD 1

1.1 STATISTIČNO STROJNO PREVAJANJE 2

1.1.1 Oris delovanja 3

1.1.2 Vrste statističnega strojnega prevajanja 3

1.1.3 Strojni prevajalniki 5

1.2 PREGLED VSEBINE 7

2 STATISTIČNO STROJNO PREVAJANJE NA OSNOVI FRAZ 8

2.1 MODEL, KI TEMELJI NA FRAZAH 8

2.2 METODE ZA EVALVACIJO STROJNIH PREVODOV 11

2.2.1 Ročna evalvacija 12

2.2.2 Povratni prevod 13

2.2.3 Samodejna evalvacija 13

3 UGLAŠEVANJE PARAMETROV IN DIFERENCIALNA EVOLUCIJA 17

3.1 SORODNA DELA 17

3.1.1 Minimum Error Rate Training (MERT) 17

3.1.2 Margin Infused Relaxed Algorithm (MIRA) 18

3.1.3 Pairwise Ranking Optimization (PRO) 18

3.2 DIFERENCIALNA EVOLUCIJA 18

3.2.1 Inicializacija 20

3.2.2 Mutacija 21

3.2.3 Križanje 22

3.2.4 Popravljanje 23

3.2.5 Selekcija 24

4 EKSPERIMENT 25

4.1 NAČRTOVANJE EKSPERIMENTA 25

4.1.1 Zbirka JRC-ACQUIS 25

4.1.2 Orodje Moses 29

4.1.3 Sistem za strojno prevajanje 30

4.2 IZBOLJŠAVA SISTEMA ZA STROJNO PREVAJANJE 30

4.3 NAŠ PRISTOP ZA UGLAŠEVANJE PARAMETROV 31

4.3.1 Inicializacija *32*
4.3.2 Ocenjevanje posameznikov *33*
4.3.3 Ustvarjanje novih posameznikov *34*

5 REZULTATI 36

5.1 PREDOBDELAVA KORPUSA 36
5.2 OPIS PARAMETROV 36
5.3 REZULTATI UGLAŠEVANJA 36
5.4 REZULTATI TESTIRANJA 40
5.5 STATISTIČNA PRIMERJAVA Z ORODJEM MULTEVAL 41

6 SKLEP 43

VIRI 44

KAZALO TABEL

Tabela 1: Tabela prevodov fraz 10
Tabela 2: Zbirka dokumentov ACQUIS (verzija 2.2) v 21 jezikih 27
Tabela 3: Parametri pred uglaševanjem z metodo MERT 38
Tabela 4: Parametri po končanem uglaševanju z metodo MERT 38
Tabela 5: Parametri pred uglaševanjem z metodo MIRA 38
Tabela 6: Parametri po končanem uglaševanju z metodo MIRA 38
Tabela 7: Parametri pred uglaševanjem z metodo PRO 38
Tabela 8: Parametri po končanem uglaševanju z metodo PRO 38
Tabela 9: Parametri pred uglaševanjem z našim pristopom 39
Tabela 10: Parametri po končanem uglaševanju z našim pristopom 39
Tabela 11: Rezultati uglaševanja (ocena BLEU in čas v urah) 39
Tabela 12: Rezultati testiranja 40
Tabela 13: Prikaz statistične primerjave z orodjem MultEval 42

KAZALO SLIK

Slika 1: Primer prevajanja po frazah 10
Slika 2: Oblika dokumenta DTD (Document Type Definition) 28
Slika 3: Izboljševanje najboljšega posameznika 39

KAZALO ALGORITMOV

Algoritem 1: Osnovni algoritem diferencialne evolucije 20
Algoritem 2: Inicializacija 21
Algoritem 3: Mutacija 22
Algoritem 4: Križanje 23
Algoritem 5: Popravljanje 24
Algoritem 6: Selekcija 24
Algoritem 7: Glavni program našega pristopa 32
Algoritem 8: Inicializacija 33
Algoritem 9: Ocenitvena funkcija 34
Algoritem 10: Vsebina skripte eval.sh 34
Algoritem 11: Spremenjena strategija DE/rand/1/bin 35

SEZNAM KRATIC

MT – machine translation

SMT – statistical machine translation

BLEU – bilingual evaluation understudy

METEOR – metric for evaluation of translation with explicit ordering

TER – translation error rate

TERp – translation error rate plus

ALPAC – automatic language processing advisory committee

ARPA – advanced research projects agency

GTM – general text matcher

ROUGE – recall-oriented understudy for gisting evaluation

WER – word error rate

NIST – national institute of standards and technology

MERT – minimum error rate training

MIRA – margin infused relaxed training

PRO – pairwise-ranked optimization

DE – differential evolution

EU – european union

JRC – joint research centre

AC – acquis communutuaire

TEX – text encoding initiative

1 UVOD

Naravni jezik je eden izmed najkompleksnejših sistemov, kar se jih je razvilo skozi človekovo evolucijo. Prevajanje pa je kot prenašanje sporočil med temi sistemi eden izmed najbolj zapletenih miselnih procesov, pri katerem ostajajo mnoga vprašanja s psihološkega in jezikoslovnega vidika nepojasnjena.

Razvijanje računalniških tehnologij, ki bi znale ta proces simulirati ali vsaj optimizirati nekatere njegove dele, že dolgo predstavlja izziv tako jezikoslovcem kot tudi računalničarjem.

Strokovnjaki na področju strojnega prevajanja že dolgo poskušajo opisati jezik. Naravni jezik pa je živ in se nenehno spreminja, pravila so kompleksna in ne upoštevajo kreativnosti, zato jezika z vsemi izjemami ne morejo ukleščiti v svoje stroge okvire. To je že na začetku stoletja ugotovil Edward Sapir in napisal, da "vse slovnice puščajo" [43].

Strojno prevajanje (angl. *machine translation*) [23] je postopek, pri katerem računalniški program analizira besedilo v izvornem jeziku in brez posredovanja človeka proizvede besedilo v ciljnem jeziku. Sistemi za strojno prevajanje vključujejo eno- ali večjezične leksikone, programe za morfološko analizo in sintezo, programe za sintaktično analizo in sintezo, programe za razreševanje večpomenskosti, programe za prepoznavanje večbesednih semantičnih enot in druge kompleksne mehanizme, ki služijo avtomatizaciji prevajalskega procesa. Čeprav je avtomatizacija glavna značilnost strojnega prevajanja, pa tovrstni prevajalni sistemi pogosto zahtevajo posredovanje človeka, in sicer v obliki predhodnega urejanja in kasnejših popravkov. Pri semantičnih ali leksikalnih težavah nekateri sistemi vključujejo tudi interakcijo z uporabnikom že med samim prevajalskim procesom.

Prevajanje je zahtevno in ustvarjalno dejanje. Strojno prevajanje lahko v nekaterih primerih prevajalcu delo olajša ali pa ga celo popolnoma nadomesti:

- če potrebujemo le grob prevod, ki ga bo kasneje pregledal in popravil prevajalec,
- kot osnutek, ki služi kot pomoč pri prevajanju ali
- če gre za določene besedilne vrste, pri katerih je izrazje močno omejeno (vremenske napovedi, navodila za uporabo, računalniški programi, inventurni zapisniki, zdravniška poročila ipd.).

Razlikujemo več metod strojnega prevajanja:

- na osnovi pravil,
- na osnovi podatkov,
- statistična metoda,
- na osnovi primerov in
- hibridni sistemi.

1.1 Statistično strojno prevajanje

Statistično strojno prevajanje je vrsta strojnega prevajanja, ki temelji na večji količini vzporednih besedil, iz katerih se s statističnimi algoritmi izračunavajo verjetnosti prevodne ekvivalence za posamezne jezikovne enote. Besedilo je prevedeno glede na verjetnostno porazdelitev. Izbran je tisti prevod, ki ima najvišjo verjetnost, ta pa se običajno računa po posameznih povedih [58]. Statistične metode so se prvotno ukvarjale s prevajanjem posameznih besed, v zadnjih letih pa so napredovale na raven besednih zvez. Največje prednosti so neodvisnost od jezikovnega para, manjši stroški kot pri metodah, ki temeljijo na pravilih, in bolj naravni prevodi ter neobčutljivost na napake.

Začetki statističnega strojnega prevajanja segajo v leto 1949, ko je ameriški znanstvenik Warren Weaver pri prevajanju uporabil teorijo informacij [51]. Leta 1991 pa so raziskovalci iz podjetja IBM za to vejo strojnega prevajanja vzbudili veliko zanimanje.

1.1.1 Oris delovanja

Sistemi statističnega strojnega prevajanja so osnovani na parametričnih statističnih modelih, ki so naučeni na poravnanih dvojezičnih korpusih (učnih primerih). Namesto da bi prevajalnik − kot pri strojnem prevajanju na osnovi pravil − razčlenjeval stavke po slovničnih pravilih, ta išče splošne vzorce, ki se pojavljajo v jezikovni rabi. Besedilo je prevedeno glede na verjetnostno porazdelitev – izbran je tisti prevod, ki ima najvišjo verjetnost, katera se običajno računa po posameznih povedih. Statistične metode so se prvotno ukvarjale s prevajanjem posameznih besed, v zadnjih letih pa so napredovale na raven besednih zvez.

1.1.2 Vrste statističnega strojnega prevajanja

Prevajanje po besedah

Pri tem tipu prevajanja je osnovna prevodna enota beseda nekega naravnega jezika. Število besed v izhodiščni in ciljni povedi je običajno različno. Razmerje med dolžinami prevedenih povedi se imenuje plodnost (angl. *fertility*) [2]. Ta nam pove, koliko besed v ciljnem jeziku proizvede vsaka beseda v izhodiščnem jeziku. Informatika predpostavlja, da med seboj ustrezajoči si leksemi pokrivajo isti pojem, resnica pa je pogosto drugačna. Tako se lahko slovensko besedo *kót* v španščino prevede z besedo *rincón* ali *esquina* − odvisno, ali gre za notranji ali zunanji kot.

Primer sistema za strojno prevajanje po besedah je program GIZA++ [37]. Ta preprosti način prevajanja pa ni ustrezen za prevajanje med jeziki z različno plodnostjo. Sicer je relativno preprosto ustvariti sistem strojnega prevajanja na osnovi besed, ki se lahko kosa z visoko plodnostjo, saj ene besede ni težko prevesti z več besedami; težava se pojavi v nasprotni smeri – prepoznavanju večbesednih enot in prevajanju le-teh z eno besedo.

Spodnji primer [48] prikazuje pravilen prevod prevajalnika Google iz slovenščine v angleščino ter napačnega iz angleščine v slovenščino, pri katerem prevajalnik kot povedek

ni prepoznal fraznega oziroma sestavljenega glagola *call off = cancel = preklicati*, temveč le njegov del *call = poklicati*.

On je preklical poroko. – *He canceled the wedding*.
He called off the wedding. – Poklical je off poroko.

Prevajanje po besednih zvezah

Bolj kot po besedah, se v zadnjem času prevaja po (različno dolgih) besednih nizih, ki jih imenujemo fraze. Cilj tega je, da bi se zmanjšale omejitve prevajanja po besedah. Z "nizi" niso mišljeni stavki kot slovnične strukture, temveč nizi besed, ki jih v korpusu prepoznajo sistemi, ustvarjeni po statistični metodi. Dokazano je bilo, da omejevanje nizov na slovnične stavke (slovnično motivirane skupine besed) zmanjšuje kakovost prevoda.

Prevajanje po slovnici

Prevajanje po slovnici v nasprotju s prevajanjem po besedah in besednih nizih temelji na prevajanju slovničnih enot, tj. slovnično razčlenjenih povedih. Ta pristop je na področju strojnega prevajanja prisoten že dlje, komaj v 90-ih letih 20. stoletja pa so začeli nastajati močni naključni slovnični razčlenjevalniki (angl. *stochastic parsers*).

Prednosti statističnih metod

Največje prednosti statistične metode v primerjavi z bolj tradicionalnimi pristopi so:

- lažje dostopni viri: veliko materiala v naravnem jeziku je dostopnega v digitalni obliki,
- neodvisnost od jezikovnega para (četudi je ta metoda pri določenih jezikovnih parih ustreznejša (mednje ne spada slovenščina [55])),
- manjši stroški kot pri metodah, ki temeljijo na pravilih in
    - sistemi za strojno prevajanje na osnovi pravil zahtevajo ročno oblikovanje − razvijanje jezikovnih pravil, kar je neekonomično,

    - pravil prav tako pogosto ni mogoče posplošiti oziroma prenesti na druge jezike,
- naravnejši prevodi: sistemi, ki temeljijo na pravilih, pogosto prevajajo dobesedno, statistično strojno prevajanje pa se tej težavi izogiba. Težava pri zanašanju na verjetnostno porazdelitev (v nasprotju s slovarji in slovničnimi pravili) pa je, da statistično prevedena besedila pogosto vključujejo že na prvi pogled nesmiselne in očitne napake.

Problemi

Najpogostejše težave, ki se pojavijo pri statističnem strojnem prevajanju, so:

- poravnava povedi: v vzporednih korpusih je marsikatera poved izhodiščnega besedila prevedena v več povedi ciljnega besedila in obratno,
- prevajanje sestavljenk,
- prevajanje frazeologije,
- razlike v oblikoslovju različnih jezikov: oblikoslovne težave so precej pogoste pri slovenščini, saj je ta morfološko zelo bogata,
- razlike v besednem redu: različni jeziki imajo različen besedni red. Do določene mere se ta lahko določi z običajnim vzorcem osebek-povedek-predmet, tako lahko govorimo o jezikih SVO (subject-verb-object), VSO (verb-subject-object) ipd. Dodatne razlike v besednem redu se pojavijo pri prilastkih ter med trdilnimi in vprašalnimi povedmi,
- besede zunaj besedišča (angl. *out of vocabulary*): sistemi za statistično strojno prevajanje imajo v svojih bazah podatkov različne besedne oblike shranjene kot posebne simbole brez medsebojne povezave. Besednih oblik in stavkov, ki niso v bazi podatkov, ni mogoče prevesti. Do tega pride zaradi pomanjkanja besedilnih virov, razlik v oblikoslovju različnih jezikov ipd.

1.1.3 Strojni prevajalniki

Prvi spletni prevajalnik, ki je bil dostopen le naročnikom poštnih storitev mreže Minitel iz Francije, je nastal konec 80-ih let 20. stoletja. Leta 1997 se je na medmrežju predstavil prvi

brezplačni prevajalnik Babel Fish, ki je nastal s sodelovanjem podjetja Systran Software Inc. in brskalnika AltaVista. Leta 2006 ga je pod svoje okrilje vzel Yahoo, zato se zdaj imenuje Yahoo Babel Fish. Sprva je prevajalnik vključeval deset jezikovnih kombinacij, do leta 2006 pa je svoje storitve razširil na kar 38 jezikovnih kombinacij. Danes je na spletu na voljo kar nekaj takih prevajalnikov, pojavljajo pa se tudi lažni strojni prevajalniki (angl. *spoof machine translation*), namenjeni predvsem zabavi uporabnikov, ki po mnenju nekaterih strokovnjakov slabo vplivajo na ugled pravih spletnih prevajalnikov [40].

Na spletu sta v kombinaciji s slovenščino na voljo naslednja brezplačna strojna prevajalnika, ki delujeta po statistični metodi:

- prevajalnik Google [58] in
- prevajalnik Bing [55].

Dobri rezultati se pri strojnem prevajanju naravnih jezikov kažejo predvsem pri prevajanju sorodnih jezikov.

Prevajalnik Google

Prevajalnik Google je tipičen predstavnik sistemov statističnega strojnega prevajanja. Sistem ne uporablja dodatnega jezikovnega znanja, zanaša se samo na korelacijo med znanimi pari že prevedenih vzporednih besedil. Statistične metode zahtevajo ogromne količine besedil in veliko računalniške moči za obdelavo teh besedil. Google ima oboje, besedila nabrana za izdelavo iskalnika in velike gruče računalnikov, ki omogočajo hitro izdelavo sistemov za strojno prevajanje z zavidljivo kakovostjo. Opisane lastnosti so omogočile Googlu izdelavo prevajalnih sistemov za kar 58 svetovnih jezikov, torej za kar 3364 jezikovnih parov. O natančnosti delovanja prevajalnika Google ni veliko znanega, znane so le osnovne metode.

Prevajalnik Bing

Prevajalnik Bing je hibridni sistem za strojno prevajanje naravnih jezikov. Sistem temelji na statističnem strojnem prevajalniku, ki uporablja tudi pravila, ki so odvisna od jezika in

določene mere analize izvornega besedila. Microsoft imenuje ta sistem kot "jezikovno obveščeno statistično strojno prevajanje" (angl. *linguistically informed statistical machine translation*). Sistem je v osnovi statistični sistem za strojno prevajanje, ki temelji na frazah in vključuje jezikovno odvisno analizo besedila, drevesa odvisnosti (angl. *dependency tree*) in drevesa izpeljave (angl. *parse tree*) ter pravila za poravnavo besed (angl. *word alignment rules*) za generalizacijo naučenih fraz.

1.2 Pregled vsebine

V magistrskem delu se bomo osredotočili na uglaševanje sistema za strojno prevajanje. V okviru tega bomo vzpostavili sistem za strojno prevajanje, ki temelji na statističnih modelih. Omejili se bomo na prevajanje iz slovenščine v angleščino. Rezultate bomo ocenili z metriko BLEU [39]. Pričakujemo, da bo naš pristop primerljiv z ostalimi metodami za uglaševanje statističnih modelov prevajanja. V drugem poglavju bomo predstavili statistični model za strojno prevajanje, ki temelji na frazah, in metode za evalvacijo strojnih prevodov. V tretjem poglavju bomo predstavili sorodne pristope za uglaševanje parametrov pri statističnem strojnem prevajanju in opisali diferencialno evolucijo, ki je osnova našega pristopa. V četrtem poglavju bomo opisali naš eksperiment. V petem poglavju so predstavljeni rezultati tako sorodnih pristopov kot tudi našega in njihova primerjava. Delo je zaključeno s sklepnim poglavjem, v katerem je povzeto jedro magistrskega dela in so predstavljene smernice za raziskovanje nadaljnjega dela.

2 STATISTIČNO STROJNO PREVAJANJE NA OSNOVI FRAZ

Sistem za strojno prevajanje je sistem, ki uporablja modele prevajanja. Obstaja več modelov prevajanja:

- modeli, ki temeljijo na besedah (angl. *word-based models*),
- modeli, ki temeljijo na drevesih (angl. *tree-based models*),
- modeli, ki temeljijo na frazah (angl. *phrase-based models*) in
- faktorizirani modeli (angl. *factored models*).

Trenutno najuspešnejši sistemi statističnega strojnega prevajanja uporabljajo modele, ki temeljijo na frazah (angl. *phrase-based models*), zato smo tudi mi uporabili ta model v našem sistemu za strojno prevajanje.

2.1 Model, ki temelji na frazah

V statističnem strojnem prevajanju se uporabljajo verjetnostni modeli za iskanje najboljšega možnega prevoda *e** med vsemi možnimi prevodi *e* za podan stavek *f* [36]. Iskanje najboljšega prevoda se imenuje dekodiranje. Verjetnostni modeli so ocenjeni iz dvojezičnih in enojezičnih učnih podatkov in lahko vključujejo modele prevajanja, jezikovne modele, modele prerazporeditve itd. V praksi se uporablja diskriminativni linearni model z logaritemskimi verjetnostmi. Če so komponente modela $ħ_1, \dots, ħ_r$, katere so odvisne od *e* in *f*, potem dobimo najboljši prevod z

$$e^*(\lambda) = arg\max_e P(e,f) = arg\max_e \sum_{i=1}^{r} \lambda_i ħ_i(e,f)$$

in je odvisen od uteži komponent $\lambda_1, \dots, \lambda_r$. Komponenta $ħ_i$ je lahko neko število ali funkcija, npr. štetje besed, ki bo kaznovalo dolge ali kratke stavke. Problem se pojavi, kako najti množico uteži, ki bo ponujala najboljšo kakovost prevajanja.

V našem eksperimentu, ki smo ga podrobneje opisali v poglavju 4, je verjetnost $P(e, f)$, ki je dodeljena prevodu, produkt verjetnosti štirih modelov. Vsak od teh modelov prispeva informacijo o eni lastnosti dobrega prevoda:

- tabela prevodov fraz (angl. *phrase translation table*):
    - zagotavlja, da so angleške in slovenske fraze dobri prevodi med seboj,
- jezikovni model (angl. *language model*):
    - zagotavlja, da je izhod tekoča angleščina oz. drugi jezik,
- model popačenja (angl. *distortion model*):
    - omogoča prerazporeditve za vhodni stavek, ampak s stroškom: več je preurejanja, dražji je prevod,
- kazen za besede (angl. *word penalty*):
    - zagotavlja, da prevodi niso predolgi ali prekratki.

Matematični zapis:

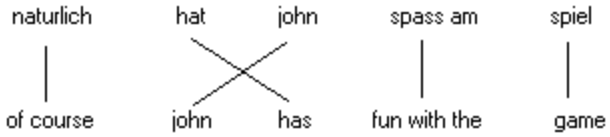
$$P(en|sl) = tm(sl|en)^{utež_{tm}} * \ lm(en)^{utež_{lm}} * \ dm(en, sl)^{utež_{dm}} * \ w(en)^{utež_{w}},$$

kjer je *tm* model za prevajanje, *lm* jezikovni model, *dm* model popačenja in *w* kazen za besede. Vsaki izmed teh komponent se lahko dodeli utež, katera določa njeno pomembnost.

Z nastavljanjem teh uteži na boljše vrednosti izboljšamo kakovost prevodov. Kakovost je odvisna od korpusa in jezikovnega para. Najenostavnejša metoda je, da poskusimo z velikim številom možnih nastavitev in izberemo najboljšo. Dobre vrednosti za tabelo prevodov fraz, jezikovni model in model popačenja so na intervalu 0.1 in 1. Pri kazni za besede so dobre vrednosti na intervalu -3 in 3, kjer negativne vrednosti dajejo prednost daljšemu prevodu, pozitivne vrednosti pa krajšemu prevodu [52]. Prevajanje po tem modelu bomo prikazali na naslednjem primeru.

Vhodni nemški stavek je segmentiran v kakršnekoli večbesedne enote, t. i. fraze. Nato je vsaka fraza prevedena v angleško frazo, te fraze pa lahko kasneje med seboj drugače

razporedimo (angl. *reordering*). Na sliki 1 imamo primer, kjer imamo šest nemških besed in osem angleških, ki so preslikane v pet parov fraz.



Slika 1: Primer prevajanja po frazah

Angleške fraze je treba razporediti tako, da glagol sledi predmetu. Da lahko prevajamo fraze, moramo imeti tabelo prevodov fraz (angl. *phrase translation table*), kot je prikazano v tabeli 1. Za vsako frazo obstaja več prevodov in vsak prevod ima svojo verjetnost. Izberemo tisti prevod fraze, ki ima večjo verjetnost.

Tabela 1: Tabela prevodov fraz

Prevod	**Verjetnost $P(en|de)$**
of course	0.5
Naturally	0.3
of course ,	0.15
, of course ,	0.05
…	…

Tabela prevodov fraz se lahko ustvari s pomočjo poravnave besed. Vsi pari fraz, ki so skladni s poravnavo besed, so dodani v tabelo. Za drugačno razporeditev fraz uporabimo model leksikalne prerazporeditve (angl. *lexicalized reordering model*). Ta model napoveduje usmerjenost (angl. *orientation*) fraze:

- monotona (angl. *monotone*),
- nezvezna (angl. *discontinuous*) in
- zamenjava (angl. *swap*).

Originalni model za preverjanje se da razširiti z dodatnimi komponentami modela:

- verjetnost dvosmernega prevoda (angl. *bidirectional translation probability*),

- oteževanje besed (angl. *lexical weighting*),
- kazen za besede (angl. *word penalty*) in
- kazen za fraze (angl. *phrase penalty*).

2.2 Metode za evalvacijo strojnih prevodov

Za spremljanje napredka strojnega prevajanja je potrebno ocenjevanje kakovosti prevodov. Ocenjevanje prevodov je zelo subjektivno in kompleksno, zato univerzalna metoda ocenjevanja še ni določena. Znana evalvacija strojnih prevajalnikov, ki je izšla leta 1966, je poročilo ALPAC (Automatic Language Processing Advisory Committee [1]). Kriterije ocenjevanja iz tega poročila so pri evalvaciji prevodov uporabljali številni raziskovalci.

Hutchins in Somers [23] pri analizi besedila navajata tri kriterije:

- informativnost: ali prevod posreduje enake informacije kot izvirnik,
- ustreznost: ali so besede v prevodu primerne glede na vsebino in namen ter
- razumljivost: ali je prevod jasen.

Pogosto so nerazumljivi prevodi zvesti izvirniku, popolnoma razumljivi prevodi pa posredujejo nepopolne ali celo napačne informacije. Prav tako ni dovolj, da besedilo lepo teče, a obenem ni primerno po namenu in vsebini. Pri analizi so torej pomembni vsi trije vidiki, ki niso nujno soodvisni [23].

Ena možnost ocenjevanja kakovosti prevoda je preverjanje vsebine z nalogami za bralno razumevanje. Ustreznost besedila je mogoče preprosto preveriti pri prevodih navodil za uporabo. Bralca se preprosto vpraša, ali bi z dotičnimi navodili (prevodom) lahko izvedel želeno dejanje. Med možnostmi sta tudi povratni prevod ter primerjava informativnosti izvirnika in strojnega prevoda.

Do oprijemljivejše ocene je mogoče priti s statističnim pristopom, pri katerem se primerja število napak različnih vrst. Čeprav se s štetjem napak pridobi objektivne številčne rezultate, tudi ta pristop ni povsem objektiven, saj vsak ocenjevalec napake vidi drugače.

Nekdo na neko stilistično pomanjkljivost, ki ne vpliva na razumljivost in točnost, ne gleda kot na napako, kdo drug pa [23].

Prav tako je kot način ocenjevanja pomembno, kdo podaja oceno strojnega prevajalnika. Mnenje raziskovalca bo do neke mere drugačno od mnenja uporabnika, saj raziskovalca zanima potencial orodja na dolgi rok, uporabnika pa trenutna uporabnost [23].

2.2.1 Ročna evalvacija

Leta 1966 je ALPAC objavil raziskavo na temo ročne evalvacije, v kateri so ocenjevali človeške in strojne prevode. Ocenjevalci so bili usposobljeni posebej za raziskavo. Ocenjevali so prevode iz ruščine v angleščino z vidika razumljivosti (angl. *intelligibility*) in natančnosti (angl. *fidelity*). Najprej so z ocenami od 1 do 10 vrednotili berljivost prevoda − v kolikšni meri je bil prevod razumljiv, ne da bi ocenjevalci videli original. Nato so z ocenami od 1 do 10 ocenjevali informativnost, tj. semantično pravilnost prevoda v primerjavi z izvirnikom − ali so vsi podatki oziroma vse informacije, ki jih vsebuje izvirnik, ohranjene ali katera manjka, ali je v prevodu kakšen podatek dodan, ali so katere besede ostale neprevedene ipd. Raziskava je pokazala, da so bile razlike med ocenjevalci majhne, kljub temu pa priporočajo, da pri evalvaciji sodelujejo vsaj trije ali štirje ocenjevalci. Ocenjevalci so zlahka ločili, ali gre za človeški ali za strojni prevod.

V podjetju ARPA (Advanced Research Projects Agency) so leta 1991 pod okriljem projekta Human Language Technologies Program vzpostavili evalvacijski program, ki deluje še danes, in izdelali metodologijo za evalvacijo strojnih prevajalnikov. Glavni izziv pri evalvaciji je bil zmanjšati subjektivnost – ocenjevanje mora biti intuitivno in hkrati kar se da objektivno, kar se kaže v minimalnih odstopanjih med ocenjevalci. Najprimernejše metode, ki so jih izbrali za nadaljnjo uporabo, so vključevale evalvacijo razumljivosti s pomočjo testov razumevanja, evalvacijo primernosti, ki so jo izvedli naravni govorci angleščine, in evalvacijo, ki temelji na primernosti, berljivosti in informativnosti [58].

2.2.2 Povratni prevod

Povratni prevod (angl. *round-trip translation*) je prevod, ki smo ga s pomočjo istega strojnega prevajalnika najprej prevedli v ciljni jezik, nato pa nazaj v izvirni jezik. Izvirni jezik je tisti jezik, iz katerega želimo prevesti, ciljni jezik pa je tisti, v katerega želimo prevesti. Največja težava pri tem je, da ne moremo vedeti, ali je sistem naredil napako ob prevajanju v drug jezik, ali ob prevajanju nazaj v izvirni jezik [40].

Spodnji primeri prikazujejo, kako je lahko povratno prevajanje za evalvacijo zavajajoče. V prvem primeru [40] je prevod v italijanščino popolnoma sprejemljiv, medtem ko je v povratnem prevodu kar nekaj napak. V drugem primeru [40] je povratni prevod identičen izvirniku, medtem ko je prevod v portugalščino brezpredmeten. Pri tretjem primeru, ki je prikazan s pomočjo prevajalnika Google, je prevod slovenskega frazema v angleščini popolnoma nesmiseln, povratni prevod pa je sicer slovnično pravilen, a nepomemben, saj se ob zamenjani informaciji izgubi pomen fraze.

Angleški izvirnik: **Tit for tat.**
Prevod v slovenščino: **Milo za drago.**
Povratni prevod: **Tit for tat.**

Slovenski izvirnik: **Ne tič ne miš.**
Prevod v angleščino: **Not fish not fowl.**
Povratni prevod: **Ne rib ne kokoš.**

2.2.3 Samodejna evalvacija

Ko je v začetku devetdesetih let ameriška vlada sponzorirala tekmovanje med strojnimi prevajalniki, so vse prevode ocenjevali ročno. Visoki stroški, subjektivnost in porabljen čas so spodbudili številne raziskovalce, da so začeli iskati objektivnejšo in hitrejšo rešitev [58]. Metrike za samodejno evalvacijo so na voljo kot brezplačna računalniška orodja, napisana v različnih programskih jezikih, kot so Perl, Java itd.: BLEU, METEOR, GTM, TER, ROUGE, WER in NIST.

BLEU

Bilingual Evaluation Understudy (BLEU) [39] je bila prva in še vedno najbolj razširjena metrika za evalvacijo kakovosti prevodov sistemov strojnega prevajanja. Kakovost prevodov je predstavljena kot natančnost ujemanja prevodov sistemov za strojno prevajanje z referenčnimi prevodi poklicnih prevajalcev. Vrednosti so izračunane za posamezne prevedene odseke, navadno povedi, in povprečne za celoten testni korpus. Berljivost in slovnična pravilnost nista upoštevani. BLEU uporablja spremenjeno različico natančnosti (angl. *precision*), ki predstavlja število pravilno klasificiranih elementov (angl. *true positives*) za primerjavo prevoda z enim ali več referenčnimi prevodi. Sprememba z osnovno natančnostjo naj bi poskrbela za lastnost sistemov strojnega prevajanja, ki težijo k daljšim prevodom.

METEOR

Metric for Evaluation of Translation with Explicit ORdering (METEOR) [12] temelji na harmonični sredinski natančnosti in priklicu unigramov (angl. *unigram precision and recall*), kjer je priklic močneje utežen kot natančnost. Vsebuje še več metod jezikovnih tehnologij, ki niso prisotne pri ostalih samodejnih metrikah strojnega prevajanja, kot so krnjenje in ujemanje sinonimov kot pomoč pri iskanju ujemanja besed. Krnjenje je predvsem primerno za visoko pregibne jezike, saj omejuje vpliv napačne uporabe pregibanja, npr. napačne uporabe sklona pri samostalnikih.

GTM

General Text Matcher (GTM) [20] meri podobnost med dvema besediloma na podlagi skupnih ngramov med strojnim in referenčnim prevodom prevajalcev. GTM je še ena izmed metrik strojnega prevajanja, katera ne dodeljuje enakovredno utež ngramom, ampak jih nagradi.

ROUGE

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [32] je množica metrik in programskih paketov, ki se uporabljajo za ocenjevanje samodejnih povzetkov in strojnega prevajanja v procesiranje naravnih jezikov. Metrike primerjajo strojno generiran povzetek oz. prevod z referenčnimi povzetki oz. prevodi.

Na voljo je naslednjih pet metrik:
ROUGE-N je metrika, ki gleda na sopojav ngramov. Sočasna pojavnost ngramov je običajno izvedena odsek po odsek, kjer je odsek najmanjša enota skladnosti prevoda, ponavadi en ali več stavkov.

ROUGE-L je metrika, ki temelji na najdaljšem skupnem podnizu (angl. *longest common subsequence*). Ideja te metrike je, da dolgi skupni podnizi izražajo veliko prekrivanje med strojnim in referenčnim prevodom.

ROUGE-W je metrika, ki prav tako temelji na LCS, le da dodeli večje uteži podnizom, ki imajo manj prekrivanj. Vendar ti dve metriki še vedno ne razlikujeta med prevodi z istim najdaljšim skupnim podnizom, pač pa z različnim številom krajših podnizov.

Zato obstaja metrika ROUGE-S, ki poskuša popraviti ta problem z združevanjem natančnosti in priklica (angl. *precision and recall*) skipbigrama referenčnih prevodov s strojnimi prevodi. Skipbigram je vsak par besed v stavku.

ROUGE-SU je metrika, ki gleda na sopojav unigramov in skipbigramov.

NIST

National Institute of Standards and Technology (NIST) je metoda za ocenjevanje kvalitete besedila, ki je bilo strojno prevedeno. Temelji na metriki BLEU, vendar z nekaj spremembami. Medtem ko BLEU preprosto izračuna natančnost ngramov in vsakemu

doda enakovredno težo, NIST računa, kako informativen je določen ngram. Torej, ko najdemo pravi ngram, mu bomo dodelili večjo težo, če je redkejši.

Npr., če imamo bigram "kako si", bo dobil manjšo težo kot bigram "informacijske tehnologije", saj se bo slednji pojavil manjkrat.

TER in TERp

Translation Error Rate (TER) [44] je metrika za napake pri strojnem prevajanju, ki meri število urejanj, potrebnih za spremembo sistemskega izhoda v enega izmed referenčnih.
TER-Plus (TERp) [44] je metrika, ki je bila predstavljena na delavnici MATR 2008. Je močnejše orodje kot TER, ki je bilo zasnovano tako, da je v korelaciji s človeškimi presojami.

3 UGLAŠEVANJE PARAMETROV IN DIFERENCIALNA EVOLUCIJA

V tem poglavju opišemo sorodna dela s področja uglaševanja parametrov in algoritem diferencialne evolucije.

3.1 Sorodna dela

3.1.1 Minimum Error Rate Training (MERT)

Algoritem Minimum Error Rate Training (MERT) [36] je trenutno najpopularnejši način za uglaševanje parametrov pri statističnem strojnem prevajanju. Algoritem je razumljiv, enostaven za implementacijo in hiter, a se lahko obnaša nepredvidljivo in ne podpira uporabe več kot 20-30 lastnosti. To je pomembna pomanjkljivost, saj zavira sisteme, da ne morejo uporabljati več kot 30 lastnosti.

MERT deluje tako, da poišče tiste uteži, ki minimizirajo napako (angl. *error measure*) ali maksimizirajo določeno metriko. Glavna značilnost tega algoritma je, da izhaja iz n-najboljših prevodov za podan stavek. S tem dosežemo hitro konvergenco optimizacijskega procesa. Najpogosteje se uporablja metrika BLEU, načeloma pa se lahko uporabi katerakoli samodejna metrika. MERT je vedno v nevarnosti, da se ujame v lokalni maksimum in ker uporablja n-najboljše sezname kot približek za izhod dekodirnika, ne more raziskati dejanskega prostora parametrov. Kljub vsem tem omejitvam daje MERT dobre rezultate.

Da bi odpravili pomanjkljivost z lastnostmi, so začeli razvijati druge pristope, s katerimi bi rešili ta problem. Eden izmed teh algoritmov je Margin Infused Relaxed Algorithm (MIRA).

3.1.2 Margin Infused Relaxed Algorithm (MIRA)

MIRA [21] se deli na dva pristopa:

- Online MIRA
- Batch MIRA (paralelizacija Online MIRA)

3.1.2.1 Online MIRA

Online MIRA je metoda, ki zahteva tesnejšo integracijo z dekodirnikom. Dekodirnik dekodira stavek po stavek iz razvojne množice, uteži modela prevajanja se posodobijo glede na izhod dekodirnika, nato gre na naslednji stavek. Ta proces se lahko ponovi večkrat. Ker je to metodo zelo težko paralelizirati, so razvili še Batch MIRA.

3.1.2.2 Batch MIRA

Celotna razvojna množica je dekodirana, običajno se ustvari seznam n-najboljših (angl. *n-best list*), nato so uteži modela prevajanja posodobljene glede na izhod dekodirnika. Ta iterativen proces se ponavlja, dokler ni zadoščeno konvergenčnemu kriteriju. Te množice so dovolj male, da jih lahko naložimo v pomnilnik in je zato učenje zelo hitro.

3.1.3 Pairwise Ranking Optimization (PRO)

Pri tem pristopu gledajo na optimizacijo kot na razvrstitveni problem, kjer je izrecni cilj naučiti, kako pravilno razvrstiti različne prevedene stavke za izvorni stavek. Razvrstitveni problem je zmanjšan na opravilo binarnega klasifikacijskega (angl. *binary classification task of deciding*) odločanja med prevodi parov.

3.2 Diferencialna evolucija

Algoritem diferencialne evolucije (angl. *differential evolution*) [50] [3] [4] [11] [35] je enostaven in učinkovit algoritem za globalno optimizacijo. Zaradi svoje enostavnosti in

učinkovitosti je uporabljen za reševanje različnih praktičnih problemov. Algoritem DE lahko opišemo kot preprost matematični model kompleksnega evolucijskega procesa. Uvrščamo ga med evolucijske algoritme. Matematični model diferencialne evolucije je zasnovan na uporabi razlik med vektorji oziroma posamezniki. Razlike med posamezniki se določajo s pomočjo hitrih in enostavnih aritmetičnih operacij, ki ločujejo DE od ostalih metod za globalno optimizacijo.

Algoritem DE je realno kodiran populacijski algoritem, ki v vsaki generaciji G vsebuje trenutno populacijo $\vec{P}_{x,G}$ in poskusno populacijo $\vec{P}_{x',G}$. Trenutna populacija vsebuje posameznike, ki so preživeli prejšnjo generacijo, medtem ko poskusna populacija vsebuje posameznike, ki jih je ustvaril algoritem DE v trenutni generaciji in bodo tekmovali proti posameznikom iz trenutne generacije. Obe populaciji vsebujeta N_p posameznikov in njihovih ocenitev. Posamezniki so predstavljeni v obliki D dimenzionalnih vektorjev. Elementi vektorjev so realna števila, katerih vrednosti se nahajajo na določenih intervalih. Intervale parametrov določa uporabnik s pomočjo spodnjih in zgornjih mej.

Algoritem DE vsebuje nekaj parametrov, s katerimi lahko spreminjamo lastnosti algoritma in ga prilagajamo različnim problemom:

- parameter mutacije (F),
- parameter križanja (C_r),
- velikost populacije (N_p),
- dimenzija problema (D),
- strategija algoritma DE (s),
- spodnje meje parametrov ($\vec{x}_{min}$),
- zgornje meje parametrov ($\vec{x}_{max}$) in
- maksimalno število generacij ($GENMAX$).

Parametri F, C_r, N_p, s in G predstavljajo krmilne parametre in vplivajo na obnašanje algoritma. Parametre $D, \vec{x}_{min}$ in $\vec{x}_{max}$ določa problem, ki ga rešujemo. Glede na tip problema, lahko izberemo med več različnimi strategijami algoritma DE. Strategija določa način mutacije in križanja.

V nadaljevanju predstavimo algoritem DE z eno najbolj razširjenih in uporabljenih strategij *DE/rand/1/bin*. Prikazan algoritem je preprost, vsebuje inicializacijo in operacije, kot so mutacija, križanje, popravljanje in selekcija.

Algoritem 1: Osnovni algoritem diferencialne evolucije

```
// *** DEAlgorithm() ***
initialization(P, N_p, D, x_min , x_max )

for G := 0 to GENMAX
begin
       for i := 0 to N_p
       begin
              m = mutation(P_x,G, i, F, N_p, D)
              k = crossover(x_i, m, C_r, D)
              x' = repair(k, D, x_min , x_max )
       end for
selection(P_x,G , P_x',G , N_p , D)
end for
```

3.2.1 Inicializacija

Preden algoritem začne uglaševati parametre, inicializiramo vektorje z naključnimi vrednostmi na intervalih, ki jih določata $\vec{x}_{min}$ in $\vec{x}_{max}$. Za inicializirane vektorje se nato še izračuna njihova ocena (e_i). Enačba za inicializacijo vektorjev:

$$x_{i,j,0} = x_{min,j} + \left(x_{max,j} - x_{min,j}\right) * rand_{i,j},$$

kjer spremenljivka $rand_{i,j}$ predstavlja naključno vrednost na intervalu [0, 1).

Algoritem 2: Inicializacija

```
// *** initialization() ***
for i := 1 to N_p
begin
       for j := 0 to D
       begin
              x_{i,j} = x_{min_j} + (x_{max_j} − x_{min_j}) * rand()
       end for

       e_i = fitness(x_i)
end for
```

3.2.2 Mutacija

Operacija mutacije izbere vektor iz trenutne populacije kot osnovni vektor $\vec{x}_{r_{1,i},G}$ in ga premakne v določeno smer. Smer premikanja vektorja določa vektor razlike, ki je določen z razliko med dvema naključno izbranima vektorjema iz trenutne populacije $\vec{x}_{r_{2,i},G}$ in $\vec{x}_{r_{3,i},G}$. Ta vektor še skaliramo s pomočjo krmilnega parametra F, ki je običajno manjši od 1 in je definiran na intervalu [0, 2]. Ta skaliran vektor razlike se prišteje osnovnemu vektorju in kot rezultat dobimo mutiran vektor $\vec{m}$. Enačba za mutacijo:

$$\vec{m} = \vec{x}_{r_{1,i},G} + F * (\vec{x}_{r_{2,i},G} - \vec{x}_{r_{3,i},G}),$$

kjer $r_{1,i}, r_{2,i}$ in $r_{3,i}$ predstavljajo indekse naključno izbranih posameznikov iz populacije, ki so med seboj različni in hkrati različni od indeksa i.

Algoritem 3: Mutacija

```
// *** mutation() ***
do
        r_{1,i} = rand {0, N_p − 1}
while r_{1,i} == i

do
        r_{2,i} = rand {0, N_p − 1}
while r_{2,i} == i or r_{2,i} == r_{1,i}

do
        r_3 = rand {0, N_p − 1}
while r_{3,i} == i or r_{3,i} == r_{2,i} or r_{3,i} == r_{1,i}

for j := 0 to D
begin
        m_j = x_{r_{1,i},j} + F * (x_{r_{2,i},j} − x_{r_{3,i},j})
end for
```

3.2.3 Križanje

Algoritem DE ima dva načina križanja:

- binarno in
- eksponentno.

Predstavili bomo binarno križanje. Za vsak element vektorja izračuna naključno vrednost *rand()* na intervalu [0, 1). Če je ta vrednost večja od krmilnega parametra $C_r \in [0, 1]$, se elementu križanega vektorja k_j priredi element mutiranega vektorja m_j. V nasprotnem primeru element križanega vektorja dobi vrednost elementa vektorja iz trenutne populacije $x_{i,j,G}$. Enačba za križanje:

$$k_j = \begin{cases} m_j, & \text{če } rand() < C_r \text{ ali } rand_j == i \\ x_{i,j,G}, & drugače \end{cases},$$

kjer $rand_j$ vrne vrednost na intervalu {1, D} in je odgovorna, da križan vektor vsebuje vsaj en element mutiranega vektorja.

Algoritem 4: Križanje

```
// *** crossover() ***
rand_j = rand{0, D − 1}
for j := 0 to D
begin
        rand_{i,j} = rand[0,1)
        if rand_{i,j} < C_r or rand_j == j
        begin
                k_j = m_j
        else
                k_j = x_{i,j}
        end if
end for
```

3.2.4 Popravljanje

Po operaciji križanja dobimo vektor križanja, katerega elementi k_j se lahko nahajajo izven določenih intervalov $[x_{min,j}, x_{max,j}]$, zato je vektor križanja potrebno popraviti. Pri popravljanju imamo dve strategiji:

- vrednosti, ki so izven intervalov, nastavimo na bližnjo spodnjo oz. zgornjo mejo ali
- vrednosti, ki so izven intervalov, prezrcalimo nazaj v interval.

Enačba za zrcaljenje:

$$x'_{i,j,G} = \begin{cases} x_{min,j} + (x_{min,j} - k_j), & \text{če je } k_j < x_{min,j} \\ x_{max,j} + (k_j - x_{max,j}), & \text{če je } k_j > x_{max,j} \\ k_j, & drugače. \end{cases}$$

Popravljen vektor je hkrati tudi poskusni vektor (angl. *trial vector*).

Algoritem 5: Popravljanje

```
// *** repair() ***
if k_j < x_min,j
begin
        x_i,j = x_min,j + (x_min,j − k_j)
else
        if k_j > x_max,j
        begin
                x_i,j = x_max,j − (k_j − x_max,j)
        end if
else
        x_i,j = k_j
end if
```

3.2.5 Selekcija

Selekcija določa, kateri posamezniki bodo preživeli v naslednjo generacijo. V ta namen se primerjata posameznika z enakim indeksom *i* iz trenutne in poskusne populacije. Odločitev o preživetju določa ocena (angl. *fitness*) posameznikov. To vrednost določa problem, ki ga rešujemo. V naslednjo generacijo preživi posameznik, ki ima boljšo oceno. Če imata isto oceno, v naslednjo generacijo preživi posameznik iz trenutne populacije. Enačba selekcije za problem, kjer iščemo globalni maksimum:

$$x_{i,G+1} = \begin{cases} x'_{i,G}, & \text{če je } f(x'_{i,G}) > f(x_{i,G}) \\ x_{i,G}, & drugače. \end{cases}$$

Algoritem 6: Selekcija

```
// *** selection() ***
for i := 0 to N_p
begin
        et_i = fitness(x'_i)
        if et_i > e_i
        begin
                x_i = x'_i
                e_i = et_i
        end if
end for
```

4 EKSPERIMENT

V tem poglavju najprej opišemo načrt eksperimenta, nato opišemo izboljšavo sistema za strojno prevajanje. Sledi podroben opis našega pristopa za uglaševanje parametrov.

4.1 Načrtovanje eksperimenta

4.1.1 Zbirka JRC-ACQUIS

Pred vstopom v Evropsko unijo (angl. *European Union*) je potrebno novim državam članicam (angl. *New Member States*) prevesti in potrditi obstoječo zakonodajo EU, ki je sestavljena iz izbranih besedil, napisanih med leti 1950 in 2005. Ta organ zakonodajnega besedila, ki je sestavljen iz približno osem tisoč dokumentov in zajema različne domene, se imenuje pravni red EU (angl. *Acquis Communatuaire)*. V začetku leta 2005 je bilo v EU 20 uradnih jezikov, tako da obstajajo vzporedna besedila (besedila in prevodi le-teh). Ti jeziki so: češčina, danščina, nemščina, grščina, angleščina, španščina, estonščina, finščina, francoščina, madžarščina, italijanščina, litovščina, latvijščina, malteščina, nizozemščina, poljščina, portugalščina, slovaščina, slovenščina in švedščina.

Nekatere države, ki so kandidatke za EU, so že začele prevajati AC, tako da so nekateri dokumenti že na voljo. Hrvaščina in bolgarščina trenutno nista del distribucije.

Jezikovni raziskovalni interes pravnega reda Evropske unije

V računalniškem jezikoslovju so vzporedna besedila koristni viri, ki so uporabljena za različne aplikacije in namene. Večina vzporednih besedil obstaja za majhno število jezikov. Acquis Communatuaire (AC) je največja obstoječa zbirka vzporednih besedil, če upoštevamo velikost in število jezikov, ki jih zajema.

Ekipa jezikovnih tehnologij skupnega raziskovalnega centra (angl. *Joint Research Centre*) [27] je poskušala poiskati dokumente, ki so del AC, jih prenesla in pretvorila v obliko XML. V naslednjih korakih so iz besedil odstranili nogo in priloge in jih poravnali po stavkih (angl. *sentence-aligned*) za vsak jezikovni par.

Za nekatere dokumente so bili na voljo le predhodni prevodi. Za spletna besedila so bili v nekaterih jezikih prevedeni samo naslovi, a prikazano besedilo je bilo v angleščini. Uporabljeno je bilo orodje za samodejno zaznavanje jezika, da izločijo tista besedila, ki so prikazana kot en jezik. Ročno preverjanje ni bilo opravljeno.

Statistika

V tabeli 2 imamo statistiko za zbirko dokumentov ACQUIS (verzija 2.2), kjer je 21 jezikov. Sedaj je na voljo že nova verzija zbirke dokumentov ACQUIS (verzija 3.0), kjer je 22 jezikov.

Vsi dokumenti so bili preneseni iz spleta [14] [8] in imajo številčno oznako (kodo CELEX), prikazano na sliki 2. Ta oznaka pomaga poiskati isto besedilo v različnih jezikih. Vsi preneseni dokumenti so bili pretvorjeni v obliko XML in kodirani z UTF-8.

Tabela 2: Zbirka dokumentov ACQUIS (verzija 2.2) v 21 jezikih

Jezik	Število dokumentov	Telo dokumenta			Podpis	Priloga	Število vseh besed (Dokument + podpis + priloga)
		Število vseh besed	Število vseh znakov	Povprečno število besed	Število vseh besed	Število vseh besed	
CS	7983	5979261	38479314	749	609441	2100301	8689003
DA	7939	6548461	44444011	825	691894	1599456	8839811
DE	7914	6576633	47047334	831	571928	1506847	8654608
EL	7782	7377316	47715936	948	559487	1628451	9565254
EN	7972	7512013	45150120	942	667978	1752545	9932536
ES	7809	**7964255**	**48281455**	**1020**	709279	1832745	**10506279**
ET	7944	4925361	38603952	**620**	439184	1819226	7183771
FI	7735	5134294	43705813	664	565226	**1180877**	6880397
FR	7862	7812577	45609935	994	673061	1726720	10212358
HU	7489	5391810	40601868	720	539967	1887476	7819253
IT	7872	7264126	46792286	923	707467	1704221	9675814
LT	7966	5386359	39936370	676	625365	1948354	7960078
LV	7980	5656335	39290110	709	461736	2011426	8129497
MT	7639	7230538	43919981	947	505324	2288013	10023875
NL	7882	7339465	47699598	931	**712255**	1710041	9761761
PL	7968	5974605	43160945	750	668248	2070687	8713540
PT	7848	7851904	47225710	1001	648180	1838833	10338917
RO	5792	5122354	33681450	884	**402929**	**4047393**	9572676
SK	**5278**	**3911895**	**26077956**	741	413511	1381471	**5706877**
SL	**7984**	5989322	37844883	750	573052	2153138	8715512
SV	7731	6472717	42990411	837	560188	1424887	8457792
Povprečje	**7636**	**6353410**	**42340925**	**831**	**585947**	**1886338**	**8825695**

```
<TEI.2 id="jrcCELEX-LG" n="CELEX" lang="LG">
<teiHeader lang="en" date.created="DATE">
<fileDesc>
  <titleStmt>
    <title>JRC-ACQUIS CELEX LANGUAGE</title>
    <title>Document Title</title>
  </titleStmt>
  <extent>nb_of_paragraphs paragraph segments</extent>
  <publicationStmt>
    <distributor>
      <xref url="http://wt.jrc.it/lt/acquis/">http://wt.jrc.it/lt/acquis/</xref>
    </distributor>
  </publicationStmt>
  <notesStmt>
    ....
  </notesStmt>
  <sourceDesc>
      <bibl>Downloaded from <xref url="Downloading_URL">Downloading_URL</xref> on <date>Downloading_DATE</date></bibl>
  </sourceDesc>
</fileDesc>
<profileDesc>
    <textClass>
        <classCode scheme="eurovoc">Eurovoc_Code</classCode>
          .....
    </textClass>
</profileDesc>
</teiHeader>
<text>
  <body>
    <head n="1">Document Title</head>
    <div type="body">
      <p n="paragraph_number">... TEXT...</p>
      .......
    </div>
    <div type="signature">
      <p n="paragraph_number">... signature text...</p>
        ....
    </div>
    <div type="annex">
      <p n="paragraph_number">... annex text...</p>
        ....
     </div>
  </body>
</text>
</TEI.2>
```

Slika 2: Oblika dokumenta DTD (Document Type Definition)

Poravnava stavkov za vse jezike

Zbirka besedil je poravnana na nivoju odstavka. En odstavek je običajno kratek in vsebuje en stavek ali samo del stavka. Za poravnavo sta na voljo naslednja dva pristopa:

- Vanilla in
- Hun.

Za vsak jezikovni par (vseh možnih parov za 21 jezikov je 210) so ustvarjene datoteke, ki vsebujejo informacije o poravnavi določenega jezikovnega para. Te datoteke vsebujejo kazalec na odstavke, ki so med seboj prevodi. Uporabljena je bila oblika TEX (Text

Encoding Initiative). Zaradi velikosti zbirke in števila jezikovnih parov, te datoteke ne vsebujejo samega besedila.

4.1.2 Orodje Moses

Orodje Moses [28] je odprtokodno orodje za statistično strojno prevajanje (angl. s*tatistical machine translation*). Sestavljeno je iz vseh komponent, ki so potrebne za vnaprejšnjo obdelavo podatkov in učenje jezikovnih ter prevajalnih modelov. Vsebuje tudi orodja za optimizacijo modelov z uporabo pristopov Minimum Error Rate Training (MERT) in Margin Infused Relaxed Algorithm (MIRA). Za evalvacijo prevodov uporablja metriko Bilingual Evaluation Understudy (BLEU). Za nekatera opravila se uporabijo zunanja orodja, kot so:

- GIZA++ [37] za poravnavo besed in
- SRILM [49], IRSTLM [19], KENLM [22] za modeliranje jezika.

Ker ta opravila običajno porabijo veliko procesorske moči, je bilo orodje zasnovano za delo s paralelnim okoljem Sun Grid Engine za povečanje prepustnosti (angl. *throughput*). Glavna komponenta orodja je dekodirnik. Razvili so ga kot zamenjavo za zaprto kodni dekodirnik Pharaoh [29] – zelo popularni frazni dekodirnik. Orodje Moses ima vse zmogljivosti orodja Pharaoh. Med razvijanjem orodja Moses so se razvijalci držali naslednjih principov:

- dostopnost,
- enostavno za vzdrževanje,
- fleksibilnost,
- enostavno za porazdeljeno delo v skupini in
- prenosljivost.

Razvit je bil v programskem jeziku C++, zaradi učinkovitosti in objektno-orientiranega pristopa. Orodje od začetka gostuje in se razvija pod okriljem sourceforge.net ter ima aktivno raziskovalno skupnost. Dosega primerljive rezultate z najbolj konkurenčnimi in pogosto uporabljenimi statističnimi strojnimi prevajalnimi sistemi, tako v kvaliteti prevajanja kot v času delovanja.

4.1.3 Sistem za strojno prevajanje

Po navodilih iz [5] smo zgradili sistem za strojno prevajanje iz slovenščine v angleščino. Za delo z jezikovnimi modeli smo uporabili orodje IRSTLM [19], za poravnavo besed pa orodje GIZA++ [37]. Za učenje sistema smo uporabili slovenski in angleški korpus JRC-ACQUIS. Vzeli smo prvih 1.160.000 stavkov, ki so poravnani v obeh jezikih. Nato smo besedilo segmentirali. Nato smo odstranili stavke, ki so daljši od 80 besed (predolgi stavki običajno dajejo slabše rezultate) in dobili 1.068.691 stavkov.

Na podlagi nove učne množice smo zgradili petgramski jezikovni model (angl. *language model),* ki uporablja *Kneser Ney* glajenje frekvenc. Ngram je zaporedje n-tih elementov iz besedila ali govora. Elementi so lahko fonemi, zlogi, črke, besede ali bazni pari v skladu z aplikacijo. Ngram velikosti 1 se imenuje *unigram,* ngram velikosti 2 se imenuje *bigram* in ngram velikosti 3 se imenuje *trigram*. Večje velikosti pa se imenujejo po *n,* torej ngram za velikost 4 je štirigram, ngram za velikost 5 je petgram itd. Nato smo zgradili model za prevajanje (angl. *translation model*). Uporabili smo poravnavo *grow dialog final and* in prerazporeditev *msd bidirectional fe*. Tako smo dobili tabelo prevodov fraz in tabelo prerazporeditve, ki sta potrebni za prevajanje.

Sistem za strojno prevajanje lahko tudi izboljšamo glede na:

- hitrost:
    - omejimo iskalni prostor dekodirnika (s tem dobimo hitrejše prevajanje, vendar slabše prevode) in
- kakovost:
    - dobra tabela prevodov fraz (ustvari se v fazi učenja) in/ali
    - dobre uteži (parametra) modela.

4.2 Izboljšava sistema za strojno prevajanje

V našem delu se bomo osredotočili na kakovost prevajanja z uglaševanjem parametrov modela prevajanja.

Uglaševanje parametrov se nanaša na postopek iskanja optimalnih uteži za linearni model, kjer so optimalne uteži tiste, ki povečajo učinkovitost prevajanja na majhni razvojni množici vzporednih stavkov (angl. *held-out set*). Učinkovitost prevajanja običajno merimo z metriko BLEU.

V statističnem strojnem prevajanju obstajata dva razreda algoritmov za uglaševanje:

- *Batch*: Celotna razvojna množica je dekodirana, ponavadi se ustvari seznam n-najboljših prevodov (angl. *n-best list*), potem so uteži modela posodobljene glede na izhod dekodirnika. Ta iterativen proces se ponavlja dokler ni zadoščeno konvergenčnemu kriteriju.
- *Online*: Te metode zahtevajo tesnejšo integracijo z dekodirnikom. Slednji dekodira stavek po stavek iz razvojne množice, uteži se posodobijo glede na izhod dekodirnika, nato gre na naslednji stavek. Ta proces se lahko ponovi večkrat.

4.3 Naš pristop za uglaševanje parametrov

Naš pristop za uglaševanje parametrov pri statističnem strojnem prevajanju temelji na algoritmu DE.

V procesu uglaševanja je potrebno posameznike oceniti. Glede na dobljeno oceno se s pomočjo selekcije določi, kateri posamezniki bodo preživeli v naslednjo generacijo. Za ocenjevanje posameznikov smo v algoritem DE vključili kvaliteto prevoda. Za vsakega posameznika prevedemo besedilo iz izvornega jezika v ciljni jezik. Prevedeno besedilo nato primerjamo z referenčnimi besedili in s pomočjo metrike BLEU [39] dobimo odstotek podobnosti, kar nam predstavlja oceno.

V glavnem programu najprej nastavimo začetne uteži oz. parametre. Nato za posamezno generacijo uporabimo operacije, kot so mutacija, križanje in popravljanje, da dobimo nove posameznike. Nove posameznike ocenimo in jih primerjamo s trenutnimi posamezniki iz generacije. Če je novi boljši, preživi v naslednjo generacijo, drugače pa ostane trenutni posameznik.

Algoritem 7: Glavni program našega pristopa

```
// *** main() ***
inicializacija(x⃗_min , x⃗_max )

for G := 0 to GENMAX
begin
       for i := 0 to N_p
       begin
              rand1bin(i,x⃗_min , x⃗_max , x⃗ ') // strategija DE/rand/1/bin

              // selection
              e_x⃗' = fitness(x⃗')
              if e_x⃗' > e_i
              begin
                     x⃗_i = x⃗'
                     e_i = e_x⃗'
              end if
       end for
end for

best = GetBest()
```

4.3.1 Inicializacija

Pri inicializaciji za prvega posameznika uporabimo parametre, ki jih uporablja orodje Moses (moses.ini), ostalim posameznikom iz populacije pa se določijo naključno na definiranih intervalih. Vrednosti za vsakega posameznika še normaliziramo (vsota = 1) in ga ocenimo.

Algoritem 8: Inicializacija

// *** inicializacija() ***
$min_0 = 0.0$
$max_0 = 3.0$

```
for i := 1 to D
begin
```
$min_i = 0.0$
$max_i = 1.0$
```
end for
```

$\vec{x}_0 =$ nastavimo privzete vrednosti iz datoteke moses.ini
$e_0 = fitness(\vec{x}_0)$ // pri ocenjevanju uporabimo metriko BLEU

```
for i := 1 to N_p
begin
    sum = 0
    for j := 0 to D
    begin
```
$x_{i,j,0} = x_{min_j} + \left(x_{max_j} - x_{min_j}\right) * rand()$
$sum += abs(x_{i,j,0})$
```
    end for

    // normalizacija
    for j := 0 to D
    begin
```
$x_{i,j,0} = \frac{x_{i,j,0}}{sum}$
```
    end for
```

$e_i = fitness(\vec{x}_i)$
```
end for
```

4.3.2 Ocenjevanje posameznikov

Da bi ocenili posameznike, je potrebno za vsakega prevesti izvorno besedilo v ciljno besedilo in ga nato primerjati z referenčnimi besedili. Uporabili smo slovenski in angleški korpus JRC-ACQUIS, ki sta med seboj poravnana po stavkih. Izvorno besedilo nato prevedemo z našim prevajalnim sistemom in ga ocenimo s pomočjo evalvacijske metrike BLEU. V sami ocenjevalni funkciji *fitness* kličemo zunanjo skripto *eval.sh*, ki vrne oceno o kakovosti prevoda s pomočjo metrike BLEU. V skripti *eval.sh* uporabimo dekodirnik moses iz orodja Moses za prevajanje iz izvornega v ciljni jezik. Vhod v dekodirnik so parametri posameznika in besedilo, ki ga želimo prevesti. Izhod je prevedeno besedilo, katerega nato primerjamo z referenčnim besedilom, ki so ga prevedli priznani prevajalci, in izračunamo oceno BLEU. Rezultat zapišemo v datoteko *ocena.txt*.

Algoritem 9: Ocenitvena funkcija

```
// *** fitness() ***
eval = system(eval.sh)
return eval
```

Algoritem 10: Vsebina skripte eval.sh

```
#!/bin/bash

moses -f moses.ini -parametri < razvojni_korpus.izvorni_jezik > razvojni_korpus.translated.ciljni_jezik

multi-bleu.perl -lc razvojni_korpus.ciljni_jezik < razvojni_korpus.translated.ciljni_jezik > ocena.txt
```

4.3.3 Ustvarjanje novih posameznikov

Posamezniki se ustvarjajo s pomočjo mutacije in križanja. Uporabili smo spremenjeno strategijo *DE/rand/1/bin*, kjer za vsakega posameznika v populaciji določimo naključno vrednost krmilnih prametrov F_i na intervalu [0.3, 0.6] in $C_{r,i}$ na intervalu [0, 1]. Vrednosti še normaliziramo (vsota = 1) za vsakega posameznika v populaciji.

Algoritem 11: Spremenjena strategija DE/rand/1/bin

```
// *** rand1bin() ***
r_{1,i}, r_{2,i}, r_{3,i}    // izračunamo indekse kot pri Algoritmu 3

F_i = rand(0.3, 0.6)
C_{r,i} = rand(0, 1)
jrand = rand{0, D - 1}

for j := 0 to D
begin
       if rand(0, 1) < C_{r,i} or j == jrand
       begin
              x'_{j,G} = x_{r_{1,i},j} + F * (x_{r_{2,i},j} - x_{r_{3,i},j})

              if x'_{j,G} < x_{min_j}
              begin
                     x'_{j,G} = x_{min_j} + (x_{min_j} - x'_{j,G})
              else
                     x'_{j,G} = x_{max_j} - (x'_{j,G} - x_{max_j})
              end if
       else
              x'_{j,G} = x_{i,j}
       end if

       sum += abs(x'_{j,G})
end for

for j := 0 to D
begin
       x'_{j,G} = x'_{j,G} / sum
end for
```

5 REZULTATI

5.1 Predobdelava korpusa

Iz zbirke JRC-ACQUIS smo vzeli 5663 stavkov za razvojno množico (angl. *held-out test*) in 5000 stavkov za testno množico. Obe množici smo ustrezno segmentirali ter odstranili prekratke in predolge stavke (stavke, ki so krajši od 8 in daljši od 60 besed), saj le-ti običajno dajejo slabše rezultate. Tako je ostalo 3475 stavkov pri razvojni množici, med katerimi smo zaradi časovne zahtevnosti našega pristopa, ki smo ga opisali v poglavju 4.2, vzeli prvih 1000 stavkov za vse metode (MERT, MIRA, PRO in naš pristop). Za testno množico je ostalo 3297 stavkov, s katerimi smo testirali uspešnost sistema za strojno prevajanje.

5.2 Opis parametrov

Za vsak model v modelu prevajanja imamo različno število uteži (parametrov). Število parametrov modela prevajanja je štirinajst, razdeljeni pa so na štiri dele:

- kazen za besede, ki ima en parameter (*w*),
- jezikovni model, ki ima en parameter (*lm*),
- model popačenja, ki ima sedem parametrov (*d*) in
- model prevajanja, ki ima pet parametrov (*tm*).

5.3 Rezultati uglaševanja

Sistem za strojno prevajanje smo uglaševali s tremi metodami za uglaševanje (MERT, MIRA in PRO) in z našim pristopom.

V tabeli 3 imamo parametre pred začetkom uglaševanja z metodo MERT. Ocena BLEU je 92,40 %. V tabeli 4 imamo parametre po končanem uglaševanju z metodo MERT. Ocena BLEU je 93,85 %. Opazimo, da se je kakovost prevajanja izboljšala.

V tabeli 5 imamo parametre pred začetkom uglaševanja z metodo MIRA. Ocena BLEU je 97,02 %. V tabeli 6 imamo parametre po končanem uglaševanju z metodo MIRA. Ocena BLEU je 97,51 %. Opazimo, da se je kakovost prevajanja izboljšala.

V tabeli 7 imamo parametre pred začetkom uglaševanja z metodo PRO. Ocena BLEU je 92,40 %. V tabeli 8 imamo parametre po končanem uglaševanju z metodo PRO. Ocena BLEU je 93,85 %. Opazimo, da se je kakovost prevajanja izboljšala.

V tabeli 9 imamo parametre pred začetkom uglaševanja z našim pristopom. Ocena BLEU je 76,44 %. V tabeli 10 imamo parametre po končanem uglaševanju z našim pristopom. Ocena BLEU je 79,37 %. Opazimo, da se je kakovost prevajanja izboljšala. Pri našem pristopu so na razvojni množici ocene BLEU dosti nižje kot pri ostalih sorodnih metodah, so pa zato boljši rezultati na testni množici, kar bomo prikazali v nadaljevanju.

Slika 3 prikazuje izboljševanje najboljšega posameznika v generaciji skozi vseh petdeset generacij.

V tabeli 11 imamo rezultate uglaševanja tako sorodnih metod kot našega pristopa. Prikazana je ocena BLEU v odstotkih in čas uglaševanja. Iz tabele je razvidno, da je naš pristop časovno zelo zahteven. Razlika je v tem, da MERT, MIRA in PRO prevedejo razvojno množico največ 26-krat, medtem ima naš pristop 51 generacij (vključno z začetno, ki je naključna) in v vsaki generaciji ima 15 posameznikov – torej je treba prevesti razvojno množico kar 765-krat. Naš osnovni namen je raziskati, ali lahko izboljšamo prevajanje, potemtakem časovna komponenta pri našem delu nima bistvenega pomena.

Iz tabele 11 je razvidno, da se naš rezultat uglaševanja močno razlikuje od sorodnih pristopov. Takšna razlika v rezultatih je zaradi uporabe različne metrike BLEU. Mi uporabljamo metriko BLEU, ki ovrednoti celotno razvojno množico (angl. *corpus-based*

BLEU), medtem ko sorodni pristopi uporabljajo metriko BLEU, ki ovrednoti vsak stavek oz. vrstico posebej in nato izračuna povprečje ocen vseh stavkov (angl. *sentence-based BLEU*).

Tabela 3: Parametri pred uglaševanjem z metodo MERT

w	**lm**	**d**	**d**	**d**	**d**	**d**
0,305884	0,0335813	0,0820533	0,00517957	0,00480919	0,140353	0,144638
d	**d**	**tm**	**tm**	**tm**	**tm**	**tm**
0,00313721	0,00541555	0,0340291	0,00110149	0,214733	-0,023130	0,00195432

Tabela 4: Parametri po končanem uglaševanju z metodo MERT

w	**lm**	**d**	**d**	**d**	**d**	**d**
0,120246	0,0516285	0,149959	0,0579584	0,112491	0,02487	0,0167933
d	**d**	**tm**	**tm**	**tm**	**tm**	**tm**
0,0304671	0,105837	0,0361942	0,0124244	0,151099	-0,013966	0,116066

Tabela 5: Parametri pred uglaševanjem z metodo MIRA

w	**lm**	**d**	**d**	**d**	**d**	**d**
0,127833	0,0591762	0,129880	0,066814	0,127833	0,0591762	0,129880
d	**d**	**tm**	**tm**	**tm**	**tm**	**tm**
0,026709	0,096332	0,107609	0,038050	0,026709	0,096332	0,107609

Tabela 6: Parametri po končanem uglaševanju z metodo MIRA

w	**lm**	**d**	**d**	**d**	**d**	**d**
0,128907	0,052240	0,109092	0,087393	0,128907	0,052240	0,109092
d	**d**	**tm**	**tm**	**tm**	**tm**	**tm**
0,014802	0,063417	0,088099	0,016616	0,014802	0,063417	0,088099

Tabela 7: Parametri pred uglaševanjem z metodo PRO

w	**lm**	**d**	**d**	**d**	**d**	**d**
0,073475	0,092890	0,126220	0,037556	0,073475	0,092890	0,126220
d	**d**	**tm**	**tm**	**tm**	**tm**	**tm**
0,069503	0,085655	0,162849	0,060279	0,069503	0,085655	0,162849

Tabela 8: Parametri po končanem uglaševanju z metodo PRO

w	**lm**	**d**	**d**	**d**	**d**	**d**
0,046032	0,074807	0,153645	0,166634	0,046032	0,074807	0,153645
d	**d**	**tm**	**tm**	**tm**	**tm**	**tm**

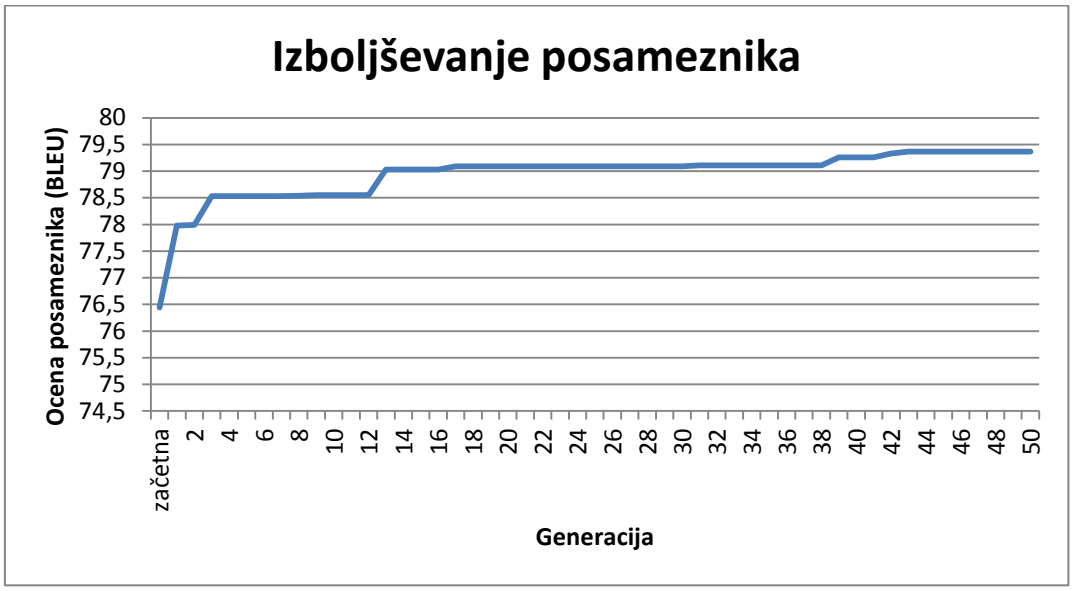
0,029611	0,031340	0,140247	0,021172	0,029611	0,031340	0,140247

Tabela 9: Parametri pred uglaševanjem z našim pristopom

w	**lm**	**d**	**d**	**d**	**d**	**d**
0,049206	0,114956	0,093271	0,099459	0,050806	0,043820	0,127831
d	**d**	**tm**	**tm**	**tm**	**tm**	**tm**
0,091974	0,097407	0,054818	0,008038	0,11067	0,053142	0,0045942

Tabela 10: Parametri po končanem uglaševanju z našim pristopom

w	**lm**	**d**	**d**	**d**	**d**	**d**
0,046114	0,085760	0,084082	0,057160	0,12605	0,092868	0,036349
d	**d**	**tm**	**tm**	**tm**	**tm**	**tm**
0,053884	0,029918	0,157649	0,015083	0,107181	0,005535	0,102362



Slika 3: Izboljševanje najboljšega posameznika

Tabela 11: Rezultati uglaševanja (ocena BLEU in čas v urah)

METODA	**BLEU (%)**	**Čas (ure)**
MERT	93,85	1
MIRA	**97,51**	3
PRO	97,06	3
NAŠ PRISTOP	79,37	200

5.4 Rezultati testiranja

V tabeli 12 imamo rezultate osnovnega sistema, treh sorodnih metod in našega pristopa. Parametre za vsako metodo, ki smo jih dobili z uglaševanjem, smo uporabili pri modelu prevajanja in prevedli na testni množici ter ocenili z metriko BLEU. Iz tabele je razvidno, da naš pristop daje primerljive oz. boljše rezultate kot ostale metode. Poudarimo, da so rezultati, ki so dobljeni na testiranju (tabela 12), precej bolj pomembni kot tisti, ki smo jih dobili ob uglaševanju (poglavje 5.3).

Tabela 12: Rezultati testiranja

METODA	**BLEU (%)**
Baseline	63,02
MERT	63,65
MIRA	64,01
PRO	65,03
NAŠ PRISTOP	**65,71**

V nadaljevanju bomo prikazali nekaj primerov prevajanj z osnovnim sistemom, s sorodnimi metodami in z našim pristopom.

Slovenski izvirnik: nelegirane jeklene cevi in fitingi za transport vodnih tekočin vključno s pitno vodo – tehnični dobavni pogoji

Referenčni prevod: wastewater lifting plants for buildings and sites — principles of construction and testing — part 3 : lifting plants for wastewater containing faecal matter for limited applications

Prevod v angleščino (Baseline): wastewater lifting plants for buildings and sites — principles of construction and testing — part 3 : lifting plants for wastewater containing faecal matter for limited applications

Prevod v angleščino (MERT): wastewater lifting plants for buildings and sites — principles of construction and testing — 3 . part : lifting plants for wastewater containing faecal matter for limited applications

Prevod v angleščino (MIRA): wastewater lifting plants for buildings and sites — principles of construction and testing — part 3 : lifting plants for wastewater containing faecal matter for limited applications

Prevod v angleščino (PRO): wastewater lifting plants for buildings and sites — principles of construction and testing — part 3 : lifting plants for wastewater containing faecal matter for limited applications

Prevod v angleščino (DE): wastewater lifting plants for buildings and sites — principles of construction and testing — 3 . part : lifting plants for wastewater containing faecal matter for limited applications

5.5 Statistična primerjava z orodjem MultEval

Orodje MultEval [34] izračuna oceno kakovosti, standardno deviacijo in p-vrednosti strojnih prevodov za tri popularne metrike (BLEU, METEOR, TER). Orodje je namenjeno za pomoč pri ocenjevanju učinka notranjih eksperimentalnih variacij na kakovosti prevodov. MultEval ne naredi segmentacije besedila, zato je potrebno besedila prej posebej segmentirati.

V tabeli 13 imamo primerjavo referenčnega besedila z osnovnim sistemom (Baseline) in z metodami za uglaševanje (MERT, MIRA, PRO ter naš pristop). Tabela 13 prikazuje izhod, ki ga ustvari orodje MultEval, ki je priznano orodje na področju statističnega strojnega prevajanja. V tabeli 13 imamo znaka ↑ in ↓, ki pomenita sledeče:

- znak ↑ kaže, da so višje vrednosti pri dani metriki boljše,
- znak ↓ kaže, da so nižje vrednosti pri dani metriki boljše.

Najboljše vrednosti smo označili v tabeli 13, kjer je razvidno, da je naš pristop po metrikah BLEU in METEOR boljši, pri metriki TER pa je boljša metoda PRO. Metrika LENGTH pove dolžino strojnih prevodov v primerjavi z dolžino referenčnega prevoda. Pri metriki METEOR lahko opazimo, da se naš pristop signifikantno ne razlikuje od osnovnega sistema za strojno prevajanje. Največ uporabljena metrika pri statističnem strojnem prevajanju je metrika BLEU, ki je ena izmed prvih metrik, ki je dosegla najvišjo korelacijo s človeško presojo o kakovosti prevoda [7] [13].

Tabela 13: Prikaz statistične primerjave z orodjem MultEval

Metric	**System**	**Avg**	$\overline{s}_{sel}$	s_{Test}	p**-value**
5*BLEU ↑	baseline	61,9	0,6	-	-
	mert	62,4	0,6	-	0,01
	mira	62,8	0,6	-	0,00
	pro	63,9	0,6	-	0,00
	naš pristop	**64,6**	0,6	-	0,00
5*METEOR ↑	baseline	47,6	0,3	-	-
	mert	46,6	0,3	-	0,00
	mira	46,7	0,3	-	0,00
	pro	47,4	0,3	-	0,03
	naš pristop	**47,7**	0,3	-	0,22
5*TER ↓	baseline	27,7	0,5	-	-
	mert	25,6	0,5	-	0,00
	mira	25,7	0,5	-	0,00
	pro	**25,1**	0,5	-	0,00
	naš pristop	25,2	0,5	-	0,00
5*Length	baseline	103,9	0,3	-	-
	mert	94,0	0,2	-	0,00
	mira	95,5	0,2	-	0,00
	pro	97,1	0,2	-	0,00
	naš pristop	99,2	0,2	-	0,00

6 SKLEP

V tem magistrskem delu smo uspešno vzpostavili sistem za strojno prevajanje. Sistem smo nato uglaševali s pomočjo algoritma diferencialne evolucije (DE). Pokazali smo, da naš pristop daje primerljive oz. boljše rezultate kot sorodne metode pri statističnem strojnem prevajanju. Čeprav so rezultati boljši, še ne pomeni, da bodo tudi prevodi dejansko boljši, saj je zelo težko natančno oceniti nek prevod. Opazili smo, da je naš pristop časovno bolj zahteven kot ostale metode. Uglaševanje se izvaja *offline*, kar pomeni, da se izvede v času učenja in zato časovna komponenta tukaj ni tako zelo pomembna. Samo prevajanje pa se izvaja *online* in je časovno učinkovito.

Za nadaljnje raziskovalno delo na obravnavanem področju bi se lahko osredotočili na posamezne parametre modela za prevajanja, kateri so bolj pomembni in kateri manj in na hitrost prevajanja. Hitrost prevajanja bi pospešili z manjšo tabelo fraz (posledično slabši prevodi) ali pospešitvijo evolucijskega procesa (paralelizacija). Ker je model prevajanja sestavljen iz štirih statističnih modelov, bi lahko vsak model uglaševali posebej.

VIRI

[1] ALPAC. *Languages and Machines: Computers in Translation and Linguistics.*1966.

[2] Bar-Hillel Y. *Automatic Translation of Languages*, 1960. Dostopno na : http://www.mt-archive.info/Bar-Hillel-1960.pdf [14. 5. 2013].

[3] Bošković B. *Uglaševanje šahovske ocenitvene funkcije s pomočjo algoritma diferencialne evolucije.* Doktorska disertacija. Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, 2010.

[4] Brest J., Greiner S., Bošković B., Mernik M., in Žumer V. *Self-Adapting Control Parameters in Differential Evolution: A Comparative Study on Numerical Benchmark Problems.* IEEE Transactions on Evolutionary Computation, vol. 10, no. 6, 2006, str. 646–657.

[5] *Building baseline system.* Dostopno na: http://www.statmt.org/moses/?n=moses.baseline [14. 5. 2013].

[6] Callison-Burch C., Koehn P., Christof Monz, Matt Post, Radu Soricut, Lucia Specia, *Findings of the 2012 Workshop on statistical machine translation*, ACL, 2012, str. 10-51.

[7] Callison-Burch C., Osborne M in Koehn P. *Re-evaluating the Role of BLEU in Machine Translation Research*. 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006, str. 249-256.

[8] *CCVista Translation Database.* Dostopno na: http://ccvista.taiex.be [14. 5. 2013].

[9] Clark J., Dyer C., Lavie A. in Smith N. *Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability.* Proceedings of the Association for Computational Lingustics, 2011.

[10] Crammer K. in Singer Y. *Ultraconservative Online Algorithms for Multiclass Problems*, JMLR 3, str. 951-991, 2003.

[11] Das S. in Suganthan P. *Differential evolution: A survey of the stateof-the-art.* IEEE Transactions on Evolutionary Computation, vol. 15, no. 1, 2011, str. 27–54.

[12] Denkowski M. in Lavie A. *Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems.* Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation, 2011.

[13] Doddington G. *Automatic evaluation of machine translation quality using n-gram cooccurrence statistics*. Proceedings of the Human Language Technology Conference (HLT), 2002, str. 128-132.

[14] *Eur-Lex, dostop do zakonodaje Evropske unije*. Dostopno na: http://eur-lex.europa.eu/sl/index.htm [14. 5. 2013].

[15] *EuroVoc, večjezični tezaver Evropske unije*. Dostopno na: http://eurovoc.europa.eu [14. 5. 2013].

[16] *Evaluation of machine translation.* Dostopno na: http://en.wikipedia.org/wiki/Evaluation_of_machine_translation [14. 5. 2013].

[17] *Evalvacija strojnih prevajalnikov.* Dostopno na: http://sl.wikipedia.org/wiki/Evalvacija_strojnih_prevajalnikov [14. 5. 2013].

[18] *Evropska unija.* Dostopno na: http://europa.eu [14. 5. 2013].

[19] Federico M., Bertoldi N. in Cetollo M. *IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models.* 2008.

[20] *General Text Matcher.* Dostopno na: http://nlp.cs.nyu.edu/GTM [14. 5. 2013].

[21] Hasler E., Haddow B. in Koehn P. *Margin Infused Relaxed Algorithm for Moses.* University of Edinburgh, Institute for Language, Cognition and Computation, 2011.

[22] Heafield K. *KenLM: Faster and Smaller Language Model Queries.* Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation, 2011.

[23] Hutchins J. in Somers H. *An introduction to machine translation.* London: Academic Press - Harcourt Brace Jovanovich, 1992.

[24] Hutchins J. *The History of Machine Translation in a Nutshell.* 2005.

[25] *Jezikovni modeli*. Dostopno na: http://en.wikipedia.org/wiki/Language_model [14. 5. 2013].

[26] *Jezikovni modeli*. Dostopno na: http://homepages.inf.ed.ac.uk/lzhang10/slm.html [14. 5. 2013].

[27] *Joint Research Center* . Dostopno na: http://langtech.jrc.it [14. 5. 2013].

[28] Koehn P., Hoang H. , Birch A., Callison-burch C., Zens R., Aachen R., Constantin A., Federico M., Bertoldi N., Dyer C. , Cowan B., Shen W., Moran C. in Bojar O. *Moses: Open source toolkit for statistical machine translation system.* 2007.

[29] Koehn P. *Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models.* AMTA, 2004.

[30] Koehn P. *Statistical machine translation*, School of Informatics, University of Edinburgh, 2011.

[31] Koehn P., Birch A. in Steinberger R. *462 Machine Translation Systems for Europe*, MT Summit XII, 2009.

[32] Lin C-Y. *ROUGE: A Packagefor Automatic Evaluation of Summaries.* University of Southern California, Information Sciences Institute, 2004.

[33] Mladi za napredek Maribora. *Strojno prevajanje besedila in govora.* Raziskovalna naloga. 2009.

[34] *MultEval*. Dostopno na: https://github.com/jhclark/multeval [14. 5. 2013].

[35] Neri F. in Tirronen V. *Recent advances in differential evolution: a survey and experimental analysis.* Artificial Intelligence Review, vol. 33, no. 1–2, 2010, str. 61–106.

[36] Och F. J. *Minimum Error Rate Training in Statistical Machine Translation.* University of Southern California, Information Sciences Institute, 2003.

[37] Och F. J. in Ney H. *A Systematic Comparison of Various Statistical Alignment Models.* Computational Linguistics, volume 29, number 1. 2003, str. 19-51.

[38] O'Connell T. *Preparing your Web Site for Machine Translation*. 2001.

[39] Papineni K., Roukos S., Ward T. in Zhu W. *BLEU: a Method for Automatic Evaluation of Machine Translation.* Proceedings of the 40th Annual Meeting of the ACL. 2002, str. 311-318.

[40] Peršič L. *Evalvacija dveh strojnih prevajalnikov: Amebis Presis in Google Prevajalnik.* Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za prevajalstvo, 2009.

[41] Piron C. *Le défi des langues (The Language Challenge)*. Paris, L'Harmattan, 1994.

[42] Ralf S., Pouliquen B., Widiger A., Ignat C., Erjavec T., Tufiş D., Varga D. *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages*. Proceedings of the 5thInternational Conference on Language Resources and Evaluation LREC, 2006, str. 2142-2147.

[43] Sapir E. *Language: An introduction to the study of speech.* New York: Hartcourt Brace. 1921, str. 39.

[44] Snover M., Dorr B., Schwartz R., Micciulla L. in Makhoul J. *A Study of Translation Edit Rate with Targeted Human Annotation*. Proceedings of Association for Machine Translation in the Americas, 2006.

[45] Snover M., Madnani N., Dorr B. in Schwartz R. *Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric*. 2009.

[46] Somers H. *Round-Trip Translation: What Is It Good For?* 2005.

[47] *Statistical machine translation*. Dostopno na: http://en.wikipedia.org/wiki/Statistical_machine_translation [14. 5. 2013].

[48] *Statistično strojno prevajanje*. Dostopno na: http://sl.wikipedia.org/wiki/Statistično_strojno_prevajanje [14. 5. 2013].

[49] Stolcke A. *SRILM – An extensible language modeling toolkik.* Proceedings od the 7th International Conference on Spoken Language Processing (ICSLP), 2002.

[50] Storn R. in Price K. *Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces.* Journal of Global Optimization, vol. 11, 1997, str. 341–359.

[51] *Teorija informacij*. Dostopno na: http://sl.wikipedia.org/wiki/Teorija_informacij [25.5.2013].

[52] *Tuning for Quality*. Dostopno na: http://www.statmt.org/moses/?n=Moses.Tutorial#ntoc5 [14. 5. 2013].

[53] Turian J. P., Shen L. in Melamed I. D. *Evaluation of Machine Translation and its Evaluation*. 2003.

[54] Verdonik, D. *Jezikovni viri za strojno simultano prevajanja govora.* Zbornik Društva mladih raziskovalcev Slovenije. Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Center za jezikovne tehnologije, 2005.

[55] Vičič J. *Strojno prevajanje in slovenščina.* Zbornik Sedme konference Jezikovne Tehnologije Institut Jožef Stefan, Ljubljana, 2010.

[56] Vičič, J. *Statistično strojno prevajanje naravnih jezikov.* Magistrska naloga. Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, 2002.

[57] Vintar Š. *Računalniške tehnologije za prevajanje.* Ljubljana: Slovensko društvo Informatika. 1999, str. 17-24.

[58] Vrščaj A. *Evalvacija strojnih prevajalnikov.* Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za prevajalstvo, 2011.

[59] White J. S., O'Connell T. in O'Mara F. *The ARPA MT Evaluations Methodologies: Evolution, Lessons, and Future Approaches*. 1994.

Univerza v Mariboru

Fakulteta za naravoslovje,
matematiko in tehnologijo

IZJAVA O ISTOVETNOSTI TISKANE IN ELEKTRONSKE VERZIJE ZAKLJUČNEGA DELA IN OBJAVI OSEBNIH PODATKOV DIPLOMANTOV

Ime in priimek avtorja-ice: JANI DUGONIK

Vpisna številka: E50004465

Študijski program: RAČUNALNIŠTVO IN INFORMACIJSKE TEHNOLOGIJE

Naslov zaključnega dela: UGLAŠEVANJE PARAMETROV PRI STATISTIČNEM STROJNEM PREVAJANJU

Mentor: red. prof. dr. JANEZ BREST

Somentor: doc. dr. BORKO BOŠKOVIĆ

Podpisani-a JANI DUGONIK izjavljam, da sem za potrebe arhiviranja oddal elektronsko verzijo zaključnega dela v Digitalno knjižnico Univerze v Mariboru. Zaključno delo sem izdelal-a sam-a ob pomoči mentorja. V skladu s 1. odstavkom 21. člena Zakona o avtorskih in sorodnih pravicah dovoljujem, da se zgoraj navedeno zaključno delo objavi na portalu Digitalne knjižnice Univerze v Mariboru.

Tiskana verzija zaključnega dela je istovetna z elektronsko verzijo elektronski verziji, ki sem jo oddal za objavo v Digitalno knjižnico Univerze v Mariboru.

Zaključno delo zaradi zagotavljanja konkurenčne prednosti, varstva industrijske lastnine ali tajnosti podatkov naročnika: _____ ne sme biti javno dostopno do _____(datum odloga javne objave ne sme biti daljši kot 3 leta od zagovora dela).

Podpisani izjavljam, da dovoljujem objavo osebnih podatkov, vezanih na zaključek študija (ime, priimek, leto in kraj rojstva, datum zaključka študija, naslov zaključnega dela), na spletnih straneh in v publikacijah UM.

Datum in kraj: 30.6.2013 Podpis avtorja-ice: Jani Dugonik

Podpis mentorja: _____
(samo v primeru, če delo ne sme biti javno dostopno)

Podpis odgovorne osebe naročnika in žig:
(samo v primeru, če delo ne sme biti javno dostopno) _____

Univerza v Mariboru
Fakulteta za elektrotehniko, računalništvo in informatiko

IZJAVA O AVTORSTVU

Spodaj podpisani/-a JANI DUGONIK

z vpisno številko E5000445

sem avtor/-ica magistrskega dela z naslovom:

UGLAŠEVANJE PARAMETROV PRI STATISTIČNEM STROJNEM PREVAJANJU

(naslov magistrskega dela)

S svojim podpisom zagotavljam, da:

- sem magistrsko delo izdelal/-a samostojno pod mentorstvom (naziv, ime in priimek)

  izr. prof. dr. JANEZ BREST

  in somentorstvom (naziv, ime in priimek)

  doc. dr. BORKO BOŠKOVIĆ

- so elektronska oblika magistrskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko magistrskega dela.
- soglašam z javno objavo elektronske oblike magistrskega dela v DKUM.

V Mariboru, dne 30.6.2013

Podpis avtorja/-ice:

Jani Dugonik

Univerza v Mariboru

Fakulteta za elektrotehniko, računalništvo in informatiko

IZJAVA O USTREZNOSTI MAGISTRSKEGA DELA

Spodaj podpisani/-a izr. prof. dr. JANEZ BREST izjavljam, da je
(ime in priimek mentorja/-ice)

študent JANI DUGONIK izdelal magistrsko
(ime in priimek študenta/-ke)

delo z naslovom: UGLAŠEVANJE PARAMETROV PRI STATISTIČNEM STROJNEM PREVAJANJU

(naslov magistrskega dela)

v skladu z odobreno temo magistrskega dela, Navodili za pisanje magistrskih del na študijskih programih 2. stopnje UM FERI in mojimi navodili.

Kraj in datum: MARIBOR, 30. 6. 2013

Podpis mentorja: