UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Dean Völgyi

Strojno prevajanje s primerjavo besedilnih značilk

DIPLOMSKO DELO
NA UNIVERZITETNEM ŠTUDIJU

Mentor: doc. dr. Zoran Bosnić

Ljubljana, 2010

Univerza
v Ljubljani

Fakulteta za računalništvo
in informatiko

Tržaška 25
1000 Ljubljana, Slovenija
telefon: 01 476 84 11
faks: 01 426 46 47
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Št. naloge: 01676/2010
Datum: 14.05.2010

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kandidat: **DEAN VOELGYI**

Naslov: **STROJNO PREVAJANJE S PRIMERJAVO BESEDILNIH ZNAČILK**
MACHINE TRANSLATION BASED ON COMPARISON OF TEXT FEATURES

Vrsta naloge: Diplomsko delo univerzitetnega študija

Tematika naloge:

Strojno prevajanje med naravnimi jeziki je aktualno področje v razvoju, s katerim se spopadajo različna podpodročja računalništva. Metode za strojno prevajanje upoštevajo zakonitosti analiziranih besedil in si pomagajo z orodji kot so slovarji, metode za rudarjenje po besedilih ter z drugimi analitičnimi in sintetičnimi tehnikami.

V okviru diplomske naloge naj kandidat razišče možnost implementacije strojnega prevajanja, ki je implementirana brez prevajalnih pripomočkov (dodatnih orodij, slovarjev ipd.). Izbere naj si par besedil in implementira takšen prevajalni sistem, ki zaznava besede z enakim pomenom glede na njihovo pogostost pojavljanja, kontekst in druge statistike, ki jih definira sam. Kakovost dobljenih prevodov naj analizira in v zaključku tudi kritično ovrednoti.

Mentor:

doc. dr. Zoran Bosnić

Dekan:

prof. dr. Franc Solina

IZJAVA O AVTORSTVU
diplomskega dela

Spodaj podpisani Dean Völgyi,
z vpisno številko 63010167,

sem avtor diplomskega dela z naslovom:
Strojno prevajanje s primerjavo besedilnih značilk

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal/-a samostojno pod mentorstvom doc. dr. Zorana Bosnića,

- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela,

- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki "Dela FRI".

V Ljubljani, dne 22.05.2010 Podpis:

Zahvala

Rad bi se zahvalil vsem, ki so mi pri pisanju diplome pomagali na kakršen koli način.

Posebna zahvala gre mentorju doc. dr. Zoranu Bosniću, ki mi je ob pisanju naloge bil v veliko pomoč.

Diplomo posvečam svoji družini, ki mi je tekom celotnega študija stala ob strani.

Kazalo

Kazalo slik

Kazalo tabel

Povzetek

Cilj naloge je bil preveriti, do katere mere je možno iskati prevode med dvema naravnima jezikoma izključno na podlagi lastnosti naravnih jezikov, torej brez uporabe slovarja. K lastnostim štejemo absolutne in relativne frekvence besed, njihov kontekst, pa tudi frekvence bi- in trigramov v okviru teme podanega besedila. Za ta namen smo razvili prevajalni sistem, v katerem smo implementirali različne pristope k prevajanju. Uspešnost aplikacije smo ovrednotili s slovensko – angleškim parom besedil. Ugotovili smo, da je prevajanje na tak način do določene mere možno, vendar pa rezultati puščajo še veliko možnosti za izboljšave, preden bi se lahko tak pristop uporabljal v resne namene.

Ključne besede:

strojno prevajanje, frekvenca, naravni jezik

Abstract

The goal of the thesis was to study the feasability of translating between two natural languages exclusively on the characteristics of those languages, i.e. without using a dictionary. The term 'characteristics' as used in the study signifies the absolute and the relative frequencies of words, the context where these words are found in, as well as the frequency of bi-grams and tri-grams within the subject of a given text. For that purpose we have developed a translation system, into which we have implemened different approaches to translation. The success of the application was evaluated by a Slovenian-English pair of texts. We established that such way of translating is possible to a certain level, however, the results still leave much room for improvement so that this approach becomes useful for more serious purposes, as well.

Keywords:

machine translation, frequency, natural language

Seznam uporabljenih kratic in simbolov

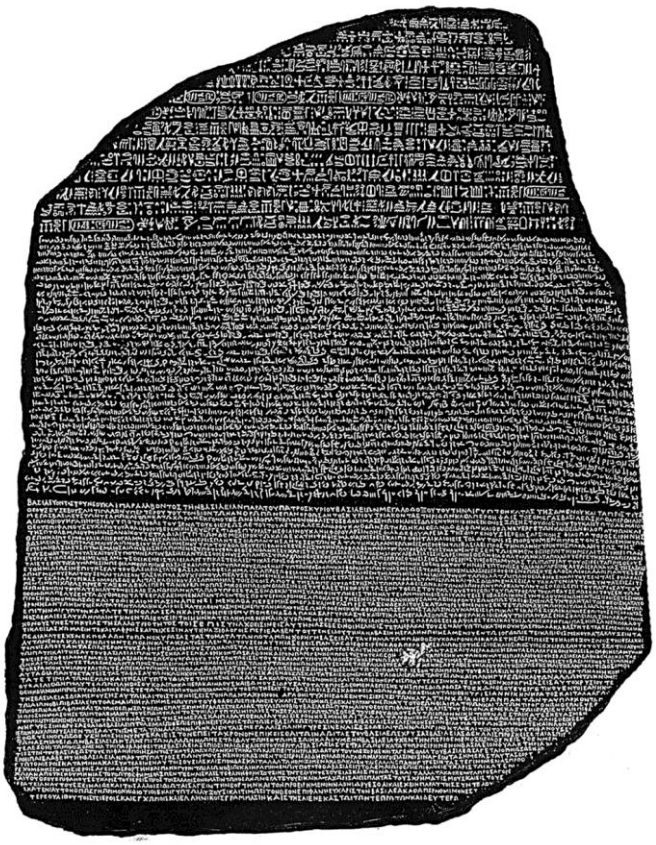
MT	machine translation
SMT	statistical machine translation
KBMT	knowledge-based MT
RBM	rule-based models
EBMT	example based MT

1 Uvod

Procesu, kjer ljudje iz enega sporočila v izvirnem naravnem jeziku tvorijo novo, ekvivalentno sporočilo v drugem naravnem jeziku, pravimo prevajanje. Pri tem procesu v naših možganih nastajajo miselni vzorci, ki pa za razumevanje in razlago niso povsem enostavni [9]. Že razumevanje samega naravnega jezika ostaja za jezikoslovce veliko vprašanje. V povezavi s tem je težko razumeti in razložiti tudi pojem prevajanja. Znanstveniki pogosto poudarjajo, da je za uspešno prevajanje s stroji najprej potrebno rešiti problem razumevanja naravnega jezika. Kako naj potem stroji, ki nimajo duše, iz naravnega jezika razberejo pomen besedil ali celo prevajajo? S tem izzivom se jezikoslovci in računalnikarji ukvarjajo že desetletja. Njihova naloga nikakor ni enostavna. Kljub temu jim je tekom zgodovine uspelo razviti metode in sisteme tako, da si tudi stroji na nek način znajo »razlagati« pomen besedil in s tem ta besedila tudi prevajati.

Izraz strojno ali avtomatsko prevajanje (angl. *machine translation*) pomeni, da proces prevajanja besedila iz enega naravnega jezika v drugega opravlja stroj oz. računalnik [3]. Pri tem procesu si pomaga z raznoraznimi tehnikami: slovarji, analizo in sintezo besedil na različnih nivojih itn. Vse to pa za brezhibne prevode še vedno ni dovolj. Stroji pri svojem delu v določeni meri še vedno potrebujejo človekovo asistenco. Kljub nepopolnosti strojnega prevajanja, pa je stroj danes že v marsikateri disciplini nadomestil človeka, predvsem tam, kjer so tudi bolj grobi prevodi že zadostni. Poklicni prevajalci si danes pomagajo z raznoraznimi avtomatizirani orodji, ki jim lajšajo delo. Po drugi strani pa je zaradi tehnološkega napredka in zaradi vse tesnejše gospodarske in politične združitve sveta potreba po prevajanju v današnjem svetu vse večja. Tako je strojno prevajanje danes za človeka postala pomembna veja na raznih področjih.

Cilj diplomske naloge je pregledati, ali je možno narediti strojni prevajalnik, ki v svojemu delovanju ne bi uporabljal že uveljavljenih metod strojnega prevajanja, temveč novi pristop, ki izkorišča lastnosti naravnega jezika. Ena od lastnosti, ki so tu mišljene, je pogostost pojavljanja besed v besedilu. Če v nekem besedilu, na primer, govorimo o strojnem prevajanju, lahko pričakujemo, da se bosta besedi *strojno* in *prevajanje* pojavili z visoko ali celo najvišjo frekvenco. In na ta način lahko vsaki besedi v podanemu besedilu pripišemo neko pričakovano frekvenco. Ravno ta pojav oz. lastnost smo v nalogi izkoristili kot osnovo za iskanje prevodov. Pričakujemo, da bi takšna ideja lahko delovala, saj je s stališča podajanja informacije popolnoma vseeno, kako, s čim ali v katerem jeziku bomo sporočilo posredovali, pomen sporočila oz. besedila bo ostal enak. Na takšen način prevajanja lahko gledamo tudi kot na prevajanje preko dekodiranja sporočil. Stvar si torej lahko predstavljamo kot iskanje stavka, besede ali znakov s pomočjo dekodiranja šifriranega sporočila. S podobnimi nalogami so se ljudje srečevali že zdavnaj, npr. II. sv. vojna, cesarjev kod, ali sporočilo zapisano na kamnu »Rosetta Stone«. Rosetta Stone, ki ga prikazuje slika 1.1., je kamen, na katerem so Egipčani zapustili zapis besedila v treh pisavah: dveh egipčanskih (hieroglifska in demotska pisava) in v klasični grški pisavi. Kamen je pomemben, saj so si znanstveniki s prevajanjem med temi tremi zapisi pomagali dešifrirati egipčansko pisavo – hieroglife.



Slika 1.1 Rosetta Stone

Cilj praktičnega dela naloge je bil realizacija aplikacije, ki bo na podlagi značilk naravnega jezika znala prevajati besede iz enega naravnega jezika v drugega. Program naj bi znal prevajati iz/v nekatere pomembnejše svetovne jezike (angleščino, nemščino, italijanščino, španščino in francoščino) in slovenščino. Pri tem procesu si bo v glavnem pomagal s prej omenjeno lastnostjo jezikov, tj. frekvenco besed. Poleg tega načina si za iskanje pravih prevodov pomaga tudi s kontekstom ter s posebnostmi v besedah. Drugi cilj praktičnega dela je realizacija metode, ki bo znala avtomatično prevesti celotno besedilo iz enega jezika v drugega.

V 2. poglavju bomo na kratko predstavili klasične pristope strojenga prevajanja, medtem ko bomo v 3. poglavju predstavili naše ideje za pristop k prevajanju ter podrobneje opisali vsako od teh idej. Poglavje 4 govori o implementaciji praktičnega dela sistema za prevajanje. Da bi dobil vtis o tem, kako dobro aplikacija sploh deluje, smo jo po dokončanju tudi testirali. Rezultati, ki smo jih dobili tukaj, so predstavljeni v 5. poglavju.

2 Pristopi k avtomatskemu prevajanju med jezikoma

2.1 Prevajanje

Prevajanje je proces, kjer iz besedila originalnega jezika tvorimo ekvivalentno besedilo v ciljnem jeziku, torej tvorimo sporočilo, ki ima enak pomen kot izvirno sporočilo. Prevajanje nikakor ni enostaven ter enoličen proces, temveč je eden od najkompleksnejših postopkov, ki jih je človeštvo v svoji evoluciji razvilo [9]. Za kvalitetne prevode so v ozadju procesa potrebne kompleksne operacije. Na splošno lahko povemo, da je prevajanje proces, sestavljen iz dveh delov: 1) dekodiranje pomena izvirnega besedila in 2) re-kodiranje tega pomena v ciljnem jeziku [11]. Prevajalec mora originalno besedilo razčleniti in analizirati; zato mora poznati slovnico, semantiko, sintakso, idiome, itn., pa tudi okoliščine in kulturo naravnega jezika iz in v katerega prevaja. V vlogi prevajalca vsak posameznik prevaja na svojevrsten način; nekateri prevajajo sproti besedo za besedo, spet drugi dajejo poudarek prevajanju po pomenu itd. Na prevajanje lahko gledamo celo kot na umetnost; na vsakem koraku prevajalnega procesa smo prepuščeni lastni izbiri. Dva človeka istega besedila iz izvirnega jezika nikoli ne bosta prevedla v isto besedilo v ciljnem jeziku. Prav zaradi kompleksnosti jezika in razumevanja miselnih procesov je proces prevajanja težko enostavno opisljiv. To pa velja še posebej v svetu računalnikov. V začetkih je bilo prevajanje izključno delo učenjakov. Z rojstvom računalnikov pa se je spremenilo tudi prevajanje kot delo; prevajalci imajo danes povsem drugačno vlogo/delo, kot so ga imeli nekoč. Profesionalni prevajalci si današnjega dela brez računalniških orodij več niti ne morejo predstavljati. Orodja kot npr. elektronski slovarji, črkovalniki ipd. jim omogočajo hitrejše in bolj kakovostno delo. Lahko rečemo, da postaja meja med strojnim in človeškim prevajanjem zabrisana.

2.2 Strojno prevajanje

O strojnem prevajanju (angl. *machine translation*) govorimo, kadar je prevajanje iz enega naravnega jezika v drugega narejeno s pomočjo analize stroja oz. računalniške programske opreme. Za osnovno strojno prevajanje lahko rečemo, da računalnik za vsako besedo iz izvirnega naravnega jezika skuša najti primerno besedo v ciljnem naravnem jeziku – gre torej le za substitucijo besed. Strojno prevajanje pa je lahko tudi izboljšano, če pri prevajanju upoštevamo razlike v lingvističnih topologijah, vključimo zbirke besedil, lingvistična pravila, upoštevamo prepoznavanje fraz, idiomov, izolacijo anomalij itn. Prevajanje deluje bolje tudi, če uporabnik računalniku asistira z raznoraznimi podatki; npr. vnos mašil, imen ipd.

Prevajanje besedila predstavlja torej stroju kompleksen proces, saj naravni jezik ni le »prazno« zaporedje besed. V njem se mora upoštevati kontekst, besedni vrstni red, časi, večpomenskost itn., to pa je stroju težje razumljivo kot človeku. Prevajalci imajo namreč, za razliko od strojev, znanje in razumevanje o svetu okoli njega. Če želimo, da bo strojno prevajanje čim bolj učinkovito, mora stroj razpolagati z eno- in večjezičnimi leksikoni, programi za morfološko in sintaktično analizo in sintezo, razreševanje večpomenskosti,

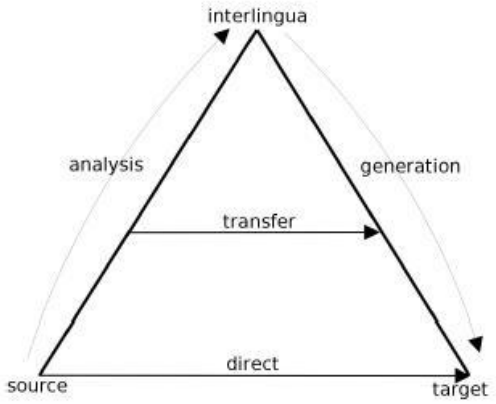
prepoznavanje večbesednih semantičnih enot ter ostale kompleksne mehanizme, ki pripomorejo k čim boljšemu avtomatskemu prevajanju s čim manj napakami [9].
Pri strojnem prevajanju je proces prevajanja torej avtomatiziran do največje možne mere. To omogoča boljše, hitrejše in cenejše prevajanje [3]. Vendar pa avtomatizacija ni popolna; prevajanje še vedno zahteva vmešavanje človeka, vsaj do določene mere. To velja predvsem za predpripravo izvirnega besedila (pre-editing) in popravek rezultatov (post-editing) [6]. Pri popravku rezultatov mislimo predvsem na popravljanje členov, predlogov, zaimkov, glagolskih časov itn. Naknadno neurejena besedila so večinoma razumljiva, a včasih tudi napačna, zavajajoča ter človeku težko berljiva [1]. Jezikoslovci menijo, da strojno prevajanje nikoli ne bo moglo popolnoma nadomestiti zmožnosti človeškega prevajalca, interpretacije finih nians, kulturnih referenc ter uporabo slengov in idiomov [10]. Čim večja je razlika med jezikoma, njuno strukturo in samo kulturo, tem večja je tudi težavnost dobrega prevajanja. Kljub temu se strojno prevajanje dandanes uporablja že v mnogih disciplinah. Tukaj ne mislimo na literarne tekste in poezijo, ampak na besedila kot so tehnični priročniki, propagandni prospekti, administrativni memorandumi, zdravniška poročila [6], tehnični dokumenti, seznam stvari (zaloge, skladišča, inventar) - skratka gre za prevode, pri katerih je tudi bolj surovo besedilo dovolj [10]. Strojno prevajanje je primerno predvsem za specifične discipline, kot so npr. vreme, šport, politika ipd. - tako je namreč v ciljnem jeziku omejeno število dovoljenih možnih substitucij.

Glavna prednost strojnega prevajanja je količina prevedenega besedila in hitrost prevajanja, vsekakor pa kvaliteta strojnega prevoda zaostaja za prevodom, ki ga zna naredi človek sam [9]. Sodobna prevajalna arhitektura se še vedno sooča z istimi morfološkimi, leksikološkimi in strukturnimi težavami, kot se je ob samem rojstvu discipline. Glavna razlika med nekoč in danes je v pristopu. Nekoč so se posluževali direktnih tehnik, medtem ko se danes raje poslužujemo bolj uspešnih interaktivnih in statističnih metod oz. metod prevajanja na podlagi korpusov (besedilnih zbirk).

Zaradi vse močnejšega vpliva globalizacije narašča tudi nuja po prevajanju iz in v različne jezike [9]. Če želi biti moderno podjetje uspešno, mora za svoj obstoj biti kompetentno; komunicirati mora s strankami in kulturami celega sveta. Eden od načinov za dosego teh ciljev je uporaba različne programske opreme za sodoben način prevajanja med različnimi jeziki sveta. Strojno prevajanje poveča efektivnost dela (npr. tako, da produkti hitreje pridejo na trg) in zmanjša stroške prevajanja. Druga velika skupina uporabnikov strojnega prevajanja pa so uporabniki interneta. Dandanes je na svetovnem spletu na voljo ogromno prosto dostopnih prevajalnikov, ki jih različni ljudje uporabljajo za različne namene.

2.3 Pregled obstoječih pristopov

Skozi zgodovino se je razvilo kar nekaj različnih pristopov avtomatiziranega prevajanja. Ti se razlikujejo predvsem po analizi vhodnega besedila in po kvaliteti prevodov.



Slika 2.1: Trikotnik prikazuje globino vmesne reprezentacije [11]

Kot samih pristopov je tudi kriterijev za klasificiranje pristopov mnogo [11]. Najbolj popularno razlikovanje je razlikovanje po globini analize besedila [4]. To lahko grafično izrazimo s pomočjo ti. prevajalne piramide - slika 2.1. Slika nam pove, da sistemi, ki so na dnu piramide, za iskanje prevodov ne izvajajo nobene lingvistične analize vhodnega besedila – so direktni. Na sredini višine piramide najdemo sisteme, ki za prevajanje izvajajo že določene analize; predvsem morfološkega in sintaktičnega značaja. Na vrhu piramide pa so sistemi z najglobljo analizo, zmožni so tudi semantične obdelave originalnega besedila. V 80-ih so znanstveniki predstavili številne nove pristope, med katerimi so največ uspeha poželi predvsem na primerih temelječ pristop, prevajanje na osnovi besedilnih zbirk (korpusov) in hibridna metoda [10]. Vsak pristop pa je v določeni meri lahko tudi mešanica s človeško asistenco.

2.3.1 Na znanju temelječ pristop

Na znanju temelječ pristop (angl. *knowledge-based MT*) prevaja po lingvističnih pravilih, ki jih stroju določijo ljudje. Sistemi KBMT imajo torej vgrajena pravila oz. neko »znanje« o zunanjem svetu [10]. Metoda, poleg velikega števila pravil za delovanje, potrebuje tudi veliko število modulov za izvajanje vseh tipov lingvističnih analiz ter reprezentacijo in generiranje form ter pomena. Ta pristop je zaradi velikega števila pravil drag in neizplačljiv, zato v realnem svetu ni preveč razširjen.

2.3.2 Na jezikovnih pravilih temelječ pristop

Na jezikovnih pravilih temelječi pristopi (angl. *rule-based models*) so bogati v »znanju«: vgrajena imajo lingvistična pravila (sintaktične, in semantične informacije) ter slovarje (splošne in kontekstualne) za vsak par jezikov posebej [14]. Pravila morajo biti implementirana v obliki, v kateri jih tudi stroji razumejo. Čim večjo bazo znanja bo sistem imel, bolj kvalitetne prevode bo generiral. Proces RBM-a bi lahko razčlenili na 3 osnovne korake [14]:

1) besedilo se razčleni oziroma slovnično opredeli,

2) iz razčlenjenega besedila dobimo vmesno simbolično reprezentacijo besedila,

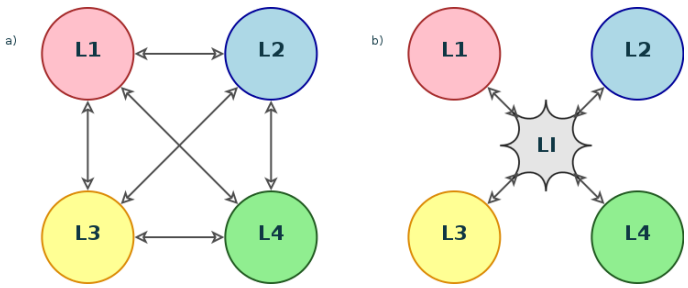
3) iz vmesne reprezentacije dobimo končno, prevedeno besedilo.

Splošna težava teh sistemov je robustnost: prevod brez slovnice ali slovarja je ponavadi slab. Slaba stran metode so tudi cena prevoda in izjeme v jeziku - ker metoda bazira na pravilih, se težave pojavljajo ravno pri izjemah pravil.

Glede na način delovanja lahko RBM razčlenimo na sledeče tri principe: direktni pristop, pristop z vmesnim jezikom in transferni pristop [11]. Oglejmo si jih podrobneje:

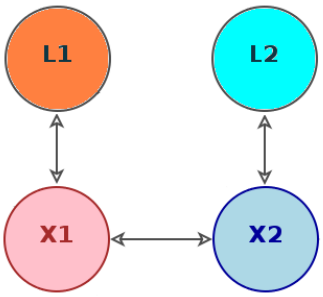
- **Direktni pristop:** pri direktnem pristopu (angl. *direct translation*) je prevajanje direktno; prevajanje ponavadi bazira le na prevodu slovarjev in pravil za prevajanje [6]. Izvirno besedilo, za razliko od drugih pristopov, ponavadi ni analizirano. Primeren je za prevajanje seznamov.

- **Pristop z vmesnim jezikom:** pristop z vmesnim jezikom (angl. *interlingua-based*) velja za enega klasičnih pristopov. Vmesni jezik naj bi bil abstraktna in jezikovno neodvisna predstavitev izvirnega besedila. Proces prevajanja poteka tako, da se besedilo iz izvirnega jezika najprej prevede v vmesni jezik, od tu pa se tvori prevedeno besedilo v ciljnem jeziku [6]. Ker je vmesni jezik neodvisen od ostalih jezikov, se besedilo iz vmesne reprezentacije (glej sliko 2.2 - desno) lahko hkrati prevede v kateri koli ciljni jezik. Prednost pristopa je ekonomičnost; dvojezični korpusi, pravila ipd. niso potrebni za vsako dvojico jezikov. Pri n različnih jezikih namesto $n*(n-1)$ dvojic potrebujemo le $2n$ dvojic prevodov [4, 6]. Za primer lahko na sliki 2.2 vidimo, da pri prevajanju med 4 različnimi jeziki z direktnim prevajanjem potrebujemo 12 slovarjev, medtem ko z metodo z vmesnim jezikom le 8.



Slika 2.2: Razlika: direktno prevajanje (levo-a) in prevajanje z vmesnim jezikom (desno-b)

- **Transferni pristop:** ideja pri transfernem pristopu (angl. *transfer-based MT*) je podobna kot pri metodi interlingue, s to razliko, da imamo tukaj namesto ene vmesne stopnje reprezentacije dve vmesni stopnji reprezentacije, in sicer eno za vsak prevajani jezik [6] – glej sliko 2.3. Prevajanje izvedemo med vmesnima reprezentacijama obeh jezikov.



Slika 2.3: Transferni pristop, L1 in L2 predstavljata originalni in ciljni jezik, X1 in X2 pa vmesni stopnji reprezentacije

2.3.3 Statistični pristop

SMT prevaja na podlagi statističnih metod. Ideja za statistično metodo (angl. *statistical machine translation*) izhaja iz informacijske teorije, ki govori o verjetnostni porazdelitvi [17, 6]. Prve ideje so bile predstavljene že leta 1949. Za razliko od RBM- a SMT nima vgrajenega »znanja« (slovar, slovnična pravila ipd.), namesto tega za vsak jezik vsebuje zbirko besedil oz. korpus tekstov. Metoda se na podlagi statistike odloči, če je nek prevod pravilen ali ne. Prve raziskave na tem področju so bile izvedene konec 80-ih z IBM-ovim projektom Candide [7]. Pred drugimi metodami ima ogromno prednosti, zato je dandanes v uporabi zelo razširjena. Njeni glavni prednosti sta [17]:

1) boljša izraba virov
    a. obstaja veliko naravnih jezikov, ki jih stroji »razumejo«,
    b. SMT sistemi niso specializirani za nobeno določeno dvojico jezikov,
    c. za razliko od RBM SMT ne zahteva ogromno vloženega dela znanstvenikov
2) prevodi so človeku bolj »naravni«

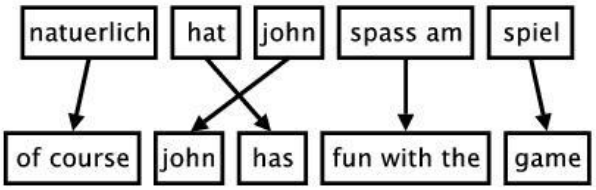
Seveda pa ima tudi ta pristop svoje pomanjkljivosti. Ena od glavnih je ta, da v besedilih nikoli ni perfektnega paralelnega nastopa besed; soočiti se mora torej z [17]:

1) idiomi (posebnostmi),
2) sestavljenimi besedami,
3) morfologijo,
4) vrstnim redom besed (ena od možnih rešitev za to je, da v vsakem stavku najdemo glagole, samostalnike, pridevnike itn., nato pa jih razvrstimo po pravilu, kot zahteva ciljni jezik),
5) sintakso,
6) ti. »besede izven slovarja«- če določene besede ni v učni množici SMT-ja, potem se ta beseda v ciljni jezik ne zna prevesti.

Kot omenjeno, je pionir na tem področju IBM [7]. Pristop, ki so ga uporabili, je bil zamenjava besede za besedo z možnostjo izbrisa ali dodajanja določene besede. Pravzaprav so na začetku vsi statistični poskusi bili tipa beseda za besedo. Kasneje so raziskovalci z izboljšavami predstavili prevajanje fraz (angl. *phrase-based trsanslations*), katerih enota ni več bila beseda, ampak fraza. Ta metoda se je obnesla bolj uspešno od predhodnega prevajanja besede za besedo. Podobno so kasneje prišli s prevajanjem na osnovi sintakse (angl. s*yntax-based translation*) do prevajanja na še višjem nivoju sintakse.

- **Pristop na osnovi besed**: glavna značilnost pristopa na metodi besed (angl. w*ord-based translation*) je, da je enota, ki jo prevajamo, beseda sama [17]. Zaradi idiomov, sestavljenk in morfologij se ponavadi število besed v ciljnem jeziku razlikuje od tistega v izvirnem jeziku.
- **Pristop na osnovi fraz:** enota, ki se pri pristopu na osnovi fraz (angl. *phrase-based trsanslations*) prevaja, ni le ena beseda, temveč celotno zaporedje besed [17], im. bloki. Postopek delovanja metode lahko opišemo na sledeč način (glej sliko 2.4) [7]:

    - vhodno besedilo razdelimo na fraze (zaporedja besed),
    - vsaka vhodna fraza se prevede v frazo v ciljnem jeziku,
    - razvrščanje besednega vrstnega reda v izhodnem jeziku.



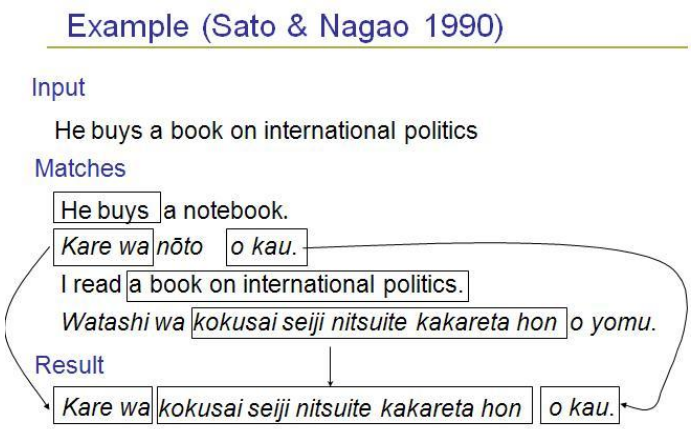
Slika 2.4: Postopek delovanja pristopa na osnovi fraz [7]

- **Pristop na osnovi sintakse:** pristop na osnovi sintakse (angl. *syntax-based translation*) deluje na ravni sintakse. Vhodni stavki se s parserjem sintaktično obdelajo [17]. Celoten proces ponavadi zajema tudi menjavo vrstnega reda besed ter vstavljanje nekaterih dodatnih besed [12].

2.3.4 Na primerih temelječ pristop

Ideja na primerih temelječega pristopa (angl. *example based MT*) deluje na osnovi analognosti besedil [13]. Osnova za prevajanje je dvojezični korpus (baza podatkov, ki vsebuje prej videne primere parov prevodov), v katerem najdemo paralelno zbirko besedil za originalni in ciljni jezik [18]. Prevajanje poteka tako, da se za vhodno besedilo iz korpusa poišče ujemanje. Proces prevajanja lahko razdelimo na tri korake (glej sliko 2.5):

1) iskanje ujemanja med vhodnim besedilom in primeri prevodov iz DB,
2) poravnava: identifikacija ustreznega fragmenta prevoda in
3) zlaganje fragmentov besedila v ciljnem jeziku.



Slika 2.5: Potek prevajanja EBMT pristopa

EBMT je relativno nova metoda - predlagana leta 1984. Spada v družino pristopov, ki sklepajo na podlagi posameznih primerov (angl. *case-based reasoning*). Tudi tukaj velja pravilo: čim obsežnejši je korpus, tem boljša bo kakovost prevodov. Prednost EBMT-ja pred RBM-om je hitrejši razvoj sistema, saj EBMT ne potrebuje vgrajevanja velikega števila pravil. Ima pa tudi veliko slabih strani. Težava nastopi pri manj prevajanih parih jezikov ali še neraziskanem področju, kjer pogosto ni dovolj razpoložljivih vzporednih korpusov. EBMT ima težavo tudi, kadar se prevaja med jezikoma, katerih strukture se zelo razlikujejo; denimo angleško – japonska dvojica. Ker ima pristop več slabosti kot prednosti, se ga uporablja manj kot SMT ali RBM.

2.3.5 Hibridni pristop

Pri hibridnem pristopu (angl. *hybrid MT*) se komplementarno dopolnjujejo prednosti statistične metode in metode, ki temelji na jezikovnih pravilih [11]. V grobem lahko povemo, da je proces sestavljen iz dveh delov [10]: pri prevajanju se besedilo najprej obdela s statističnimi metodami, nato pa še z na jezikovnih pravilih temelječim pristopom (lahko pa tudi z obratnim vrstnim redom).

2.3.6 Kontekstni pristop

Kot že samo ime pove, bazira prevajanje kontekstnega pristopa (angl. *context-based MT*) na kontekstu [8]. Le-ta se išče s pomočjo prekrivanja (dolgih) n-gramov. Glavna prednost te metode pred klasičnimi je, da klasični pristopi za strojno prevajanje potrebujejo bodisi zamudno gradnjo pravil (kot npr. RBM), bodisi zamudno gradnjo velikega paralelnega korpusa (kot npr. statistični pristop). Ta pristop pa za delovanje potrebuje le enojezični korpus in dvojezični slovar. Sicer pa metoda ni preveč razširjena.

2.4 Zgodovina/razvoj

Ideje in razvoj strojnega prevajanja so se začele prej kot si marsikdo misli. Tekom zgodovine so strojno prevajanje doleteli vzponi in padci. Venomer so se pojavljale nove ideje in pristopi. Zgodovino strojnega prevajanja lahko razdelimo na zaokrožena obdobja [6,2].

Začetne ideje (17. stol. – 1933): prve ideje o strojnem prevajanju so se pojavile že davno pred razvojem računalnikov, vendar v tistem obdobju še ni bilo tehnologije, ki bi realizacijo ideje tudi omogočila. Ideje so postale realnost šele v 20. stol. Na področju teh idej sta bila dejavna predvsem filozofa René Descartes in Gottfried Leibniz.

Začetni poskusi, predhodniki in pionirji (1933 – 1954): rojstvo strojnega prevajanja ne zaostaja veliko za rojstvom prvih računalnikov. Prvi poskusi so bili vojaške narave. Leta 1954 bil izveden eden od prvih in najznamenitejših poskusov, Georgetownov eksperiment (prvi popolnoma avtomatiziran prevod, kjer so prevedli več kot 60 stavkov iz ruščine v angleščino). V tem obdobju so še bili prepričani, da bo proces prevajanja lahko še v bližnji prihodnosti postal popolnoma avtomatiziran ter da se bodo avtomatsko lahko prevajala besedila vseh vrst oz. disciplin.

Obdobje optimizma 1954 – 1966: v tem obdobju so mislili, da se bo strojno prevajanje zaradi napredka tehnologije in analiz še dodatno izboljšalo. Raziskovalci pa so hitro naleteli na »semantične meje«. Leta 1964 so zaradi slabih uspehov strojnega prevajanja ustanovili komisijo ALPAC (Automatic Language Processing Advisory Committee), ki je 2 leti kasneje ugotovila, da je strojno prevajanje počasno, manj natančno in dvakrat dražje od človeškega prevajanja ter da za uporabo strojnega prevajanja ni predvidenega upanja.

Posledica poročila ALPAC, 1966 – 1980: ALPAC je ustavila razvoj strojnega prevajanja, predvsem v ZDA. Razvoj v Evropi pa se je kljub temu počasi nadaljeval. Prvi večji uspeh so po dolgem času dosegli šele s Systranom leta 1970.

Osemdeseta leta: v tem desetletju je projektov na področju strojnega prevajanja bilo vse več. Razvoj se je širil v vse smeri. Strojno prevajanje je začelo uporabljati vse več in več podjetij.

Devetdeseta leta: velike korporacije so povečale uporabo strojnega prevajanja in prevajalnih orodij. Povečala se je prodaja programske opreme za strojno prevajanje na osebnem računalniku. Še bolj pa se je razširila uporaba strojnega prevajanja preko svetovnega spleta (npr. AltaVista). V tem obdobju so se začele tudi raziskave na področju strojnega prevajanja iz zvočnega zapisa oz. govora.

2.5 Primeri strojnega prevajanja

Programske opreme za strojno prevajanje je danes veliko. Večina današnjih sistemov omogoča tudi že prevajanje preko spleta. Prepoznavnejša novejša orodja so [11]:

- BABEL FISH podjetja Yahoo, poganja ga sistem SYSTRAN, razvilo pa ga je podjetje Altavista. Je najbolj znan spletni prevajalnik, dostopen na naslovu www.babelfish.yahoo.com,
- Asia Online, prevajanje preko spleta,

- AppTek, na tržišču leta 2009, prevajanje preko spleta,
- Worldlingo, partner v MS Windows ter MS Mac Office,
- Babilon,
- StarDict,
- Google Language Tools, statistični sistem.

Omembe so vredni tudi starejši sistemi, ki pa dandanes seveda več niso v uporabi:

- kanadski EUROPARL, prevajanje med francoščino in angleščino,
- CANDIDE podjetja IBM, prvi statistični sistem,
- METEO, razvili so ga na univerzi Montreal University v 80. letih. Namenjen je bil prevajanju med angleščino in francoščino pri napovedovanju vremena,
- EUROTRA, prevajanje med vsemi jeziki nekdanje Evropske gospodarske zbornice.

Posebno pozornost moramo nameniti slavnemu sistemu SYSTRAN [19]. Značilno zanj je, da ima veliko bazo besed in omogoča prevajanje med številnimi svetovnimi jeziki. Google-ov SYSTRAN je osnova mnogim današnjim sistemom, npr. sistemu *Google language tools*, *produktom Yahoo*, sistemu *Babelfish* podjetja Altavista pa tudi namiznemu prevajalniku sistema Mac. Kljub temu, da je eden od prvih sistemov, velja dandanes še vedno za enega od vodilnih. Prva verzija je bila narejena v zgodnjih 70. letih. Ker so njeni začetki vezani na hladno vojno, se je v njej prevajalo predvsem med ruskim in angleškim jezikom, in sicer za Ameški US Air Force.

V naslednjem poglavju bomo predstavili naše ideje oziroma pristope za prevajanje, v 4. poglavju pa kako smo omenjene ideje implementirali v naš sistem za prevajanje.

3 Idejna zasnova sistema za strojno prevajanje

Kot smo lahko videli v prejšnjih poglavjih, je načinov za pristop k prevajanju kar nekaj. Vsaka metoda ima svojo logiko, ki stoji za njenim procesom prevajanja. Logika, ki smo jo uporabili v naši raziskavi, pa se tudi razlikuje od vseh do sedaj opisanih. Glavna razlika med klasičnimi metodami in našo je ta, da je naš cilj narediti prevode izključno na podlagi značilk naravnega jezika, torej brez pomoči slovarja ali kakršnega koli drugega orodja. Logika prevajanja naj bi delovala tako, da se za prevajano besedo iz množice vseh besed ciljnega besedila poišče najprimernejša beseda in se jo predlaga za prevod. Celotna »baza podatkov«, ki jo pri iskanju prevodov uporabljamo, pa sta le besedili, med katerima se prevaja.

Vidimo lahko, kako pomembno je, da sta si besedili čim bolj podobni. Če v ciljnem besedilu za želeno besedo ni primernega ekvivalenta, se beseda niti ne bo mogla ustrezno prevesti. V tem primeru pa se še vedno lahko zgodi, da bomo za prevod našli besedo, ki ima podobno vlogo oz. pomen. Druga pomembna razlika je ta, da z našim praktičnim poskusom ne bomo mogli prevesti poljubne besede. Izbiramo lahko samo med tistimi, ki nastopajo v besedilu. Značilnost pristopa je še ta, da se naša logika prevajanja ne bo ukvarjala z razvrščanjem besed, sklanjanjem, spreganjem, idiomi, večpomenskostjo ipd., edina naloga, ki smo si jo zadali, je najti najprimernejši prevod za izbrano besedo. Aplikacija zmore prevesti tudi celotno besedilo, ampak to prevajanje poteka po principu beseda za besedo, torej brez dodatne logike oblikovanja stavkov ali besedila.

V okviru našega pristopa smo implementirali naslednje pristope:

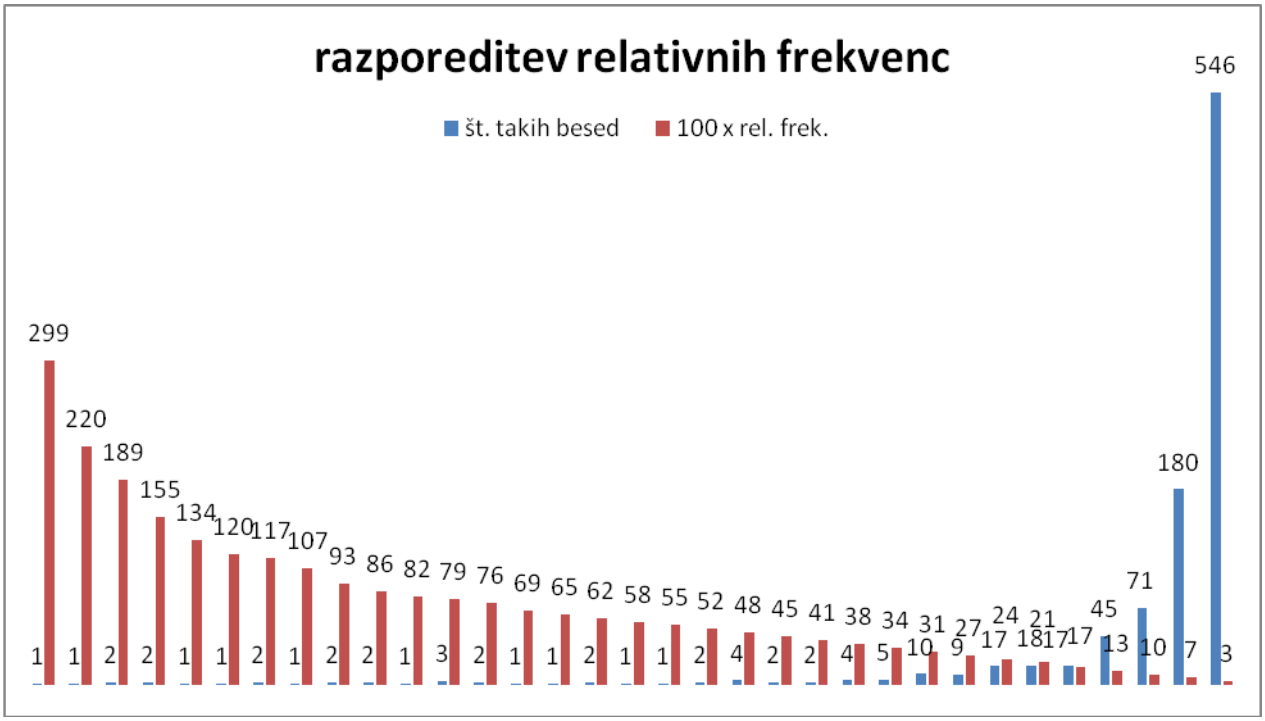
- osnova našemu pristopu predstavljajo frekvence besed. Pri iskanju možnih kandidatov bomo iz besedila v ciljnem jeziku za prevod predlagal tisto besedo, ki ima najbližjo **relativno frekvenco** prevajani besedi – glej razdelek 3.1.1,
- kot bomo videli v nadaljevanju, nam namesto relativnih frekvenc včasih pridejo prav tudi **rangi frekvenc –** poglavje 3.1.2,
- poleg relativnih frekvenc in rangov posameznih besed smo pod drobnogled vzeli tudi relativne frekvence in range **n-gramov** (bi- in trigramov) – poglavje 3.1.3,
- da bodo prevodi še boljši, smo implementirali še **metodo opazovanja konteksta** besed. Njena naloga je izkoristiti kontekst (besede v okolici) prevajane besede za iskanje iste besede v drugem besedilu preko prevodov besed iz konteksta – poglavje 3.1.4,
- za konec smo v aplikacijo implementirali še t.i. **metodo posebnih besed** – poglavje 3.1.5.

3.1 Uporabljene ideje v pristopu

3.1.1 Pristop s primerjavo relativnih frekvenc besed

Najbolj osnovna ideja v pristopu je ideja o podobnosti relativnih frekvenc med besedami dveh jezikov. To je pristop, za katerega pričakujemo, da bo deloval bolje za besede z višjo, kot tiste z nižjo frekvenco. Razlog za to je preprost. V besedilih načeloma velja pravilo, da so besede

razporejene podobno kot je to prikazano na sliki 3.1. Le-ta nam kaže stanje v izbranem angleškem besedilu, analiziranem v 4. poglavju. Rdeči stolpiči prikazujejo posamezne vrednosti relativnih frekvenc, s katerimi nastopajo besede v besedilu, modri pa število besed, ki imajo to vrednost relativnih frekvenc. Najpogostejša beseda se pojavlja z relativno frekvenco 2,99 % - taka beseda je le ena. Tretje najpogostejša se pojavlja z relativno frekvenco 1,89 %, taki besedi sta že 2. Medtem ko imajo najmanj pogoste besede relativno frekvenco le še 0,03 %, takih besed pa je kar 546. Če bomo torej prevajali besedo z najnižjo relativno frekvenco, ji bomo morali iz tega besedila za prevod dodeliti eno od besed z najnižjimi relativnimi frekvencami, se pravi eno od 546 besed – tu pa je težko določiti, katera je prava. Če pa iščemo prevod za besede z višjimi relativnimi frekvencami, bo primernih kandidatov relativno malo, pogosto le eden – tu bomo takoj vedeli, kateri prevod je pravi.



Slika 3.1: Razporeditev relativnih frekvenc besed tipičnega besedila

Kako naj bi takšna logika delovala, si oglejmo na primeru analiziranih besedil iz 4. poglavja. Iz besedil smo naredili dve tabeli (tabeli 3.1 in 3.2), ki vsebujeta vse besede besedil z izjemo mašil. Tabeli po vrsti iz leve proti desni v stolpcih vsebujeta: zaporedne range frekvenc (i), št. pojavitev (#), relativno frekvenco ter besedo samo. Besede so urejene po padajočih frekvencah in so spremenjene v njihovo osnovno obliko oz. so korenjene – več o korenjenju bomo povedali v poglavju 3.3.

i	#	relativna frekvenca	beseda
1	87	2.9907	translat
2	64	2.2	gram
3	55	1.8906	target
4	55	1.8906	word
5	45	1.5469	languag
6	45	1.5469	sourc
7	39	1.3406	cbmt
8	35	1.2031	corpu
9	34	1.1687	overlap
10	34	1.1687	synonym
11	31	1.0656	base
12	27	0.9281	candid
13	27	0.9281	score
14	25	0.8594	process
15	25	0.8594	sentenc
16	24	0.825	text
17	23	0.7906	machin
18	23	0.7906	context
19	23	0.7906	test
20	22	0.7562	dictionari
21	22	0.7562	gener
22	20	0.6875	market
23	19	0.6531	requir
24	18	0.6187	english
25	18	0.6187	phrase
26	17	0.5843	index
27	16	0.55	flood
28	15	0.5156	decod
29	15	0.5156	method
30	15	0.5156	exampl
31	15	0.5156	stock
32	14	0.4812	parallel
…	…	…	…
237	2	0.0687	klein
…	…	…	…
416	2	0.0687	oxford
417	1	0.0343	grassi
…	…	…	…
962	**1**	**0.0343**	**decalc**

Tabela 3.1: Nabor besed angleškega besedila

i	#	relativna frekvenca	beseda
1	179	5.9488	prev
2	78	2.5922	jezik
3	77	2.5589	besed
4	40	1.3293	progr
5	36	1.1964	slov
6	32	1.0634	korp
7	31	1.0302	term
8	31	1.0302	podr
9	31	1.0302	trans
10	28	0.9305	tehn
11	28	0.9305	orodj
12	27	0.8973	upor
13	25	0.8308	pomn
14	24	0.7976	račun
15	24	0.7976	razv
16	20	0.6646	inter
17	19	0.6314	stroj
18	19	0.6314	sist
19	19	0.6314	določ
20	17	0.5649	strok
21	16	0.5317	iskan
22	15	0.4985	bank
23	15	0.4985	evrop
24	15	0.4985	vzpor
25	14	0.4652	proj
26	13	0.432	drug
27	12	0.3988	štev
28	12	0.3988	ustr
29	12	0.3988	enot
30	11	0.3655	izraz
31	11	0.3655	infor
32	10	0.3323	nekat
…	…	…	…
267	2	0.0664	špel
…	…	…	…
470	2	0.0664	dipl
471	1	0.0332	spel
…	…	…	…
1010	1	0.0332	pouč

Tabela 3.2: Nabor besed slovenskega besedila

Za zgled si oglejmo primer, kjer iščemo prevod za besedo *prev* (koren besede *prevajanje*). Na podlagi relativnih frekvenc lahko predpostavimo, da bo pravi prevod v angleščini *translate*, saj je iz tabele ciljnega jezika njena relativna frekvenca najbližja prevajani besedi. Kot vidimo, je naša predpostavka za prevod bila uspešna. V praksi se tudi izkaže, da lahko besede na vrhu tabele pogosto prevedemo na ta način. Čim nižje po tabeli gremo, manj uspešni bomo pri prevajanju. Kot že rečeno, ni nujno da bo prevod vedno isti, kot bi bil prevod po slovarju.

Pogosto se zgodi, da za prevod na ta način dobimo besedo, ki ima podoben pomen prevajani besedi, ni pa popolnoma pravilen. Primer za to je beseda *gram*. Če bi jo želeli prevesti, bi v slovenščini za prevod dobili besedo *besed*. Prevod sicer ni popolnoma pravilen, je pa pomen besede dober, saj je *gram* okrajšava za n-gram, kar pa pomeni besedo sestavljeno iz *n* besed.

Včasih se lahko tudi zgodi, da imamo za možen prevod v ciljnem jeziku na razpolago 3 besede z enako relativno frekvenco. Če bi iskali prevod za besedo *parallel* bi za njo dobili 3 možne prevode, in sicer *bank, evrop* in *vzpored*. V takih primerih bodo na pomoč morale priskočiti druge ideje, ki jih bomo predstavil v naslednjih poglavjih.

Na prikazanih simboličnih primerih pristop deluje lepo in prav, v resnici pa ni vedno tako. Če bi npr. iskali prevod za besedo *corpu* (z relativno frekvenco 1,2031), bi najprimernejši kandidat za prevod bila beseda *slov* (z relativno frekvenco 1,1964). To pa seveda ni pravilno. Na srečo pa je bil pravi prevod, *korp* (z relativno frekvenco 1,0634), s svojo relativno frekvenco vseeno blizu dobljenega prevoda *slov*. Zato bi bilo najbolje, da za prevod ne predlagamo le ene, najverjetnejše besede, ampak zajamemo vse tiste, ki imajo relativne frekvence blizu relativne frekvence prevajane besede. Dovolili bi torej majhno odstopanje relativnih frekvenc. Če bi v omenjenem primeru dovolili odstopanje za 0,15 bi za možne prevode dobili *slov, progr* in *korp* (*korp* je koren besede *korpus*).

3.1.2 Pristop z rangi

Relativne frekvence besed v obeh tabelah se ne ujemajo pri vseh besedah. Pogosto se zgodi, da relativne frekvence besed iz enega besedila odstopajo od relativnih frekvenc besed v drugem besedilu. To je se je zgodilo tudi na našemu primeru iz tabel 3.1 in 3.2. Angleške besede imajo na vrhu tabele precej višjo relativno frekvenco od slovenskih. Idealno bi namreč bilo, da najbolj pogosta beseda iz originalnega besedila ustreza najbolj pogosti besedi iz ciljnega besedila, druga najbolj pogosta beseda drugi najbolj pogosti, tretja tretji itn., ne glede na to ali se relativne frekvence omenjenih parov razlikujejo ali ne.

Za takšno iskanje prevodov, ki se ne ozira na vrednost posameznih relativnih frekvenc, lahko uporabimo rang relativnih frekvenc. Pomemben naj bi bil torej le rang frekvence (v nadaljevanju: rang) in ne sama relativna frekvenca. Kot dokaz za to si ponovno oglejmo primer prevajanja besede *stroj*. Če bi namesto same relativne frekvence upoštevali le zaporedno velikost relativne frekvence, bi se 17. najbolj pogosta beseda prevedla v 17. najbolj pogosto besedo in prevod bi bil pravilen. Na enak način bi uspešno prevedli besedo *transl*. Pri parih besed *word* in *besed, languag* in *jezik, index* in *število, phrase* in *izraz,* itn., pa bi potrebovali le malo večjo množico kandidatnih rešitev in pravi prevodi bi že bili med kandidati za prevode. Da bi to dosegli, bi pri prevajanju besede *besed* morali dopustiti razliko v rangu za 1, pri prevajanju besede *languag* za 3, pri besedi *index* za 1, pri *phrase* za 5 itn.

3.1.3 Ideja z n-grami

O n-gramih govorimo, kadar na *n* zaporednih znakov ali besed gledamo kot na eno enoto. N-grami so uporabni predvsem pri kriptografiji, lingvistiki ter pri analizi besedil oziroma pri statistiki besed. Če je n=1 je n-gram pravzaprav beseda sama, če je n=2, govorimo o bigramih, če je n=3 govorimo o trigramih itn. V nalogi bomo za analizo besedil od n-gramov uporabili le bigrame in trigrame, večji *n* pa bi se splačalo uporabiti le pri daljših besedilih. Ker se relativne frekvence n-gramov obnašajo podobno kot relativne frekvence prej omenjenih posameznih besed, se nam splača tudi n-grame obravnavati na enak način kot posamezne besede. Lahko torej pričakujemo, da se bodo n-grami, ki vsebujejo istopomenske besede,

pojavljali s približno enakimi relativnimi frekvencami tako v originalnemu kot v ciljnem jeziku.

Za lažje razumevanje si ponovno oglejmo zgled na prejšnjem primeru. Primer besedil je isti kot v prejšnjem poglavju, le da bomo namesto posameznih besed analizirali bigrame. Trigramov tukaj ne bomo predstavili, saj je stvar analogna primeru z bigrami. Pristop iskanja prevodov na podlagi bigramov deluje na enak način kot pri posameznih besedah: pri prevajanju bomo kot pravilen prevod obravnavali tiste bigrame, ki nastopajo pri soležnih rangih / relativnih frekvencah prevajanega bigrama. Da je logika pristopa dobro zastavljena, si oglejmo primer iz tabel 3.3 in 3.4; če bi iskali prevod za bigram *machin translat*, bi po rangih za prevod dobili bigram *stroj prev*, kar bi seveda bil pravilen prevod. Na enak način bi pri prevajanju *parallel text* dobili *vzpored korp* (korpus je zbirka tekstov) ter z minimalno razliko v rangih pri prevajanju *target gram* dobili *prev beseda* (tukaj si bigrama sicer nista popolnoma enakovredna, imata pa podoben pomen).

i	#	relativna frekvenca	bigram
1	22	0.7565	target languag
2	20	0.6877	machin translat
3	15	0.5158	synonym synonym
4	11	0.3782	stock market
5	10	0.3438	bilingu dictionari
6	10	0.3438	sourc word
7	10	0.3438	word phrase
8	9	0.3094	blind test
9	8	0.2751	parallel text
10	8	0.2751	sourc target
11	8	0.2751	spanish english
12	8	0.2751	bleu score
13	8	0.2751	exampl base
14	8	0.2751	word phrasal
15	8	0.2751	target gram
16	7	0.2407	base machin
17	7	0.2407	translat model
18	7	0.2407	target corpu
19	7	0.2407	candid translat
20	7	0.2407	translat candid
21	7	0.2407	index target
…	…	…	
266	1	0.0343	translat jaim
…	…	…	…
2346	1	0.0343	philadelphia usa

Tabela 3.4: Nabor bigramov angleškega besedila

i	#	relativna frekvenca	bigram
1	20	0.6648	pomn prev
2	17	0.5651	stroj prev
3	13	0.4321	term bank
4	13	0.4321	jezik tehn
5	10	0.3324	progr pomn
6	8	0.2659	prev enot
7	7	0.2327	prev prev
8	7	0.2327	tehn slov
9	7	0.2327	vzpor korp
10	6	0.1994	tehn prev
11	6	0.1994	prev progr
12	6	0.1994	inst jožef
13	6	0.1994	jožef stef
14	6	0.1994	slov jezik
15	6	0.1994	določ besed
16	5	0.1662	prev besed
17	5	0.1662	jezik prev
18	5	0.1662	jezik virov
19	5	0.1662	prev ustr
20	5	0.1662	progr orodj
21	5	0.1662	trans memory
…	…	…	…
227	1	0.0332	prev špel
…	…	…	…
2548	1	0.0332	pouč predm

Tabela 3.3: Nabor bigramov sloveskega besedila

Kot smo torej videli do sedaj, lahko pričakujemo, da bo tudi pri bigramih pristop dobro deloval. Iskanje primernega prevoda s pomočjo n-grama pa vseeno ni tako enostavno, kot smo to videli pri posameznih besedah. Pri posameznih besedah smo namreč imeli enostavno nalogo, saj smo se soočali s situacijo prevajanja »beseda za besedo«. Pri n-gramih pa se soočamo s situacijo »n-besed za n-besed«. Odločiti se bomo torej morali:

a) katero od *n* besed iz izvirnega jezika bomo upoštevali in
b) katero od *n* besed iz ciljnega jezika bomo upoštevali.

Dilemo pojasnjujemo v nadaljevanju. Prevajana beseda običajno nastopa v različnih bigramih izvirnega besedila, zato se nam lahko pojavi vprašanje, katero od njih naj upoštevamo. Recimo, da s pomočjo bigramov iz prejšnjega primera želimo najti prevod za besedo *machin.*

Kot lahko vidimo, naša beseda v tabeli nastopa na dveh različnih mestih, in sicer v bigramu *machin translat* in bigramu *base machin*. Tu se nam torej postavi vprašanje, kateri bigram (oz. relativno frekvenco od tega bigrama) naj upoštevamo? Težava je še večja; če bi pogledali celotno tabelo s 2346 bigrami, bi videli, da *machin* nastopa mnogo več kot le dvakrat. V uvodu smo omenili, da je osnova za prevajanje ravno ta, da lahko vsaki besedi oz. n-gramu v okviru besedila predpostavimo neko v naprej pričakovano relativno frekvenco (torej le eno). Zato je smiselno in nujno, da se omejimo na samo en, določen bigram, ki ima svojo določeno relativno frekvenco. Ker najboljše rezultate dajejo n-grami z vrha tabele, bo najbolje, da vedno upoštevamo le ta en n-gram. Ostale n-grame pa bomo zanemarili, saj kot rečeno, dajejo tisti z dna tabele precej slabše rezultate. V našem primeru bomo torej upoštevali le bigram z rangom 2, tako da bodo kandidati za prevod le besedi *stroj* in *prev*. S tem smo množico možnih rešitev iz 4 omejili na 2.

Ker želimo prevajanje optimizirati kolikor se le da, se nam pojavi še drugo vprašanje, katero besedo od bigrama bomo upoštevali v ciljnem jeziku oz. v našem primeru: v katero od obeh besed bigrama *stroj prev* pa naj se prevede naša beseda *machin*? Tega nam bigrami na žalost ne znajo povedati. Vseeno pa obstaja način, ki nam lahko razkrije, katera beseda od obeh je verjetnejša. Vprašanje se namreč razreši že na nivoju posameznih besed. Kot smo lahko videli, se bo na nivoju posameznih besed beseda *machin* prevedla v *stroj*, tako ima beseda *machin* v primerjavi s *prev*, že prednost. Če bomo torej hoteli najti pravo rešitev, bomo morali kombinirati metodo posameznih besed z metodo n-gramov.

3.1.4 Ideja o kontekstu prevoda

Če pri opazovanju več besed skupaj ne upoštevamo točnega zaporedja besed, temveč samo njihovo prisotnost v bigramu, govorimo o kontekstu besede v bigramu. Kontekst je torej več besed skupaj ali del besedila (enota, besede, stavki, ideja, govor …), ki ima pogosto tudi nek svoj zaokrožen pomen. Podobno kot na relativno frekvenco in range besed, lahko gledamo tudi na kontekst kot na lastnost jezika. Vsaka beseda se v besedilu pojavlja v okviru določenih kontekstov; v enih bolj, v drugih manj. Če vzamemo recimo besedo *prevajanje*, lahko rečemo, da se bo večinoma pojavljala v bližini njej sodelovalnih besed, kot so npr. *strojno, poceni, avtomatsko, način, kvalitetno* ipd. Zato lahko rečemo, da jo bomo težje zasledili v bližini besed, ki nimajo kaj dosti opravka z njo, kot recimo *balon, rumen, letenje, šal* ipd.

Pristopi, ki smo jih omenjali do sedaj (relativne frekvence, rangi in n-grami), za prevode določenih besed pogosto predlagajo tudi nesmiselne prevode. Vzemimo primer, kjer iščemo 4 najverjetnejše prevode za besedo *prevajanje.* Recimo, da smo po določenih dosedanjih metodah našli naslednje 4 kandidate za prevod: *strojno, avtomatsko, rumeno* in *balon*. Človeškemu prevajalcu ne bo težko ugotoviti, katera izmed teh besed je pravi kandidat. Prav tako mu ni težko ugotoviti, da besedi *balon* in *rumeno* nimata nobene veze s prevajano besedo. Stroji pa ne razmišljajo kot ljudje, zato bi jim v temu primeru bilo potrebno pomagati izbrati, katera od 4 besed naj bi bila prava. V takšnih primerih nam pride prav ideja upoštevanja konteksta. Iz konteksta bomo namreč znali ugotoviti, da besedi *rumen* in *balon* ne moreta biti prava prevoda, saj ne nastopata v primernih kontekstih z besedo *prevajanje*. Tako bo število kandidatov iz zgoraj omenjenega primera iz 4 omejeno na 2.

3.1.5 Ideja o posebnih besedah

Kot nam je znano iz življenja, vsebujejo tehnične besede pogosto posebne znake. Takšne besede lahko razdelimo med

1) besede, ki vsebujejo posebne znake (npr. #, -, _ ipd.),

2) besede, ki vsebujejo števila,

3) besede, ki vsebujejo velike črke.

V našem sistemu smo zato implementirali metodo, ki ugotavlja, ali je beseda »posebna« v smislu, da vsebuje posebne znake, ki so navedeni zgoraj. Ta pristop je torej namenjen predvsem prevajanju tehničnih besed, saj imajo tehnične besede pogosto vsaj eno od omenjenih lastnosti. Tako bi omenjena metoda posebno pozornost namenila besedam, kot so Wi-Fi, LAN, SMT, c64, win32 ipd.

3.2 Točkovni sistem prevodov besede

Pri iskanju pravega prevoda primerjamo vsako besedo ciljnega jezika s prevajano besedo. Pri tem se sprehajamo po tabeli ciljnih besed in primerjamo lastnosti vsake besede iz tabele ciljnih besed s prevajano besedo. Na podlagi podobnosti med njima se besedam originalnega jezika dodelijo ustrezne točke. Tista ciljna beseda, ki ima bolj podobne lastnosti prevajani besedi, dobi več točk, tista z manj, pa manj točk. K lastnostim štejemo range, relativne frekvence, posebnosti v besedi ipd. Ko točkujemo vse besede, jih razvrstimo v tabelo glede na število dobljenih točk. Najprimernejši kandidat za prevod bo torej prva beseda iz tabele. Če nas recimo zanimajo prvi trije najverjetnejši prevodi, so to po vrsti prve tri besede iz vrha tabele.

Da bodo prevodi čim bolj natančni, je pomembno, da besede ciljnega jezika točkujemo na pameten način. Tukaj predvsem mislim na primerno utežitev rezultatov posameznih uporabljenih opisanih pristopov k iskanju prevodov – v primeru, da prevajamo na podlagi več metod hkrati. Če sta besedili v povprečju krajši, je pametneje dajati večji pomen metodi posameznih besed kot metodi bi- in trigramov. Če imamo v besedilih veliko tehničnih izrazov, bi bilo pametno bolj poudariti metodo posebnih besed. Ker pa vhodnih besedil do potankosti ne moremo predpisati vnaprej, smo se po testiranjih nekaj različnih besedil odločili, da bodo vse metode imele enak vpliv na točkovanje kandidatov. Tako lahko vsaka od 5-ih metod besedam prinese 20 % vseh točk. Za cilj si zadamo sistem implementirati tako, da lahko uporabnik sam določi, po katerih kriterijih oziroma metodah se bo prevod izračunal.

Da bomo dobili občutek, kako se besede točkujejo, si oglejmo primere za vsako od implementiranih metod. Algoritme iskanja prevoda smo razdelili na 2 veliki skupini:

1) osnovne metode in
2) metoda upoštevanja konteksta.

K osnovnim metodam spadajo:

1) metoda posameznih besed,

2) metoda bigramov,
3) metoda trigramov in
4) metoda posebnih besed.

Metoda posameznih besed: sestavljata jo dve (pod)metodi, ki se izvajata zaporedno, ena za drugo, to sta metoda relativnih frekvenc in metoda rangov. Vsaka beseda se naprej obdela z obema postopkoma točkovanja, točke obeh se na koncu procesa točkovanja seštejejo pri vsaki besedi. Formula, po kateri se točkujejo besede pri metodi rangov, je:

$$točke\ (beseda_i) = 1 - dif_i / (dolžina\ seznama\ -1), \quad (1)$$

kjer je

$$dif_i = |rang\ (prevajana\ beseda) - rang\ (beseda_i\ ciljnega\ besedila)| \quad (2)$$

Formula za točkovanje po metodi rangov pa je:

$$točke\ (beseda_i) = 1 - dif_i / (f(max) - f(min)), \quad (3)$$

kjer je

$$dif_i = |relativna\ frekvenca\ (prevajana\ beseda) - relativna\ frekvenca\ (beseda_i\ ciljnega\ besedila)| \quad (4)$$

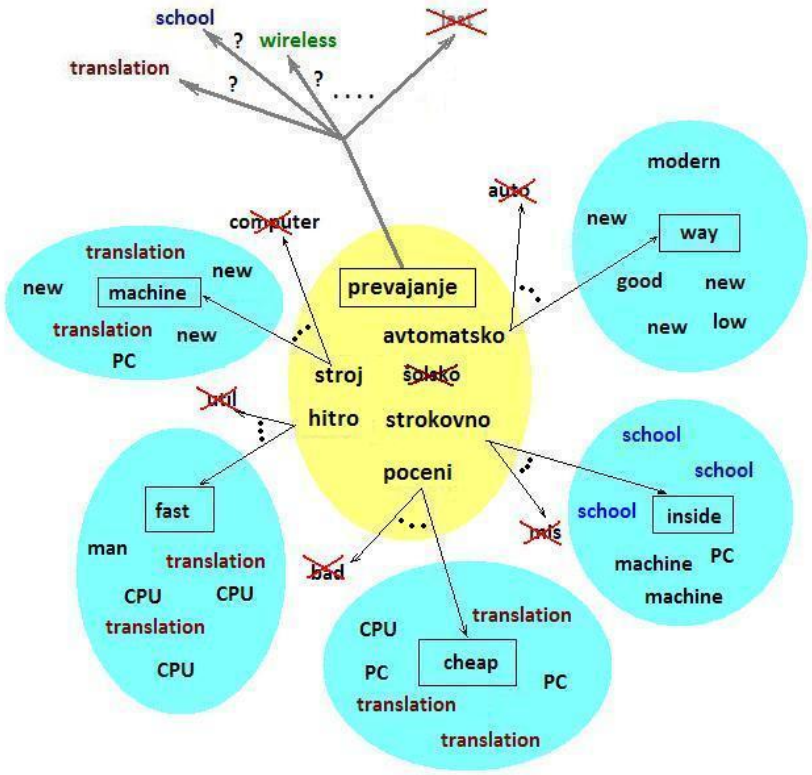
Metoda bi- trigramov: relativne frekvence in rangi n-gramov se obravnavajo na enak način kot relativne frekvence in rangi posameznih besed. To pomeni, da lahko tudi nad njimi uporabimo enake formule kot nad posameznimi besedami, torej formuli (1) in (3). Razlika je le, da v formuli namesto posameznih besed upoštevamo bi- oziroma trigrame.

Metoda posebnih besed: logika predlaganja kandidatov za prevod je tukaj zelo enostavna. Proces poteka tako, da algoritem pogleda, če beseda, ki jo želimo prevesti, vsebuje kakšno posebnost. Če jo vsebuje, se iz množice ciljnega besedila, točkujejo vse tiste besede, za katere prav tako velja, da vsebujejo kakšno posebnost.

Metoda upoštevanja konteksta: sestoji iz 8 korakov:

1) prevajanje: iskanje možnih prevodov za besedo, ki jo želimo prevesti,
2) redčenje: od prevodov, dobljenih v 1. koraku, upoštevamo le najboljše/najverjetnejše,
3) kontekst: iskanje besed, ki so v kontekstu prevajane besede,
4) redčenje: iz 3. koraka izberemo le tiste besede, katerih prevodom bomo zaupali,
5) prevajanje: iskanje prevodov za dobljene besede iz 4. koraka,
6) redčenje: izberemo samo najverjetnejše prevode, ki smo jih dobili v 5. koraku,
7) kontekst: iskanje besed, ki so v kontekstu prevodov, ki smo jih dobili v 6. koraku,
8) iskanje rešitev: iskanje skupnih besed med tistimi, ki smo jih dobili v 2. in 7. koraku.

Podrobnosti korakov algoritma si oglejmo še na primeru, ki ga opisuje slika 3.2:



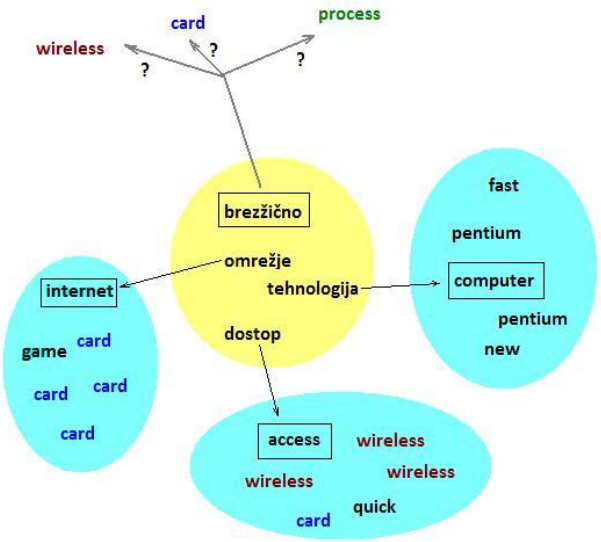
Slika 3.2: Metoda upoštevanja konteksta: Iskanje pravega prevoda. V območju znotraj rumene elipse so besede, ki spadajo v kontekst prevajane besede (v pravokotniku). Z uporabo prevodov konteksta (modre elipse) si pomagamo najti pravi prevod prevajane besede.

1. Recimo, da želimo v nekem besedilu najti prevod za besedo *prevajanje*. Recimo, da smo s pomočjo drugih metod ugotovili, da so možni kandidati za prevod (po vrsti): *school*, *wireless*, *translation*, *last* ...
2. Vse možne prevode, dobljene iz 1. koraka, omejimo le na izbrano število najbolj verjetnih (ta parameter določi uporabnik sam preko uporabniškega vmesnika). Recimo, da nam v primeru ostanejo samo še trije najverjetnejši možni kandidati: *school*, *wireless* in *translation.* Vsi ostali prevodi, kot recimo beseda *last*, pa iz nadaljnje obravnave odpadejo.
3. Algoritem pogleda, v kakšnem kontekstu nastopa beseda prevajanje. Recimo, da v njeni bližini oz. kontekstu najde naslednje besede: *stroj, avtomatsko, strokovno, hitro, šolsko* in *poceni*.
4. Besede, dobljene v 3. koraku, bo potrebno prevesti, nekateri prevodi od teh besed pa bodo po pričakovanju lahko tudi slabi. Tem ne gre zaupati, zato jih je bolje izločiti. Meja se določi tako, da uporabnik v vmesniku nastavi parametre za besede, ki izstopajo (outlier). V primeru na sliki je na ta način izpadla beseda *šolsko*.
5. Prevajalni sistem prevede besede, ki smo jih dobili v 4. koraku. Primer na sliki prikazuje le po 2 prevoda za vsako besedo, to sta najbolj- in najmanj verjeten. V resnici ima vsaka beseda tabelo vseh prevodov, kjer so le-ti razvrščeni po verjetnosti (to nam nakazujejo tri pike). Prevodi na sliki so*: machine, computer, fast, util, cheap, bad, inside, mis, auto* in *way*.
6. Kot smo omenili v 5. točki, je prevodov od vsake besede veliko. Zato jih je spet treba omejiti na le najverjetnejše. V našem primeru smo se odločili, da bomo za vsako

besedo naredili le po en, najverjetnejši prevod. Prevode, ki so manj verjetni, zanemarimo – na sliki smo jih prečrtali.

7. V tej točki mora algoritem za besede, ki smo jih dobili v prejšnji točki, poiskati kontekst besed. Za vsako od besed *machine, fast, cheap, inside* in *way* je treba poiskati besede, ki nastopajo v njihovih kontekstih. Tako dobljene besede so na sliki predstavljene znotraj modrih balončkov.
8. V zadnjem koraku moramo samo še pogledati, katere besede od možnih kandidatov za prevod (*school, wireless, translation*) nastopajo najpogosteje med tistimi, ki smo jih dobili v 7. koraku. Kot vidimo, besede *wireless* med njimi sploh ne zasledimo, beseda *school* se je pojavila le trikrat, beseda *translation* pa bo najbrž pravi kandidat, saj se pojavlja v kontekstih največ prevodov (*machine, fast* in *cheap*), poleg tega pa se v kontekstih nasploh pojavlja dokaj pogosto.

Lahko se tudi zgodi, da v vseh omenjenih korakih postopamo pravilno, vseeno pa pride do okoliščin, kjer nam metoda ne predlaga pravilne rešitve. Poglejmo si takšen primer na sliki 3.3 (za enostavnejšo razlago smo korake redčenja izpustili). Tukaj bi nam algoritem za pravi prevod predlagal *card*, saj se le-ta pojavlja največkrat. Prava rešitev pa bi bila beseda *wireless*. Logika mora zato biti zasnovana tako, da se ne osredotočimo le na najverjetnejšo rešitev in točkujemo samo to, temveč tudi ostale kandidate. Najbolje bi bilo točkovati sorazmerno s številom pojavitev v vseh kontekstih. Torej bi beseda *card* dobila 5 točk, beseda *wireless* 3 točke, *pentium* 2, *game* 1 itn. V resnici je stvar še bolj zapletena, saj je treba upoštevati relativno število pojavitev besed v celotnemu tekstu, ne samo v nastopih kontekstov. Recimo, da se *card* v celotnem besedilu pojavi 10x, *wireless* pa le 3x. Tedaj lahko rečemo, da smo v kontekst zajeli polovico vseh pojavitev besede *card* ter čisto vse pojavitve besede *wireless*. Pravi prevod bo torej *wireless*. Tako smo tudi zasnovali točkovanje v aplikaciji. Pomembno je še, da niti beseda *card* ne bo ostala brez točk; vsaka beseda dobi toliko točk, kolikokrat se pojavi v vseh kontekstih glede na razmerje z njenim skupnim številom pojavitev.



Slika 3.3: Iskanje pravega prevoda z metodo upoštevanja konteksta, kjer upoštevamo tudi relativno pogostost pojavljanja besed

3.3 Predprocesiranje besedila

Odstranjevanje mašil. Pred razlago ozadja delovanja algoritmov moramo razjasniti še nekaj osnovnih pojmov. Naloga aplikacije je prevajanje besed, in sicer predvsem besed tehnične narave. Tehnične besede so namreč za pristop, ki smo ga uporabili, primernejše, saj naj bi se v besedilih s podobno temo pojavljale s podobnejšimi relativnimi frekvencami kot ostale besede. Zato smo se odločili, da bomo eno veliko skupino netehničnih besed v resnici tudi prezrli. Te besede se imenujejo »mašila« (angl. *stopwords*), včasih jih zasledimo tudi pod imenom »prazne besede«. Njihova značilnost je, da imajo zelo visoke relativne frekvence ter da nosijo le slovnično informacijo, ne pa tudi pomenske [23]. Ker so za analizo besedil številnim aplikacijam, kot recimo spletnim brskalnikom, nezanimive, jih aplikacije med analizo ponavadi izločijo. Tako bomo naredili tudi mi.

K mašilom spadajo ti. slovnične oz. zaprte besedne vrste [23], npr. nedoločni členi – v pogovorni slovenščini (en, ena …), predlogi (v, nad, pred …), prislovi, členki (niti, ne, nikoli …), vezniki (in, ter, pa …) itn. Idealnega in univerzalnega seznama teh besed ni, temveč se pogosto gradi kar iz frekvenčnih seznamov besed, dobljenih iz čim večjega števila besedil. Idealni seznami so odvisni od aplikacije, ki jih uporabljajo – za našo uporabo bi bilo torej najbolje, da bi seznam sestavili sami. Ker pa bi to bilo preveč zamudno, se bomo v naši aplikaciji zadovoljili z obstoječimi seznami, dostopnimi na spletu. To sicer vpliva na slabše rezultate, ampak ker je jezik preveč kompleksen pojem, idealnega seznama nikoli ne bomo imeli.

Večina mašil bi se z našo logiko prevajanja zelo težko prevedla, saj se mašila v vsakem naravnem jeziku uporabljajo na drugačen način, s tem pa je tudi njihova pogostost pojavljanja različna. Nemščina, na primer, pozna 3 določne člene: *die, der* in *das*, angleščina le enega: *the*, slovenščina pa določenega člena sploh ne pozna. Iz tega primera lahko vidimo, da bi takšne skupine besed naši logiki prevajanja delale precejšnje težave. Povrhu je za mašila značilno, da se pojavljajo v velikemu številu [20], kar pomeni da imajo velik delež relativnih frekvenc.

Korenjenje besed. Drug pomemben pristop, ki se ga moramo lotiti pred samo analizo besed, je korenjenje besed. Kot vemo, se iste besede lahko uporabljajo v različnih pomenskih oblikah. Npr. beseda *prevajanje* lahko nastopa v oblikah *prevajalec, prevod, prevodu, prevodi, prevedel, prevesti, preveden* itn. Da bomo v naši raziskavi dobili pravilne rezultate, bomo morali upoštevati tudi to lastnost: najti vse oblike iste besede. Naše delo je torej, da s korenjenjem najdemo skupnega prednika besed oziroma osnovno besedo. To bomo dosegli s postopkom, ki se imenuje korenjenje (angl. *stemming*), izvajajo pa ga algoritmi za korenjenje besed (angl. *stemmers* ali *stemming algorithms*).

S korenjenjem se ukvarja tudi jezikovna morfologija, ki pravi, da je korenjenje krnjenje pregibanih (sklanjanih, spreganih …) besed na njihove osnovne korene [21]. Pri tem procesu se lahko pogosto zgodi, da kot koren dobimo besede, ki sploh niso veljavne besede, kot npr. koren *prev*, kar je koren za besedo *prevajanje*. Korenjenje ponavadi uporabljajo spletni iskalniki pri rangiranju, pogosto se uporablja tudi pri procesiranju naravnih jezikov [21]. Tudi korenjenje je dandanes že avtomatizirano, na voljo pa je kar nekaj algoritmov, še posebej za pomembnejše svetovne jezike. Za slovenski jezik na žalost to ne velja. Zato smo v aplikaciji morali uporabiti »domači« algoritem za korenjenje besed. Izvirnik [24] smo za naše potrebe prevedli v Javo.

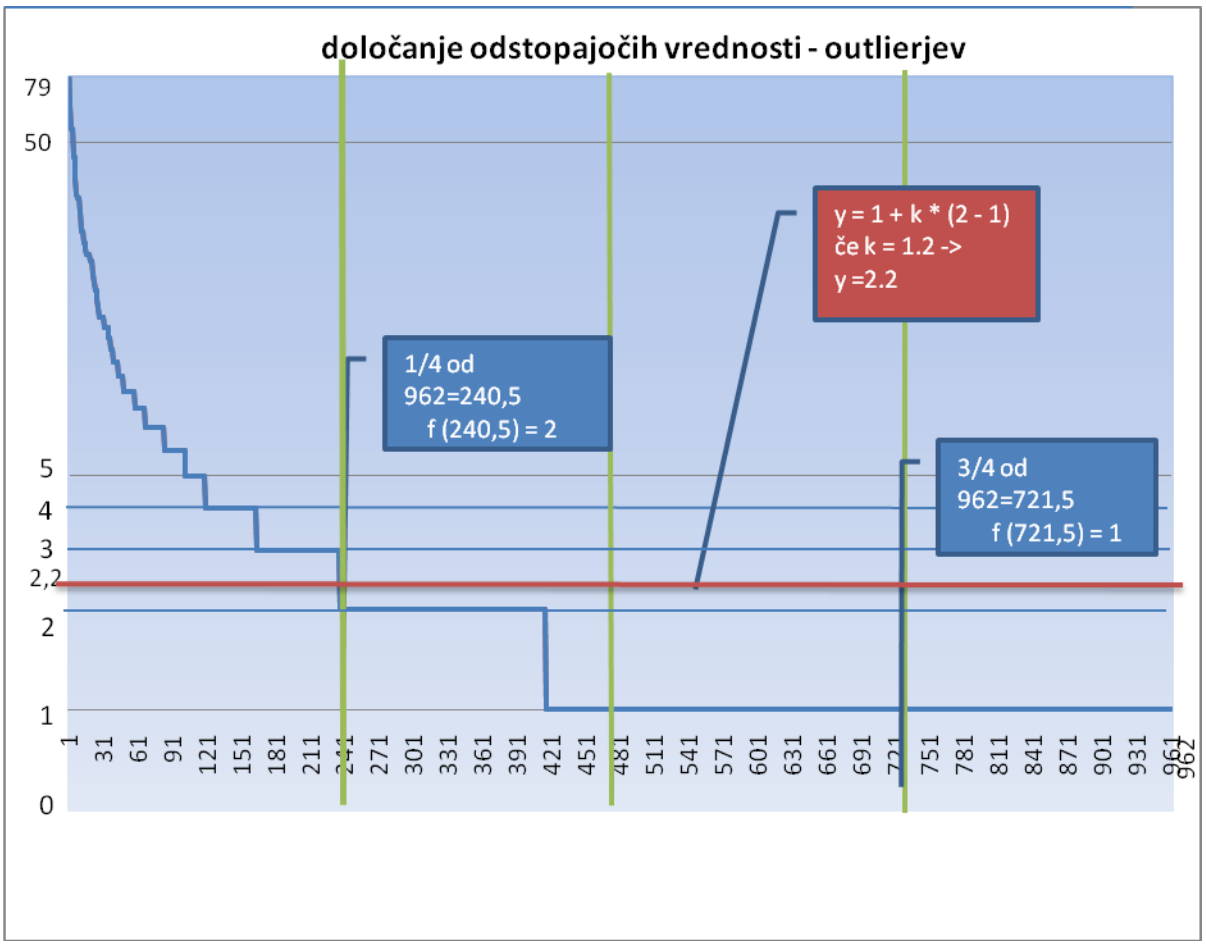
Izbor besed nad pragom zaupanja. Kot zadnje si oglejmo še način izbora boljše prevedenih besed, ki ga bomo potrebovali pri implementaciji metode upoštevanja konteksta. Metoda upoštevanja konteksta v svojem postopku potrebuje prevode drugih besed v kontekstu. Tukaj mora biti pazljiva, saj če bo upoštevala slabe prevode besed iz konteksta, iz tega niti sama ne bo znala dobro prevesti prevajane besede. Zato bomo uvedli delitev besed med tiste, katerih prevodom zaupamo, in tiste, katerih prevodom ne zaupamo.

Kot smo omenili že v uvodu, lahko besedam z višjo frekvenco bolj zaupamo kot besedam z nižjo. Pri določanju te meje si bomo pomagali s postopkom določanja odstopajočih vrednosti (angl. *outliers*), ki se ga razširjeno uporablja v statistiki [22]. Postopek se ponavadi uporablja za iskanje tistih točk, ki preveč odstopajo od ostalih, saj s tem kvarijo porazdelitev populacije, zato se iz nadaljnje analize izločijo. V našemu primeru bo ravno obratno: poiskali bomo tiste točke oziroma besede, ki se ločijo od ostalih tako, da imajo višjo relativno frekvenco od večine. Tako dobljene besede, ki torej izstopajo po relativni frekvenci v pozitivno smer, bomo upoštevali v nadaljnji analizi, ostale pa izločili.

V nalogi smo celotni nabor podatkov razdelili na štiri enako velike dele (kvartile). Označimo vrednost relativne frekvence na zgornji meji *i*-tega kvartila z $f(x_i)$. Pozitivno izstopajoči podatki so za nas potem tisti, ki ležijo nad tretjim kvartilom, ki mu prištejemo polovico interkvartilnega razmika (s čimer torej določimo pričakovano zgornjo mejo porazdelitve):

$$y = f(x_3) + 0.5 * (f(x_3) - f(x_1)) \quad (5)$$

Na primeru si oglejmo, kako bi opisana formula (5) delovala na primeru slovenskega besedila, ki ga bomo analizirali v 4. poglavju. Dobljeni graf je predstavljen na sliki 3.4. Besedilo ima vsega skupaj 962 besed (brez mašil), njihovi rangi so nanizani na vodoravni osi, medtem ko na navpični osi najdemo frekvenco vsake od teh besed. Npr. besede od ranga 417 do 962 se pojavljajo v besedilu le 1x, besede od ranga 237 do 416 2x itn. Meja med 1. in 2. kvartilom se nahaja na ¼ od 926 besed, kar je med 240. in 241. besedo – to nam označuje leva navpična zelena črta. Na podoben način dobimo mejo med 3. in 4. kvartilom med 721. in 722. besedo – desna zelena navpičnica. Pri omenjenih navpičnicah velja $f(240,5) = 1$ in $f(721,5) = 2$. Če bi koeficient bil enak 0,5, kot smo omenili v zgornji formuli, bi mejo za izstopajočo vrednosti dobili torej pri 1,5. To bi pomenilo, da bi izpadle tiste besede, ki se pojavijo manj kot 1, 5x (v praksi to pomeni, da torej izpadejo tiste besede, ki se pojavijo 1x, zaupamo pa besedam, ki se pojavijo vsaj 2x).



Slika 3.4: Določanje outlierjev – meje prevodov, ki jim zaupamo

V praksi se izkaže, da besedam, ki se ponavljajo 2x še ne moremo zaupati, zato moramo to mejo nekoliko dvignili in zaostriti faktor upoštevanja interkvartilnega razmika 0,5. Po nekaj empiričnih testih smo ugotovili, da bi ga za obravnavano besedilo bilo bolje nastaviti na 1,2. Tako z novim izračunom dobimo mejo pri 2,2 (rdeči balonček). V tem primeru nam odpadejo vse besede, ki se pojavljajo 1x in 2x, zaupali pa bomo besedam ki se pojavijo vsaj 3x. Postavljanje meje na tak način je pametno, saj upošteva razporeditev frekvenc. Koeficient 1,2 bo namreč primeren za večino besedil. Ker pa prej omenjen koeficient ni vedno primeren, je sistem za prevajanje pametno implementirati tako, da bo uporabnik lahko sam nastavil ta koeficient kot tudi število kvantilov, ki se jih upošteva pri definiciji izstopajočih vrednosti.

4 Implementacija sistema za strojno prevajanje

4.1 Uporabljena orodja

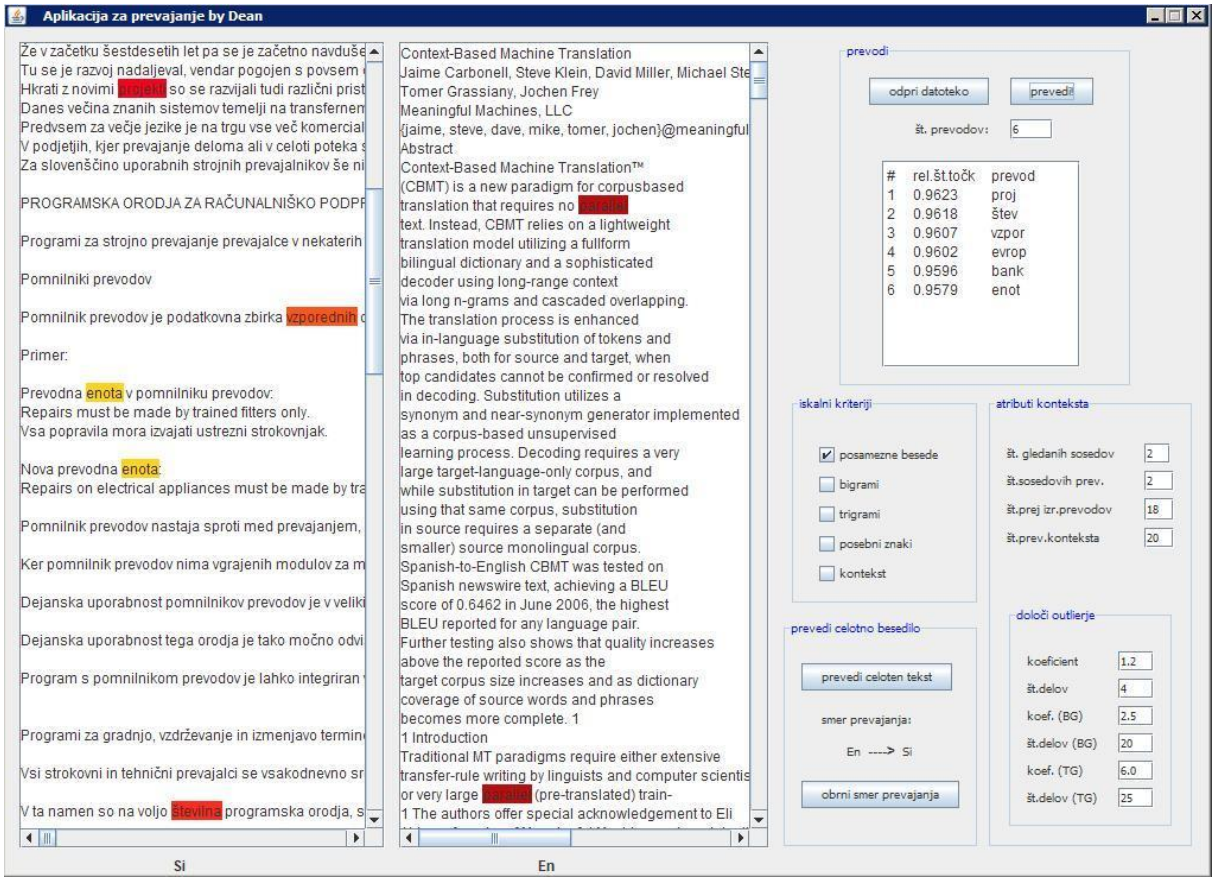
Pri izbiri programskega jezika za razvoj aplikacije, smo morali upoštevati nekatere pomembne stvari, kot so enostavnost razvoja in uporabe grafičnega vmesnika, dostopnost pomembnih orodij za delo z besedili (knjižnice za delo z algoritmom za korenjenje besed in seznami mašil), hitrost delovanja aplikacije, zanesljivost itn. Eden izmed programskih jezikov, ki podpira večino teh zahtev, je programski jezik Java. Zato smo se odločili, da bomo sistem razvili v njem. Razvojno orodje, ki smo ga uporabili v kombinaciji z Javo je Eclipse IDE for Java developers [25].

V aplikacijo smo vključili tudi knjižnico z algoritmom za korenjenje besed za jezike, ki jih bomo v naši aplikaciji želeli uporabljati. Na spletu smo našli algoritme za korenjenje besed za mnoge evropske jezike [26,27]. V aplikacijo smo implementirali le tiste, ki so za nas aktualni, torej za: angleščino, nemščino, italijanščino, francoščino in španščino. Za slovenski jezik, na žalost, prosto dostopnih algoritmov za korenjenje besed še ni, zato ni bilo druge možnosti, kot da smo algoritem za korenjenje slovenskih besed dobili iz drugega vira [24].

Za angleški jezik lahko najdemo precejšnje število prosto dostopnih seznamov mašil na spletu, prav tako tudi za pomembnejše svetovne, predvsem evropske jezike. Na spletu [27] smo našli zbirko seznamov za vse implementirane jezike razen italijanščine in slovenščine. Za ta dva jezika smo sezname poiskali posebej, na drugih straneh [28,23]. Ker smo tekom testiranja aplikacije naleteli na nekaj bistvenih pomanjkljivosti v seznamih, smo nekatere besede dodali ročno. Take besede so npr. *the* (angleška beseda *the*, ki se pogosto pojavlja tudi v slovenskih besedilih), *npr*, *itd*, *Erjavec* (imena, priimki …), *mogoče*, *lahko* itn. Vse razrede algoritmov za korenjenje omenjenih jezikov smo zbrali v skupni paket.

4.2 Funkcionalnosti sistema

Načrtovani in implementirani uporabniški vmesnik našega sistema za prevajanje je prikazan na sliki 4.1.



Slika 4.1: Uporabniški vmesnik implementiranega sistema za prevajanje: primer prevajanja besede *parallel*

Delovanje sistema za prevajanje lahko strukturno opišemo na naslednji način:

I) <u>gradnja baze podatkov za prevajanje</u>

1. iz besedila se naredi seznam vseh besed,
2. s seznama se izločijo mašila ter nezaželene besede in znaki,
3. besede se korenijo,
4. naredi se seznam bigramov in trigramov,
5. zgradi se baza besed:
    a. ugotovi se, če beseda vsebuje posebnosti,
    b. prešteje se število pojavitev enakih besed,
    c. besede v tabeli se sortirajo glede na število pojavitev,
    d. izračunajo se relativne frekvence in rangi relativnih frekvenc za vsako besedo v tabeli.

II) <u>proces prevajanja</u>

1. določi se smer prevajanja,

2. preveri se, če je za prevajanje vse nared (ne prevajamo mašil; vsi parametri so v redu; naloženi sta obe besedili; označili smo natanko eno besedo, ki jo želimo prevesti),
3. korenimo prevajano besedo,
4. iz baze besed poiščemo vse podatke za prevajano besedo,
5. iskanje prevoda,
6. izpis rezultatov v uporabniškem vmesniku.

Prva točka, gradnja baze podatkov, se zgodi ob vsakokratnem nalaganju novega besedila v aplikacijo. Kadar želimo naložiti novo besedilo v prevajalni sistem, v uporabniškem vmesniku (glej sliko 4.1) pritisnemo gumb »odpri datoteko«. Za vsako novo besedilo se mora namreč zgraditi nova baza besed in podatkov. Druga točka, dejansko prevajanje, pa se zgodi, ko v uporabniškem vmesniku z gumbom »prevedi« aktiviramo proces iskanja prevoda. Cilj procesa je točkovanje besed ciljnega jezika. Beseda, ki pri tem zbere največ točk, bo za prevod najprimernejši kandidat.

Proces prevajanja (II. točka) sprožimo tako, da v enem od obeh okenc z besedilom označimo besedo, ki jo želimo prevesti nato pa pritisnemo gumb »prevedi«.

4.3 Aplikacija za prevajanje

Aplikacijo smo implementirali v obliki datoteke *prevajalnik.jar*. V aplikaciji lahko naložimo datoteki tipa *txt*, med katerima želimo iskati pare besed, ki predstavljajo prevode. Da bo aplikacija vedela, v katerem jeziku je katera datoteka, podamo oznako jezika v imenu datoteke (De = nemščina , Es = španščina, En = angleščina, Fr = francoščina, It = italijanščina, in Si = slovenščina).

Prevajanje posameznih besed. Ko imamo naloženi obe besedili, lahko začnemo prevajati. Iz enega od obeh naloženih besedil, ki se bosta prikazali v oknu aplikacije, označimo besedo, ki jo želimo prevesti, in pritisnemo gumb »prevedi«. V angleškem besedilu se bodo pobarvale vse besede *parallel.* Ko program izračuna najverjetnejše prevode, se nam bodo le-ti izpisali v oknu »prevodi«, hkrati pa se bodo v besedilu ciljnega jezika te besede še ustrezno pobarvale. Prevajamo lahko eno besedo za drugo, prevajanje pa deluje v obe smeri. Število kandidatnih prevodov lahko tudi sami nastavimo v razdelku »prevodi«. Če s prevodi nismo zadovoljni, lahko njihovo število tudi povečamo. Če želimo preizkusiti, kako prevajajo različne implementirane metode, lahko v razdelku »iskalni kriteriji« izberemo eno ali več metod, s katerimi želimo prevajati. Privzeto je ta nastavljena na osnovno metodo, to je metodo posameznih besed. Če prevajamo z metodo upoštevanja konteksta, je temu namenjen razdelek »atributi konteksta«, kjer lahko nastavimo parametre metode upoštevanja konteksta.

Na sliki 4.1 lahko vidimo primer prevajanja besede *parallel* iz angleškega besedila. Aplikacija je zanjo našla naslednje možne prevode: *proj, štev, vzpor, evrop, bank* in *enot* – to lahko vidimo v razdelku »prevodi« na desni strani vmesnika. Tukaj so prevodi izpisani po vrsti, glede na verjetnost pravilnega prevoda. Najverjetnejši prevod je *proj*. V angleškem besedilu se nam hkrati glede na verjetnost pobarvajo omenjeni prevodi, najtemneje se pobarva najverjetnejši prevod - *proj*. Kot lahko vidimo je pravi prevod, *vzporednih*, pobarvan z oranžno, saj je le 3. mestu verjetnosti za pravi prevod.

Prevajanje celega besedila. Kot smo omenili že v uvodu, je zanimivo tudi pogledati, kako bi aplikacija prevedla celotno besedilo. To lahko naredimo z gumbom »prevedi celoten tekst«. Pri tem je smer prevajanja določena pod omenjenim gumbom. Če želimo spremeniti smer, lahko to storimo z gumbom »obrni smer prevajanja«. Kot vemo, mašil aplikacija ne zna prevesti, zato bo jih pri prevajanju celotnega besedila pustila v originalnem jeziku ter jih označila z zvezdico.

Kot smo že omenili, je najbolje, da so za izračun prevoda nekateri parametri nastavljivi s strani uporabnika. Za metodo upoštevanja konteksta so ti parametri naslednji:

- opcija »število gledanih sosedov« predstavlja obseg konteksta. Če je št. gledanih sosedov 3, naša analiza seže 3 besede naprej in 3 besede nazaj od gledane besede. Kontekst bo torej skupaj obsegal 6 besed. Primer s takšno velikostjo konteksta lahko vidimo na sliki 3.2,

- opcija »število prej izračunanih prevodov« nam pove, koliko začetnih prevodov bomo pri prevajanju za možne končne prevode sploh upoštevali. Na sliki 3.3 je ta parameter bil 3, v realnem svetu pa je bolje, da je mnogo večji, zato je v aplikaciji pod privzetem nastavljen na 20,

- s »številom sosedovih prevodov« lahko določimo, koliko najprimernejših prevodov želimo zajeti za nadaljnjo obravnavo od vsakega soseda, tj. od vseh besed v kontekstu. Na sliki 3.3 smo za vsako besedo iz konteksta upoštevali le po 1 prevod,

- na koncu moramo določiti, koliko najprimernejšim kandidatom bomo podelili točke. To nastavimo z možnostjo »število prevodov konteksta«. Točke se delijo linearno po vrstnem redu primernosti. Če bi na sliki 3.2 to število nastavili na 2, bi točke dobila *translation* in *new*, če bi to število nastavili na 4, pa bi točke dobila še *PC* in *CPU*.

Ko imamo nastavljene vse parametre konteksta, se lahko lotimo še nastavitve meje za izstopajoče besede. Tukaj najdemo po dva atributa za posamezne besede, za bigrame in za trigrame. Ta dva atributa sta koeficient *k* in število kvantilov *n*, ki smo ju omenjali v razdelku 3.3 o izračunu izstopajočih vrednosti.

5 Praktični primer in evalvacija delovanja sistema

5.1 Opis poskusa

Da dobimo vpogled v uspešnost delovanja našega pristopa, smo aplikacijo preizkusili na mnogih različnih besedilih, med različnimi jeziki. V tem poglavju bomo predstavili rezultate enega takšnega poskusa, katerega rezultate smo vzeli pod drobnogled. Za poskus smo morali izbrati dve primerni, podobni si besedili. Da bi bila tema čim bolj aktualna, smo se odločili za besedili iz področja strojnega prevajanja, in sicer eno slovensko in eno angleško besedilo.

Cilj poskusa je pogledati in oceniti, kako natančno zna naša aplikacija iskati pravilne prevode. Uspešnost prevajanja smo testirali z vsemi implementiranimi metodami razen metode posebnih besed (besedili nista primerni za to metodo). Osnovne metode smo kombinirali z metodo upoštevanja konteksta, tako da je vseh kombinacij metod bilo 6. Prevajalo se je v obe smeri, torej iz slovenščine v angleščino in obratno.

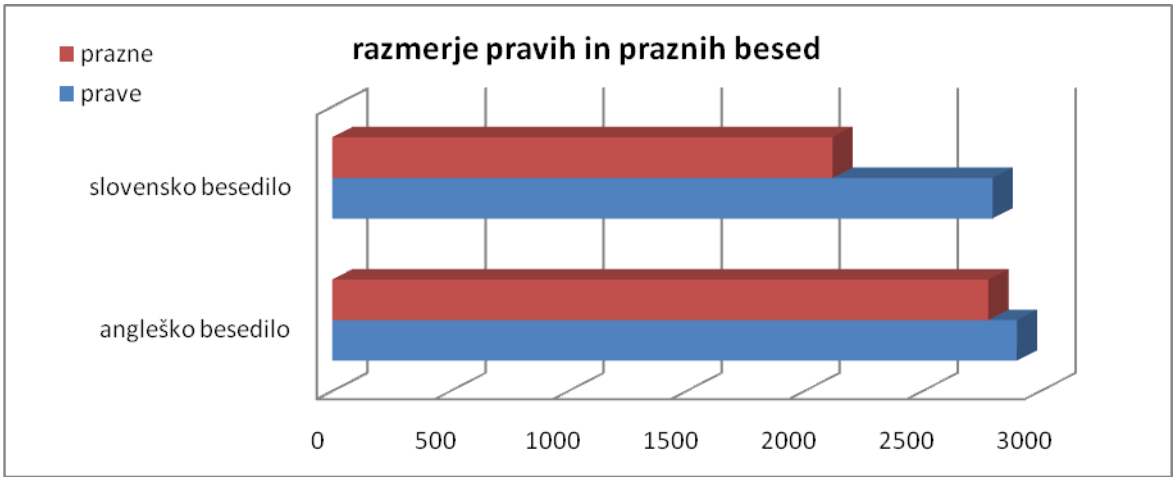
Uspešnost smo merili na 2 načina: 1) če je (najverjetnejši) prevod pravilen ter 2) če je pravi prevod med nekaj prvimi (recimo 5) najverjetnejšimi kandidati za prevod. Tako je vseh merjenj uspešnosti bilo 6 x 2 (št. jezikov) x 2(načini) = 24 - glej tabelo 5.8. Po vseh dobljenih rezultatih smo lahko primerjali še uspešnost vseh omenjenih metod med seboj.

5.2 Preliminarna analiza obeh besedil

Besedili, ki smo ju za analizo izbrali, naj bi si bili vsebinsko čim bolj podobni. Za uspeh našega pristopa je to nujen pogoj. Slovenščina je eden izmed jezikov, v katerem, na žalost, nimamo vselej na razpolago takšnega besedila, kot bi si želeli. Tako smo se morali zadovoljiti z besediloma, ki sodeč po naslovih, razen strojnega prevajanja nimata kaj dosti skupnega. Kljub temu, da eno besedilo govori o tehnologijah prevajanja, drugo pa o kontekstnem prevajanju, imata še vedno skupno lastnost, da obe govorita o strojnem prevajanju [9, 8] – glej sliko 5.1. Tako pričakujemo, da se bodo dobro prevedle besede, ki so skupne obema temama, npr. *jezik, slovar, iskanje, beseda* itn. Besede, ki pa so lastnost bodisi enega besedila (npr. beseda *kontekst*) bodisi drugega besedila (npr. besede *tehnologija*, *pristop*), pa seveda ne bodo znale najti svojega primernega ekvivalenta v drugem besedilu.

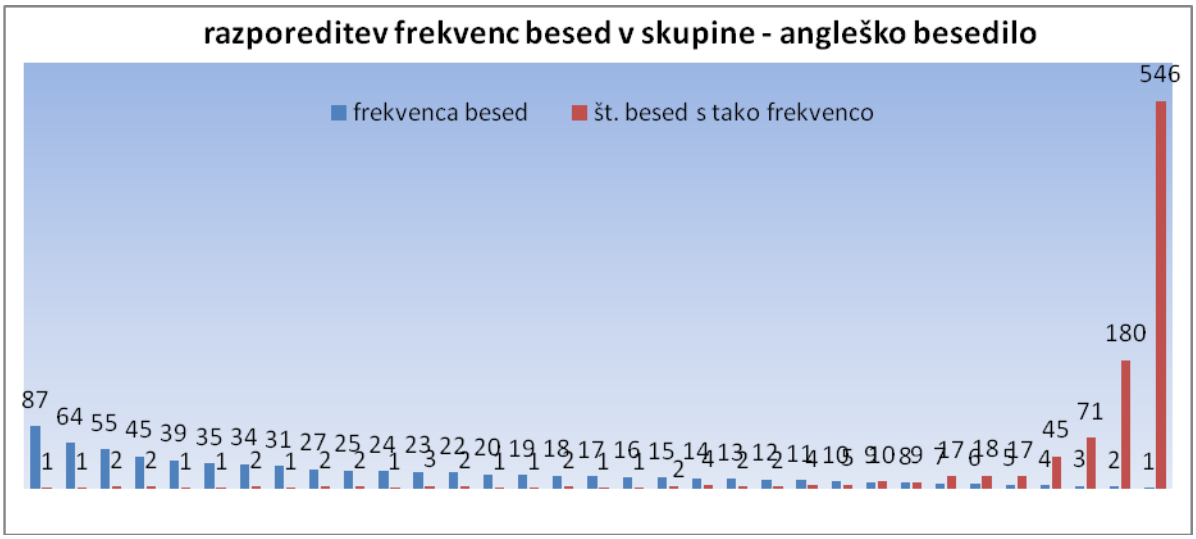
Slika 5.1: Prekrivanje tem/besed slovenskega in angleškega besedila

Oglejmo si nekaj statistik za podani besedili. Na začetku analize smo morali ločiti »prave« besede od mašil. Njihovo razmerje lahko vidimo na sliki 5.2. Slovensko besedilo sestavlja 4918 besed, od tega je le 2876 »pravih«, ostala pa so mašila. V angleškem besedilu je vsega skupaj 5681 besed, od tega le 2798 »pravih«. Vidimo, da ima angleško besedilo precej večjo množico zaznanih mašil. To je posledica razpoložljivosti seznama mašil za ta jezik kot tudi sama lastnost naravnega jezika.

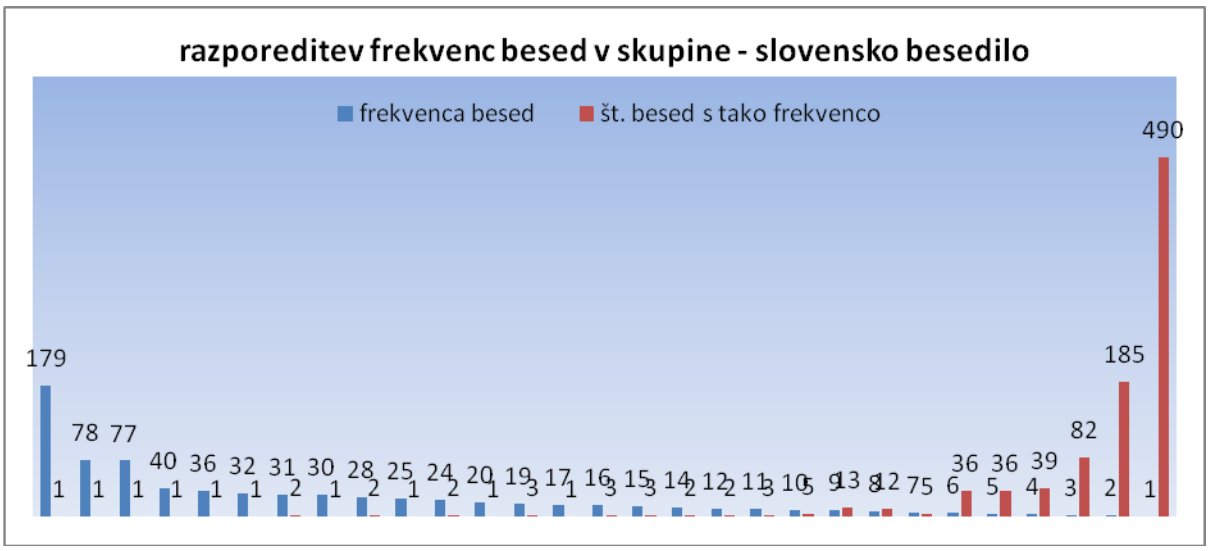


Slika 5.2: Razmerje »pravih« besed in mašil v obeh besedilih

Ko smo izločili za nas nepomembna mašila, lahko začnemo z dejansko analizo besedil. Poglejmo si, kako so porazdeljene besede, ki so ostale, torej prave besede (od tukaj naprej bomo obravnavali le še prave besede). Grafa na slikah 5.3 in 5.4 nam prikazujeta, koliko je besed, ki se ponavljajo z določenimi frekvencami. V slovenskem besedilu imamo npr. eno besedo, ki se pojavi 179x ter 490 besed, ki se pojavijo le 1x. Ni težko opaziti, da obstaja podobnost med porazdelitvijo besed v obeh besedilih.

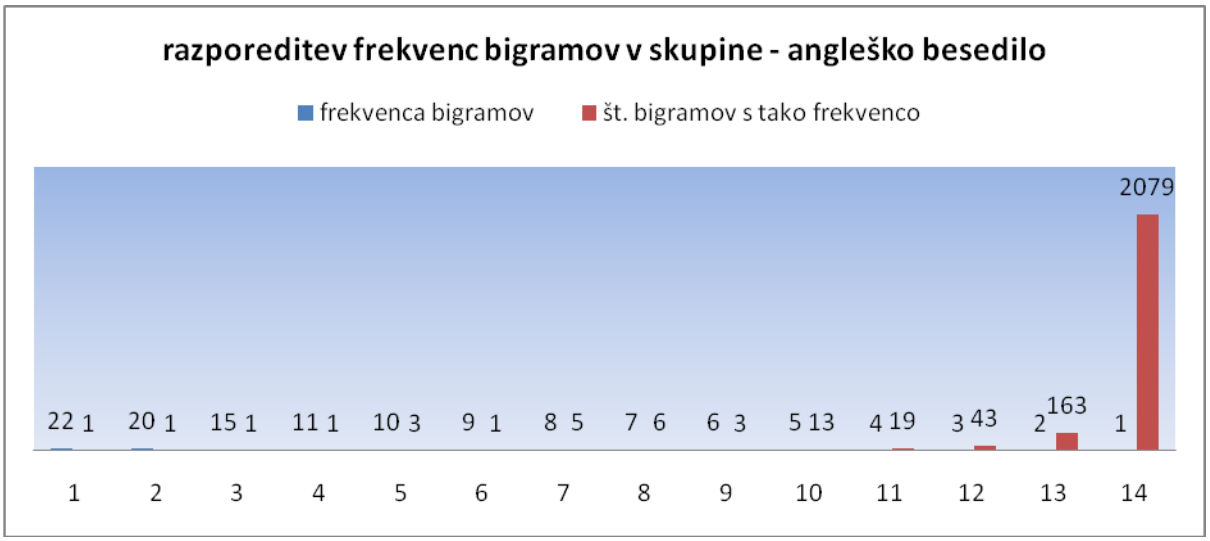


Slika 5.3: Razporeditev frekvenc skupin besed za angleško besedilo

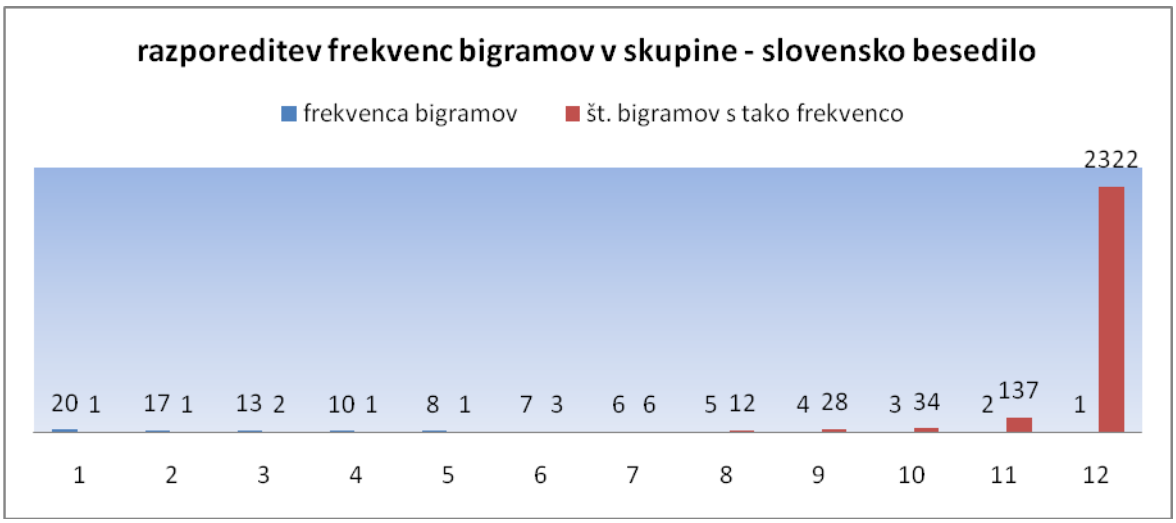


Slika 5.4: Razporeditev frekvenc skupin besed za slovensko besedilo

Če pogledamo enako razporeditev pri bigramih, vidimo, da le-ta ni več takšna, kot je bila pri posameznih besedah. Na slikah 5.5 in 5.6 lahko vidimo, da je bigramov z različno pogostostjo pojavljanja bistveno manj – v angleškem besedilu je le 14 skupin bigramov z različnimi frekvencami, v slovenskem pa le 12. Poleg tega se pojavljajo tudi z bistveno nižjimi frekvencami kot posamezne besede. Bigram, ki se pojavlja najpogosteje, se je pojavil le 20x v slovenskem in 22x v angleškemu besedilu. Lastnost bigramov je še ta, da je veliko takih, ki se v besedilih pojavijo le enkrat – v angleškem besedilu je takih 2079, v slovenskem pa celo 2322. To je seveda slabo, saj pomeni, da je ogromno takih enot, ki jim bomo težko našli primeren prevod.



Slika 5.5: Razporeditev frekvenc skupin bigramov za angleško besedilo



Slika 5.6: Razporeditev frekvenc skupin bigramov za slovensko besedilo

Opisano razporeditev pogostosti pojavljanja besed lahko podamo tudi tabelarično. Tako bomo lahko točno videli, za katere besede gre. Tabeli 5.1 in 5.2 prikazujeta stanje za prvih 20 najpogostejših besed obeh besedil (i=rang, #=št. pojavitev). Kot lahko vidimo, velja omenjena lastnost; v obeh besedilih oz. tabelah najdemo besede, ki so skupne temi »strojno prevajanje«: *jezik* in *language*, *word* in *besed*, *translat* in *prev* itn.

i	#	relativna frekvenca	beseda
1	87	2.9907	translat
2	64	2.2	gram
3	55	1.8906	target
4	55	1.8906	word
5	45	1.5469	languag
6	45	1.5469	sourc
7	39	1.3406	cbmt
8	35	1.2031	corpu
9	34	1.1687	overlap
10	34	1.1687	synonym
11	31	1.0656	base
12	27	0.9281	candid
13	27	0.9281	score
14	25	0.8594	process
15	25	0.8594	sentenc
16	24	0.825	text
17	23	0.7906	context
18	23	0.7906	machin
19	23	0.7906	test
20	22	0.7562	dictionari

Tabela 5.1: Frekvence besed angleškega besedila

i	#	relativna frekvenca	beseda
1	179	5.9488	prev
2	78	2.5922	jezik
3	77	2.5589	besed
4	40	1.3293	progr
5	36	1.1964	slov
6	32	1.0634	korp
7	31	1.0302	term
8	31	1.0302	podr
9	31	1.0302	trans
10	28	0.9305	tehn
11	28	0.9305	orodj
12	27	0.8973	upor
13	25	0.8308	pomn
14	24	0.7976	račun
15	24	0.7976	razv
16	20	0.6646	inter
17	19	0.6314	stroj
18	19	0.6314	sist
19	19	0.6314	določ
20	17	0.5649	strok

Tabela 5.2: Frekvence besed slovenskega besedila

Na enak način lahko v tabelah 5.3, 5.4, 5.5 in 5.6 pogledamo, kakšno je stanje pogostosti pojavljanja bigramov in trigramov. V tem primeru tabela obsega le prvih 10 enot, saj imamo na vrhu tabel pogosto nesmiselne n-grame.

i	#	relativna frekvenca	bigram
1	22	0.7565	target languag
2	20	0.6877	machin translat
3	15	0.5158	synonym synonym
4	11	0.3782	stock market
5	10	0.3438	bilingu dictionari
6	10	0.3438	sourc word
7	10	0.3438	word phrase
8	9	0.3094	blind test
9	8	0.2751	parallel text
10	8	0.2751	sourc target

Tabela 5.3: Frekvence bigramov angleškega besedila

i	#	relativna frekvenca	bigram
1	20	0.6648	pomn prev
2	17	0.5651	stroj prev
3	13	0.4321	term bank
4	13	0.4321	jezik tehn
5	10	0.3324	progr pomn
6	8	0.2659	prev enot
7	7	0.2327	vzpor korp
8	7	0.2327	tehn slov
9	7	0.2327	prev prev
10	6	0.1994	tehn prev

Tabela 5.4: Frekvence bigramov slovenskega besedila

i	#	relativna frekvenca	trigram
1	6	0.2063	base machin translat
2	6	0.2063	target languag corpu
3	6	0.2063	associ machin translat
4	5	0.1719	machin translat america
5	5	0.1719	translat sourc word
6	5	0.1719	exampl base machin
7	4	0.1375	requir parallel text
8	4	0.1375	form bilingu dictionari
9	4	0.1375	confer associ machin
10	4	0.1375	gram translat candid

Tabela 5.6: Frekvence trigramov angleškega besedila

i	#	relativna frekvenca	trigram
1	10	0.3325	progr pomn prev
2	6	0.1995	jezik tehn slov
3	6	0.1995	inst jožef stef
4	5	0.1662	gradn term bank
5	5	0.1662	saarbrü universitä des
6	5	0.1662	universitä des saarl
7	4	0.133	oddel prev tolm
8	4	0.133	račun podpr prev
9	4	0.133	sist stroj prev
10	4	0.133	zborn konf jezik

Tabela 5.5: Frekvence trigramov slovenskega besedila

5.3 Rezultati uspešnosti prevajanja

V tem poglavju si bomo ogledali, kakšni so rezultati prevajanja na zgoraj omenjeni dvojici besedil. Uspešnost smo izmerili za vsako metodo (razen metodo posebnih besed) posebej, nato pa še kombinacijo teh metod z metodo upoštevanja konteksta. Tukaj velja omeniti, da metoda upoštevanja konteksta prinaša 1/5 točk k točkam osnovnih metod - torej je poudarek še vedno pretežno na osnovnih metodah, metoda upoštevanja konteksta le »pomaga« pri izbiri kandidatov. Prevajali smo v obe smeri jezikov. Kombinacije testiranih metod so torej:

a) posebne besede (iz angl. v slo.)
b) bigrami (iz angl. v slo.)
c) trigrami (iz angl. v slo.)
d) pos. besede + kontekst (iz angl. v slo.)
e) bigrami + met. upoštevanja konteksta (iz angl. v slo.)
f) trigrami + met. upoštevanja konteksta (iz angl. v slo.)
g) pos. besede (iz slo. v angl.)
h) bigrami (iz slo. v angl.)
i) trigrami (iz slo. v angl.)
j) pos. besede + kontekst (iz slo. v angl.)
k) bigrami + met. upoštevanja konteksta (iz slo. v angl.)
l) trigrami + met. upoštevanja konteksta (iz slo. v angl.)

V tabeli 4.7 lahko vidimo primer, kako je sistem z metodo posameznih besed prevedel najbolj pogostih 100 besed angleškega besedila v slovenščino – torej tabela za primer a). Za merjenje uspešnosti smo uvedli parameter u_i, ki predstavlja število uspešno prevedenih prevodov od 100 najbolj pogostih besed tako, da je pravi prevod med prvimi *i*-timi kandidati za prevod. Če je npr. i=1, kot najden uspešen prevod upoštevamo le prvo besedo izmed najbolj verjetnih prevodov, in če je i=5, upoštevamo kot uspešen prevod tistega, ki ima pravilen prevod med prvimi petimi kandidati za prevod. Ravno takšen primer lahko vidimo v tabeli 4.7.

Če določimo i=1, vidimo, da je aplikacija pravilno prevedla le 2 besedi: *prev* in *besed*. Velja torej u_1= 2 %, kar je precej slabo. V nadaljevanju smo se odločili za merjenje uspešnosti z oceno u_5, ki daje višji rezultat, in sicer 12 %. Uspešnosti u_1 in u_5 za vseh 12 metod smo predstavili v naslednjem poglavju.

Podobno kot v tabeli 4.7, dobimo rezultate tudi za ostalih 11 metod, ki jih zaradi prostorskih razlogov ne bomo predstavili (rezultati so na razpolago na priloženi zgoščenki). Pri tabelah velja poudariti, da bi u_i z naraščajočo množico vzorca ustrezno padal. Nemogoče je namreč pričakovati, da bo aplikacija predlagala dobre prevode tudi za (naj)manj frekvenčne besede. Posledica tega je, da so ponavadi pravilni prevodi večinoma na vrhu tabele – to lahko vidimo tudi na tabeli 5.7.

V spodnjih delih tabel, kjer se nahajajo manj pogoste besede, pogosto nastopajo tudi nesmiselne besede. To so tiste besede, ki jih nismo dobro ločili oz. »očistili«, npr. *WWW//saarbrücke*, okrajšave *(npr. cbmt, al)*, imena *(John, New York)*, pomožne besede (*figure, tabel*), besede, ki so značilne le za eno od besedil (*market, context*) itn. Zaradi tega je za opazovani vzorec smiselno vzeti bolj pogoste besede.

i beseda	> kandidati za prevod (zloženi po verjetnosti)
1 prev	> translat, gram, target, word, languag, sourc
2 jezik	> gram, translat, target, word, languag, sourc
3 besed	> gram, translat, target, word, languag, sourc
4 progr	> cbmt, corpu, overlap, synonym, languag, sourc
5 slov	> corpu, overlap, synonym, cbmt, base, candid
6 korp	> base, overlap, synonym, corpu, candid, score
7 term	> base, candid, score, overlap, synonym, corpu
8 podr	> base, candid, score, overlap, synonym, corpu
9 upor	> base, candid, score, process, overlap, sentenc
10 tehn	> candid, score, process, sentenc, text, base
11 orodj	> candid, score, process, sentenc, text, base
12 pomn	> text, process, sentenc, context, machin, test
13 račun	> context, machin, test, text, process, sentenc
14 razv	> context, machin, test, text, process, sentenc
15 inter	> requir, market, phrase, english, dictionari, gener
16 stroj	> phrase, requir, english, market, index, flood
17 sist	> phrase, requir, english, market, index, flood
18 določ	> phrase, requir, english, market, index, flood
19 strok	> flood, index, phrase, english, decod, method
20 velik	> flood, decod, method, exampl, stock, index
21 iskan	> flood, decod, method, exampl, stock, index
22 prim	> flood, decod, method, exampl, stock, index
23 bank	> decod, method, exampl, stock, parallel, flood
24 evrop	> decod, method, exampl, stock, parallel, flood
25 vzpor	> decod, method, exampl, stock, parallel, flood
26 štev	> parallel, spanish, decod, proceed, method, model
27 proj	> parallel, spanish, decod, proceed, method, model
28 ustr	> lattic, typic, us, model, bleu, pair
29 enot	> lattic, typic, us, model, bleu, pair
30 infor	> us, bleu, result, blind, lattic, typic
31 izraz	> us, bleu, result, blind, lattic, typic
32 pogos	> us, bleu, result, blind, lattic, typic
33 proc	> bilingu, confirm, us, monolingu, bleu, improv
34 zadnj	> bilingu, confirm, us, monolingu, bleu, improv
35 stran	> bilingu, confirm, us, monolingu, bleu, improv
36 pomoč	> bilingu, confirm, us, monolingu, bleu, improv
37 podj	> bilingu, confirm, us, monolingu, bleu, improv
38 razl	> develop, bilingu, describ, confirm, us, associ

39 vrst	> develop, bilingu, describ, confirm, us, associ
40 oblik	> develop, bilingu, describ, confirm, associ, monolingu
41 trans	> develop, bilingu, describ, confirm, associ, monolingu
42 skup	> develop, bilingu, describ, confirm, associ, monolingu
43 volj	> develop, bilingu, describ, confirm, associ, monolingu
44 konk	> develop, describ, confirm, associ, bilingu, monolingu
45 možn	> develop, describ, associ, monolingu, figur, confirm
46 večj	> develop, describ, associ, figur, improv, left
47 leks	> develop, describ, associ, figur, left, human
48 anal	> develop, describ, associ, figur, left, maxim
49 potr	> describ, develop, associ, figur, left, maxim
50 angl	> associ, describ, figur, develop, left, maxim
51 orod	> figur, associ, left, form, describ, maxim
52 danes	> left, form, figur, maxim, resourc, associ
53 podat	> form, maxim, resourc, left, n-gram, refer
54 integ	> form, resourc, n-gram, refer, maxim, system
55 lastn	> form, resourc, refer, system, match, n-gram
56 povez	> form, resourc, refer, match, phrasal, search
57 poseb	> form, resourc, refer, match, search, sign
58 posam	> form, resourc, refer, match, search, chang
59 podob	> resourc, form, refer, match, search, chang
60 imam	> refer, resourc, match, form, search, chang
61 gradn	> match, refer, search, resourc, chang, form
62 sezn	> search, match, chang, refer, brown, resourc
63 elekt	> substitut, newswir, qual, chang, rule, search
64 virov	> substitut, newswir, qual, rule, brown, **origin**
65 omog	> substitut, newswir, qual, rule, origin, summit
66 kakov	> substitut, newswir, qual, rule, origin, paper
67 post	> newswir, substitut, qual, rule, origin, paper
68 letih	> qual, newswir, rule, substitut, origin, paper
69 hkrat	> rule, qual, origin, newswir, paper, substitut
70 prist	> origin, rule, paper, qual, systran, newswir
71 poved	> paper, origin, systran, rule, level, qual
72 struk	> systran, paper, level, origin, time, rule
73 zbirk	> level, systran, time, paper, compon, origin
74 izdel	> time, level, compon, systran, confer, paper
75 inst	> compon, time, confer, level, contain, systran
76 ustan	> confer, compon, contain, time, select, level
77 vnos	> contain, confer, select, compon, step, time
78 podpr	> carbonel, rang, select, perform, contain, step
79 sprem	> carbonel, rang, perform, step, linguist, select
80 štud	> carbonel, rang, perform, linguist, segment, statist
81 pres	> carbonel, rang, perform, linguist, statist, peac
82 mnog	> carbonel, rang, perform, linguist, statist, approach
83 razis	> rang, carbonel, perform, linguist, statist, approach
84 povs	> perform, rang, linguist, carbonel, statist, approach
85 kjer	> linguist, perform, statist, rang, approach, carbonel
86 svet	> statist, linguist, approach, perform, consist, rang
87 prič	> approach, statist, consist, linguist, potenti, perform
88 razr	> consist, approach, potenti, statist, function, linguist
89 izvir	> potenti, consist, function, approach, posit, statist
90 rezul	> function, potenti, posit, consist, addit, approach
91 namen	> posit, function, addit, potenti, phase, consist
92 ciljn	> addit, posit, phase, function, largest, potenti
93 večin	> phase, addit, largest, posit, deal, function
94 zagot	> largest, phase, deal, addit, comput, posit
95 jožef	> deal, largest, comput, phase, rank, addit
96 stef	> comput, deal, rank, largest, condit, phase
97 znanj	> rank, comput, condit, deal, arab, largest
98 velj	> condit, rank, arab, comput, deal, largest
99 polj	> arab, condit, rank, comput, deal, jaim
100 okvir	> arab, condit, rank, jaim, comput, corpora

Tabela 5.7: Najverjetnejši prevodi za najpogostejših 100 besed slov. besedila

5.4 Ocena uspešnosti sistema

Preden smo izvedli poskuse, smo pričakovali, da bodo rezultati boljši, kot se je dejansko izkazalo. Do slabših rezultatov je prišlo predvsem zato, ker si besedili nista podobni v

dovoljšnji meri. Druga dejavnika, ki vplivata na neidealne rezultate, sta nepopolna seznama mašil (v seznamih manjka še ogromno besed) ter ne-optimalna funkcija korenjenja. Povprečna uspešnost u_1 = 2,3 % in u_5 = 10,9 % sicer niti ni tako slaba, če upoštevamo dejstvo, da prevodi temeljijo le na podlagi podanih besedil oziroma naborov besed v teh besedilih.

metoda	smer prevajanja	u_1	u_5
pos. besede	ang -> si	**3 %**	**11 %**
	si -> ang	**2 %**	**12 %**
bigrami	ang -> si	2 %	12 %
	si -> ang	3 %	12 %
trigrami	ang -> si	2 %	8 %
	si -> ang	1 %	9 %
pos. besede + metoda upoštevanja konteksta	ang -> si	**4 %**	**13 %**
	si -> ang	2 %	12 %
bigram + metoda upoštevanja konteksta	ang -> si	2 %	8 %
	si -> ang	3 %	12 %
trigram + metoda upoštevanja konteksta	ang -> si	2 %	12 %
	si -> ang	2 %	10 %

Tabela 5.8: Uspešnosti prevajanja u1 in u5

V tabeli 5.8 so predstavljene uspešnosti u_1 in u_5. Kot lahko vidimo, so bile v povprečju metode posameznih besed bolj uspešne od metod bi- in trigramov. Prav tako lahko vidimo, da je metoda upoštevanja konteksta nekoliko doprinesla k večji uspešnosti prevodov. Najbolj uspešna metoda je tako bila metoda posameznih besed v kombinaciji z metodo upoštevanja konteksta pri prevajanju iz angleščine v slovenščino. Pričakovano slabše pa so se v povprečju odrezale metode s trigrami. Vendar pa so razlike v uspešnosti metod zelo majhne - v kakšnem drugem paru besedil bi številke bile nekoliko drugačne.

5.5 Ovrednotenje rezultatov in ideje za izboljšave

Pri rezultatih moramo upoštevati dejstvo, da smo poskus izvajali na relativno kratkih besedilih. Če bi bili besedili daljši, bi bili tudi rezultati boljši. Druga, podobna možnost za izboljšavo je, da bi namesto dveh podobnih si besedil imeli več podobnih si besedil, ki bi jih sestavili v dve (za vsak jezik po eno) zelo dolgi besedili. V teoriji bi nas velik obseg besedil privedel do boljših rezultatov - s tem bi se relativne frekvence besed nekako ustalile oz. postale naravne.

Kaj nam je še povzročalo težave in kako bi jih lahko odpravili? Besed, ki so značilne le za eno od besedil, seveda ne moremo prevesti. Na takšno težavo naletimo npr. pri imenih, kot so

Univerza Ljubljana, *John Meadow*, *Inštitut Jožef Štefan* itn. Pomembno je omeniti tudi, da imamo v besedilu pogosto tudi dodatne elemente, kot so viri, slike, tuje besede ali celo povzetki v drugem jeziku itn., kar vse slabo vpliva na uspešnost prevajanja. Če bi v tem kontekstu želeli doseči boljše rezultate, bi morali omenjene besede pred prevajanjem izločiti.

Izboljšave bi lahko dosegli tudi pri metodi upoštevanja konteksta tako, da bi za besedilo bolje nastavili parametre v vmesniku aplikacije. Predstavljeni rezultati so namreč dobljeni na podlagi privzetih atributov, bilo pa bi bolje, če bi imeli sistem samodejne izbire najbolj ustreznih parametrov.

6 Zaključek

Namen diplomske naloge je bil preveriti, ali je na podlagi značilk naravnih jezikov možno prevajati med dvema naravnima jezikoma ter do kolikšne mere. V ta namen smo razvili aplikacijo, ki zna na omenjen način narediti pare besed med besedili različnih jezikov. Rezultati naše analize so pokazale, da se nekatere besede, predvsem bolj pogoste v jeziku, prevajajo uspešno, ostale besede pa manj. Rezultate uspešnosti si lahko vsak posameznik razlaga po svoje, dejstvo pa je, da tak pristop še ni dovolj dober, da bi ga lahko uporabljali v resne namene. Pot do tega bi bila še dolga, če ne celo nemogoča. Za dosego boljših rezultatov bi bilo aplikacijo potrebno nadgraditi; predvsem izpopolniti sezname mašil in implementirati boljše algoritme za korenjenje. Kot pomoč pri prevajanju bi bila možna tudi implementacija novih metod iskanja prevodov, kot npr. iskanje prevodov na podlagi dolžine besed (pogostejše besede so ponavadi krajše). Vmesnik bi lahko tudi razširili tako, da bi bilo možno vnesti več iskalnih parametrov. Takšnih idej je še kar nekaj, a je v okviru naše naloge zanje žal zmanjkalo časa.

Ni presenetljivo, da aplikacija strojnega prevajanja ne prevaja tako dobro, kot bi znal to prevajalec. Strojno prevajanje vsebuje kompleksne operacije, kjer je potrebno upoštevati razlike med naravnimi jeziki. Prevajanje je nekaj, kar umetna inteligenca najbrž nikoli ne bo povsem obvladala; težko je namreč zajeti vsako specifičnost naravnega jezika. Tako kljub velikemu napredku področja v zadnjih desetletjih idealnih rezultatov še ne moremo pričakovati. To nam preprečujejo prevelike kulturne razlike in s tem posledično prevelike razlike med jeziki samimi. Vsak jezik ima svoje, unikatne lastnosti. Vsak, ki od strojev pričakuje enako dobre prevode kot od profesionalnega prevajalca, verjetno pričakuje preveč; pa tudi če prevajamo s še tako uveljavljenimi sistemi, kot je npr. Babelfish [5].

Kljub temu to ne pomeni, da strojno prevajanje ni uporabno, saj ima v praksi svojo uporabno vrednost.

Literatura

[1] Automatic Language Processing Advisory Committee, National Academy of Sciences, National Research Council, *Language and machines: computers in translation and linguistics*, Washington D.C.: National Academy of Sciences, 1966, str.19

[2] William John Hutchins, *Early years in machine translation: memoirs and biographies of pioneers*, Amsterdam: John Benjamins Publishing Co, September 2000, pogl. 2.

[3] M.A. Khan, *Cataloguing In Library Science*, New Delhi: Sarup & Sons, 1997, str. 181.

[4] D. Maxwell, K. Schubert, T. Witkam, *New directions in machine translation*, Budapest: Foris Publications, 1988 Aug., str.9

[5] S. Yates, »Scaling the Tower of Babel Fish: An Analysis of the Machine Translation of Legal Information«, *Law Library Journal*, Vol. 98, str. 481, 2006.

[6] Machine Translation: History and general principles, dostopno na: http://www.hutchinsweb.me.uk/EncLangLing-1994.pdf

[7] (2008, Dec.) Moses - statistical machine translation system, dostopno na: http://www.statmt.org/moses/?n=Moses.Background

[8] Context-Based Machine Translation, dostopno na: http://www.meaningful.com/press/MM%20-%20Context%20Based%20MT%20-%20AMTA%202006%20final.pdf

[9] Računalniške tehnologije za prevajanje, dostopno na: http://www2.arnes.si/~svinta/ui.rtf

[10] (2006, Feb.) Machine Translation Techniques, dostopno na: http://www.globalsecurity.org/intell/systems/mt-techniques.htm

[11] (2010, Apr.) Wikipedia, Machine Translation, dostopno na: http://en.wikipedia.org/wiki/Machine_translation

[12] Phrase Based Machine Translation, dostopno na: http://ltrc.iiit.ac.in/winterschool08/presentations/sivajib/winter_school.ppt

[13] Example based machine translation, dostopno na: http://personalpages.manchester.ac.uk/staff/harold.somers/EBMT%20MT%20marathon.ppt

[14] (2010) Systran, What Is Machine Translation?, dostopno na: http://www.systran.co.uk/systran/corporate-profile/translation-technology/what-is-machine-translation

[15] (2010, Jan.) Wikipedia, Interlingual Machine Translation, dostopno na: http://en.wikipedia.org/wiki/Interlingual_machine_translation

[16] (2010, Jan.) Wikipedia, Transfer-based Machine Translation, dostopno na:
http://en.wikipedia.org/wiki/Transfer-based_machine_translation

[17] (2010, Apr.) Wikipedia, Statistical Machine Translatin, dostopno na:
http://en.wikipedia.org/wiki/Statistical_machine_translation

[18] (2010, Mar.) Wikipedia, Example-based Machine Translation, dostopno na:
http://en.wikipedia.org/wiki/Example-based_machine_translation

[19] (2010, Apr.) Wikipedia, Systran, dostopno na:
http://en.wikipedia.org/wiki/SYSTRAN

[20] (2010, Mar.) Wikipedia, Stoppwort, dostopno na:
http://de.wikipedia.org/wiki/Stoppwort

[21] (2010, Mar.) Wikipedia, Stemming, dostopno na:
http://en.wikipedia.org/wiki/Stemming

[22] (2010, May) Wikipedia, Outlier, dostopno na:
http://en.wikipedia.org/wiki/Outlier

[23] (2000, Jan.) Prazne besede slovenskega jezika, dostopno na:
http://nl.ijs.si/et/project/GNUsl/Stop

[24] (2000, Sep.) Slovenian Stemmer, dostopno na:
http://snowball.tartarus.org/archives/snowball-discuss/0670.html

[25](2010) Eclipse, dostopno na:
http://www.eclipse.org/

[26] (1970) Snowball – download, dostopno na:
http://snowball.tartarus.org/download.php

[27] (2010) kea – algorithm, dostopno na:
http://code.google.com/p/kea-algorithm/downloads/list

[28] (2010) culturitalia, dostopno na:
http://culturitalia.uibk.ac.at/stop_ita.htm