UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

JOŽE KADIVEC

Prilagoditev statističnega strojnega prevajalnika za specifično domeno v slovenskem jeziku

MAGISTRSKO DELO

Ljubljana, 2016

UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

JOŽE KADIVEC

Prilagoditev statističnega strojnega prevajalnika za specifično domeno v slovenskem jeziku

MAGISTRSKO DELO

Mentor: izr. prof. dr. Marko Robnik Šikonja
Somentor: prof. dr. Špela Vintar

Ljubljana, 2016

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Večna pot 113
1000 Ljubljana, Slovenija
telefon: 01 47 98 100
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Številka: 141-MAG-ISO/2016
Datum: 29. 02. 2016

Jože KADIVEC, univ. dipl. org. inf.

Ljubljana

Fakulteta za računalništvo in informatiko Univerze v Ljubljani izdaja naslednjo magistrsko nalogo

Naslov naloge: **Prilagoditev statističnega strojnega prevajalnika za specifično domeno v slovenskem jeziku**
Domain specific adaptation of a statistical machine translation engine in Slovene language

Tematika naloge:
Zaradi velikih količin besedil, ki nastajajo v različnih jezikih, se pojavlja potreba po hitrejšem, po možnosti avtomatskem prevajanju. Sistemi za strojno prevajanje vsebujejo več analitičnih komponent in jezikovnih virov, s pomočjo katerih analizirajo izvorno besedilo in ga brez sodelovanja ljudi prevedejo v ciljno besedilo. V zadnjem času se je z dostopnostjo obsežnih vzporednih večjezičnih korpusov težišče raziskav preneslo na statistično strojno prevajanje. Statistični strojni prevajalniki preiskujejo prostor mogočih prevodov in izberejo najverjetnejši prevod na podlagi verjetnostnih modelov jezika in modelov zvestobe prevoda. Za uspešnost prevajalnika se je torej potrebno osredotočiti predvsem na ti dve komponenti. Eden najuspešnejših odprtokodnih statističnih strojnih prevajalnikov je sistem Moses, ki ga želimo prilagoditi splošnemu slovenskemu jeziku in enemu od jezikovno specifičnih podpodročij.
Kandidat naj preuči področja strojnega prevajanja, obstoječe raziskave strojnih prevajalnikov za slovenski jezik in prilagajanje strojnega prevajanja za določeno področje. Za uporabo v sistemu Moses naj poišče in prilagodi obstoječe splošne korpuse za slovenski jezik kot osnovo za gradnjo primerjalnega jezikovnega modela. Za področje farmacevtskih besedil naj poišče obstoječe angleško-slovenske prevode in druge jezikovne vire, ki bodo služili kot učna množica za učenje strojnega prevajalnika. V okolju statističnega strojnega prevajalnega sistema Moses naj ovrednoti vpliv različnih jezikovnih virov na kakovost dobljenega strojnega prevoda za področje farmacije. Ovrednotenje naj temelji predvsem na avtomatskih merah kakovosti prevodov, končni rezultat pa naj ocenijo tudi potencialni uporabniki sistema. Pričakovan prispevek naloge je analiza vpliva različnih virov pri specializaciji strojnega prevajalnika, vzpostavljen splošni jezikovni model za slovenski jezik in vzpostavljen statistični strojni prevajalnik za področje farmacevtskih besedil.

Mentor:
izr. prof. dr. Marko Robnik Šikonja

Somentorica:
prof. dr. Špela Vintar

Dekan:
prof. dr. Nikolaj Zimic

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

I Z J A V A O A V T O R S T V U

magistrskega dela

Spodaj podpisani Jože Kadivec, z vpisno številko 63020364, sem avtor magistrskega dela z naslovom

Prilagoditev statističnega strojnega prevajalnika za specifično domeno v slovenskem jeziku

S svojim podpisom zagotavljam, da:

- sem magistrsko delo izdelal samostojno pod vodstvom mentorja

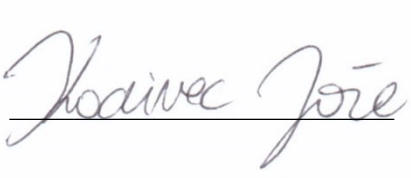
  izr. prof. dr. Marka Robnika Šikonje

  in somentorstvom

  prof. dr. Špele Vintar

- so elektronska oblika magistrskega dela, naslova (slov., angl.), povzetka (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko magistrskega dela in

- soglašam z javno objavo elektronske oblike magistrskega dela v zbirki »Dela FRI«.

V Ljubljani, dne 22. 8. 2016 Podpis avtorja/-ice: 

Zahvala

Zahvaljujem se mentorju izr. prof. dr. Marku Robniku Šikonji za pomoč in usmerjanje pri izdelavi tega magistrskega dela ter koordiniranje s predstavniki sodelujočih institucij, somentorici prof. dr. Špeli Vintar za vse jezikoslovne nasvete, dr. Andreji Čufar (JAZMP), Liljani Čop in Tomiju Laptošu (Lekarna UKC Ljubljana) za nasvete in pomoč pri pridobivanju farmacevtskih zbirk ter pregledu strojnih prevodov, Urošu Palminu za pomoč pri reševanju sistemskih zagat, Janezu Kadivcu za pomoč pri statističnih analizah, dr. Simonu Kreku in dr. Mihi Grčarju za pomoč z označevalnikom besedil, Manici Posega za lektoriranje ter vsem prevajalcem za ocenjevanje strojnih prevodov.

Za pomoč, potrpežljivost in podporo pri izdelavi magistrskega dela se zahvaljujem tudi ženi Mojci in družini.

Kazalo

Seznam uporabljenih kratic

Kratica	Angleško	Slovensko
ARPA	Advanced Research Projects Agency	Agencija za napredne raziskovalne projekte
BLEU	Bilingual Evaluation Understudy	Algoritem za samodejno ocenjevanje strojnih prevodov
CAT	Computer Aided Translation	Računalniško podprto prevajanje
CBZ	Central Medicine Database	Centralna baza zdravil
DGT	Directorate General for Translation	Generalni direktorat za prevajanje
EMA	European Medicines Agency	Evropska agencija za zdravila
EU	European Union	Evropska unija
EURAMIS	European advanced multilingual information system	Evropski napredni večjezikovni informacijski sistem
GIZA++	Segment alignment program name	Ime programa za poravnavanje segmentov
IRSTLM	Istituto per la Ricerca Scientifica e Tecnologica Language Model	Jezikovni model Inštituta za znanstvene in tehnološke raziskave v Povu
JAZMP	Public Agency of the Republic of Slovenia for Medicinal Products and Medical Devices	Javna agencija za zdravila in medicinske pripomočke
KENLM	Kenneth Language Model	Jezikovni model Kennetha Heafielda
MERT	Minimum Error Rate Training	Algoritem za uglaševanje z minimalno stopnjo napak
MIRA	Margin Infused Relaxed Algorithm	Algoritem za uglaševanje strojnih prevodov z mejno stopnjo
OCR	Optical Character Recognition	Optično prepoznavanje znakov
OPUS	Open Paralel Corpus	Odprtokodni korpus paralelnih besedil
OTM	Open Translation Manager	Ime odprtokodnega programa za prevajanje
PDF	Portable Document Format	Format datoteke, neodvisen od računalniškega okolja
PubMed	Public Medical documents	Javni medicinski dokumenti
SMT	Statistical Machine translation	Statistično strojno prevajanje (SSP)
SRILM	Stanford Research Institute Language Model	Jezikovni model raziskovalnega inštituta v Stanfordu
TAUS	Translation Automation User Society	Združenje uporabnikov za avtomatizacijo prevajanja
UKC	University Clinical Center	Univerzitetni klinični center

Povzetek

Strojno prevajanje, še posebej statistično strojno prevajanje, se je v zadnjih letih zelo razširilo, zahvaljujoč predvsem vse večjemu številu večjezičnih jezikovnih virov. Večina javno dostopnih strojnih prevajalnikov nam omogoča, da dobimo osnovno razumevanje vsebine v tujem jeziku, medtem ko ti niso dovolj natančni za specifične domene. Študije za nekatere tuje jezike kažejo izboljšanje strojnega prevajalnika, če je za učenje uporabljena domenska zbirka. Za slovenski jezik podobna študija še ni izvedena, kar predstavlja motivacijo za naše delo. Dodatno motivacijo predstavlja neobstoj javno dostopnega splošnega jezikovnega modela za slovenski jezik.

V magistrskem delu se osredotočimo na prilagajanje statističnega strojnega prevajalnika za specifično domeno v slovenskem jeziku. Opišemo različne pristope k prilagajanju za specifično domeno. Vzpostavimo sistem za strojno prevajanje Moses in poiščemo ter prilagodimo obstoječe splošne korpuse za slovenski jezik kot osnovo za gradnjo primerjalnega jezikovnega modela. Iz označenega in neoznačenega korpusa slovenskega jezika ccGigafida izdelamo jezikovni model slovenskega jezika. Za področje farmacevtskih besedil poiščemo in prilagodimo obstoječe angleško-slovenske prevode in druge jezikovne vire, ki služijo kot učna množica za učenje strojnega prevajalnika. Ovrednotimo vpliv različnih jezikovnih virov na kakovost dobljenega strojnega prevoda za področje farmacije. Ovrednotenje izvedemo samodejno z metriko BLEU, nekatere testne prevode pa ročno ocenijo tudi strokovnjaki in potencialni uporabniki sistema. Analiza ocen pokaže, da testni prevodi, prevedeni z domenskim modelom, dosežejo precej boljše ocene od prevodov, prevedenih s splošnim modelom, medtem ko večji, kombinirani model, ne prinese boljših ocen od manjšega domenskega modela. Analiza ročnih ocen berljivosti in ustreznosti pokaže, da prevodi, ki dosežejo visoko oceno BLEU, lahko dosežejo nižje ocene berljivosti ali ustreznosti od testnih prevodov, ki so sicer dosegli nižjo oceno BLEU. Ugotovimo tudi, da uporaba strokovnega slovarja doprinese 1 oceno BLEU in zagotovilo, da je uporabljeno želeno izrazoslovje.

Ključne besede: strojno prevajanje, statistično strojno prevajanje, prilagajanje strojnega prevajalnika za specifično domeno, prilagajanje statističnega strojnega prevajanja za področje farmacije, faktorski model, Moses, model na osnovi besednih zvez, Cohenova kappa, Fleissova kappa, strinjanje ocenjevalcev

Abstract

Machine translation, especially statistical machine translation gained a lot of interest in recent years, mainly thanks to the increase of publicly available multilingual language resources. In terms of obtaining the basic understanding of the target language text, the majority of free machine translation systems give us satisfactory results but are not accurate enough for specific domain texts. For some foreign languages, research shows increases in the quality of the machine translation if trained with the in-domain data. Such research has not yet been conducted for the Slovenian language which presents the motivation for our research. Additional motivation is the nonexistence of a publicly available language model for the Slovenian language.

This master thesis focuses on a statistical machine translation system adaptation for a specific domain in the Slovenian language. Various approaches for the adaptation to a specific domain are described. We set up the Moses machine translation system framework and acquire and adapt existing general corpora for the Slovenian language as a basis for building a comparative linguistic model. Annotated and non-annotated Slovenian corpus, ccGigafida, is used to create a linguistic model of the Slovenian language. For the pharmaceutical domain, existing English-Slovenian translations and other linguistic resources have been found and adapted to serve as a learning base for the machine translation system. We evaluate the impact of various linguistic resources on the quality of machine translation for the pharmaceutical domain. The evaluation is conducted automatically using the BLEU metrics. In addition, some test translations are manually evaluated by experts and potential system users. The analysis shows that test translations, translated with the domain model, achieve better results than translations that are generated using the out-of-domain model. Surprisingly, bigger, combined model, does not achieve better results than the smaller domain model. The manual analysis of the resulting fluency and adequacy shows that translations that achieve a high BLEU grade can achieve lower fluency or adequacy grades than the test translations that otherwise achieved a lower BLEU grade. The experiment with the addition of the domain-based dictionary to the in-domain translation model shows a gain of 1 BLEU grade and assures the use of the desired terminology.

Keywords: machine translation, statistical machine translation, statistical machine translation system adaptation for specific domain, statistical machine translation system adaptation for pharmaceutical domain, factor model, Moses phrase based model, Cohen’s kappa, Fleiss kappa, agreement of ratters

1. Uvod

Prevajanje sega že daleč v preteklost do antičnih časov, ko se med seboj srečajo pripadniki različnih narodov in začutijo potrebo po medsebojnem komuniciranju. Narodom na različnih delih sveta omogoča napredek in razvoj, saj s prevodi knjig in besedil lahko pridejo do novega znanja, ki jim prej še ni znano. Prevajanje danes definiramo kot »postopek prenosa pomena ter jezikovnih in kulturnih značilnosti iz izvirnega v ciljni jezik« [54].

S časom se količina znanja povečuje, še posebej pa se količina informacij začne povečevati s pojavom prvih računalnikov. Tedaj tudi prevajanje dobi računalniško podporo s prvimi CAT orodji. Kaj kmalu pa strokovnjaki vidijo možnosti, da bi računalnik sam prevajal namesto človeka. Začetnih idej ne morejo dodobra preizkusiti, predvsem zaradi premalo zmogljivih računalnikov, vse do preloma tisočletja, ko pride do razmaha strojnega prevajanja.

Adam Lopez definira strojno prevajanje kot »avtomatsko prevajanje iz enega naravnega jezika v drugega z uporabo računalnika« [30]. Strojno prevajanje, morda še posebej statistično strojno prevajanje, ima pomemben vpliv ne samo na prevajanje, temveč tudi na medsebojno komuniciranje in druga področja. V prvi vrsti pomaga prevajalcem pri naslednjem:

- prevajanju osnutka, ki služi kot pomoč pri prevajanju,
- pripravi in prevajanju pomnilnikov prevodov,
- izdelavi in prevajanju terminoloških zbirk in slovarjev [11].

Z javno dostopnimi strojnimi prevajalniki lahko ljudje razumemo vsebino dokumentov in spletnih strani, ki so jih morali v preteklosti prevajalci predhodno prevesti. Na področju medsebojnega komuniciranja strojni prevajalniki nastopajo tudi v govornih prevajalnikih, kjer računalnik govor pretvori v besedilo, ga strojno prevede in sintetizira govor v ciljnem jeziku, ter na ta način podpirajo medsebojno komuniciranje med ljudmi.

Lahko rečemo, da je glavna prednost strojnega prevajanja njegova hitrost. Ob predpostavki, da imamo dobre učne zbirke, lahko pričakujemo tudi dokaj natančne prevode. V eksploziji informacij novejše dobe je za vse narode pomembno, da lahko sledijo toku globalnega dogajanja. Tudi za Slovence.

Za slovenski jezik je na voljo nekaj javno dostopnih splošnih strojnih prevajalnikov. Za prevajalnike Amebis, Bing in Google Translate je narejena raziskava [49], ki pokaže njihovo

uporabnost za prevajanje splošnih besedil zgolj do določene mere in manjšo uporabnost za specifična strokovna besedila. Ista raziskava pokaže, da boljše ocene dobijo statistični strojni prevajalniki. Kakovost statističnega strojnega prevajanja v slovenski jezik je mogoče izboljšati z uglaševanjem ustreznih, s prevajanjem povezanih parametrov, vendar je kakovost prevoda še vedno odvisna od kakovosti učnih podatkov [11].

Na tem mestu se vprašamo, kako na kakovost prevoda vpliva učna zbirka, ki je bližje ciljnemu področju? Koehn in Schroeder [28] izvedeta vrsto poskusov z namenom prilagajanja statističnega strojnega prevajanja za specifično domeno za prevajanje iz angleškega v francoski jezik. Kot osnovo vzameta veliko splošno zbirko prevodov (*angl. out-of-domain*) in nato manjšo zbirko za specifično domeno (*angl. in-domain*), s katero želita optimizirati učinkovitost prevajanja za specifično domeno in primerjati rezultate. Za slovenski jezik podobna študija še ni izvedena, kar predstavlja motivacijo za naše delo. Dodatno motivacijo predstavlja neobstoj javno dostopnega splošnega jezikovnega modela za slovenski jezik in za sistem Moses ter praktična vrednost prilagoditve za področje farmacije.

Naš namen je v sodelovanju s farmacevti Lekarne UKC Ljubljana in Javne agencije za zdravila in medicinske pripomočke (JAZMP) izdelati prototip prevajalnega sistema, ki bo poenostavil in olajšal prevajanje farmacevtskih besedil. Bolnišnične lekarne morajo namreč občasno same prevajati navodila za uporabo specifičnih zdravil, ki jih naročijo pri proizvajalcih zdravil znotraj EU, vendar zaradi njihove ozke namembnosti nimajo slovenskih navodil za uporabo.

Ker je za uspeh statističnega strojnega prevajanja ključnega pomena ustrezna količina kakovostnih podatkov, načrtujemo s pomočjo sodelujočih farmacevtov zbrati javno dostopne dvojezične zbirke podatkov, slovarje s področja farmacije, kemije, medicine in biokemije, ki služijo za izboljšanje prevajalnega modela. Za gradnjo jezikovnega modela želimo poleg enojezičnih dokumentov, ki jih priskrbita Lekarna UKC in JAZMP, uporabiti tudi kakovostne javno dostopne korpuse slovenskega jezika, kot je denimo ccGigafida [29].

Za prevajalni sistem, ki ga nameravamo izdelati v sistemu Moses, načrtujemo na koncu izvesti tudi samodejno in ročno oceno kakovosti.

1.1. Predstavitev strojnega prevajanja

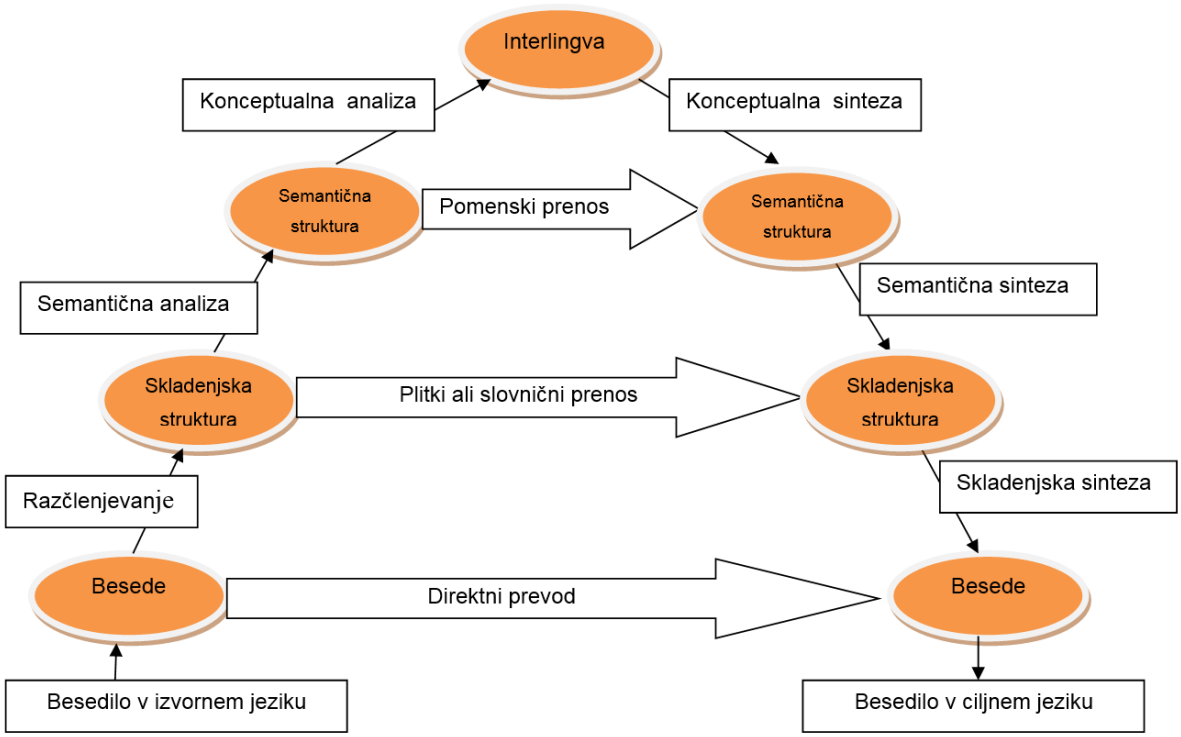
Zanimanje za strojno prevajanje je skoraj tako staro kot elektronski računalnik. Prve ideje za današnji razvoj leta 1949 zapiše Warren Weaver, le nekaj let potem, ko je predstavljen prvi računalnik ENIAC [30]. Prvi ročno zgrajeni model na osnovi slovnice doseže le omejen uspeh. Področje strojnega prevajanja doseže preobrazbo, ko IBM-ovi raziskovalci od kanadskega parlamenta dobijo veliko količino angleško-francoskih vzporednih besedil, ki jim omogočijo statistično analizo prevodov besed in zaporedij besed ter zgraditi verjetnostni model strojnega prevajanja [20]. Z rastjo količine dostopnih vzporednih besedil in pocenitvijo računalniških kapacitet se zanimanje za statistično strojno prevajanje še poveča. Namesto prevajanja po besedah, je predstavljen model prevajanja po besednih zvezah, ker imajo majhne skupine besed značilnejše prevode. Franz Och uporabi ta model pri razvoju danes pogosto uporabljenega spletnega prevajalnika Google Translate [15].

Strokovnjaki, ki delajo na področju strojnega prevajanja, razvijajo različne načine strojnega prevajanja, ki jih lahko v glavnem razdelimo v tri skupine:

- prevajanje na osnovi pravil,
- statistično strojno prevajanje,
- hibridni sistemi strojnega prevajanja [22].

Prevajanje na osnovi pravil

Sistemi strojnega prevajanja na osnovi pravil so prvi komercialni sistemi za strojno prevajanje in temeljijo na jezikovnih pravilih, ki omogočajo postavljanje besed na različna mesta, da imajo ustrezen pomen glede na kontekst. Ta tehnologija uporablja velike zbirke jezikovnih pravil v treh različnih fazah – analizi, prenosu in generiranju prevoda. Pravila razvijajo strokovnjaki na področju naravnih jezikov ter programerji, ki pravila razumejo in znajo ustvariti preslikave med dvema jezikoma [22].



Slika 1.1: Vauquoiev trikotnik za predstavitev strojnega prevajanja na osnovi pravil [7]

Strojno prevajanje na osnovi pravil lahko v grobem razdelimo na tri načine:

- direktni prevod,
- transferno strojno prevajanje,
- interlingva pristop.

Splošen način za predstavitev teh treh pristopov je Vauquoiev trikotnik (Slika 1.1.), ki prikazuje, kako se povečuje globina potrebne analize, ko se premikamo od direktnega pristopa do transfernega pristopa in na vrhu do pristopa z vmesnim jezikom.

Statistično strojno prevajanje

»Korpusni pristopi in z njimi statistično strojno prevajanje (SSP) so se pojavili v 90. letih 20. stoletja kot izziv prej dominantnemu pristopu na osnovi pravil. Od takrat je SSP postalo osrednje področje raziskav številnih raziskovalnih skupin.« [4]

Sistem je popolnoma neodvisen od jezikovnega znanja in temelji na tehniki učenja strojnega prevajalnika z obsežnimi zbirkami eno- in dvojezičnih podatkov oz. korpusov ter verjetnosti. Izračun verjetnosti je odvisen od dveh modelov. Prvi je t. i. prevodni model, s katerim sistem

izračuna verjetnost, da se bodo besede v S (izvirnem jeziku) zamenjale z besedami v T (ciljnem jeziku). Drugi, t. i. jezikovni model, pa predstavlja verjetnost, da bodo te besede v ciljnem jeziku ustrezno uporabljene.

Za prevodni model algoritem tako potrebuje vzporedna besedila v obeh jezikih, iz takšnega poravnanega korpusa pa je mogoče za vsako besedo/večbesedno enoto izvirnega jezika izluščiti niz najverjetnejših prevodnih ustreznic. Ker pa sam prevodni model ne zagotavlja slovnično pravilnih stavkov ciljnega jezika, je potreben še jezikovni model, ki se prav tako gradi iz obsežnih zbirk besedil [4].

Metode statističnega strojnega prevajanja uporabimo v tem magistrskem delu in jih podrobneje opišemo v poglavju 2.

Hibridni sistemi

Angleška Wikipedia [54] navaja, da hibridni sistemi lahko zajamejo dobre strani statistične metode in metode na osnovi pravil. Hibridni sistemi se med seboj ločijo glede na to, kako uporabijo katero metodo:

- Najprej metoda na osnovi pravil nato statistična – sistem najprej prevede z metodo na osnovi pravil, nato s statistično metodo naredi popravke in prilagoditve v prevodu.

- Statistična metoda, ki jo uglašena z metodo na osnovi pravil – z metodo na osnovi pravil sistem pripravi besedilo, da je prevajanje z uporabo statistične metode učinkovitejše. Metodo na osnovi pravil ponovno uporabi za normalizacijo prevoda.

1.2. Opis problema farmacevtskih besedil

Za besedila s specifičnih področij je splošno znano, da uporabljajo strokovno izrazoslovje, poseben slog, pogosto pa imajo celo zakonsko predpisano obliko. Vsa ta dejstva moramo upoštevati pri prevajanju tovrstnih besedil. Tudi besedila s področja farmacije, medicine in biotehnologije spadajo v skupino visoko strokovnih besedil, za katera je pri prevajanju treba upoštevati več določil in pravil.

V prvi vrsti je za farmacevtska besedila značilna uporaba strokovnih farmacevtskih in medicinskih izrazov ter posebnega tehničnega sloga, ki jim je pri prevajanju treba slediti. Za

zagotavljanje kakovosti prevoda moramo uporabljati predpisane slovarje in obstoječe pregledane prevode.

Ker je od prevoda farmacevtskega ali medicinskega besedila včasih odvisno tudi življenje, naročniki pogosto pri prevajanju tovrstnih besedil zahtevajo poseben pristop k prevajanju, ki zahteva, da se besedilo, ki je prevedeno in pregledano s strani strokovnjaka, prevede nazaj v izvirni jezik s strani tretjega strokovnjaka. Temu pravimo povratni prevod [42]. Povratno prevajanje je časovno zamudno in predstavlja dvakratni strošek pri skupni ceni prevajanja.

Preden lahko proizvajalci farmacevtskih in medicinskih pripomočkov začnejo s trženjem svojih izdelkov, morajo s strani Evropske agencije za zdravila (EMA) dobiti dovoljenje. Za pridobitev dovoljenja morajo v prvi fazi med drugim v odobritev poslati tudi povzetek značilnosti medicinskega izdelka (SmPC), nalepke in navodila za uporabo. Predstavniki agencije poslano vsebino dobro pregledajo, za kar imajo na voljo 110 dni, preden izdajo izjavo o kakovosti informacij za izdelek. Proizvajalci nato informacije izboljšajo in pošljejo v vnovično presojo. Ko dobijo odobrene angleške izvorne dokumente, imajo na voljo malo časa, da jih prevedejo v vse jezike držav, v katerih želijo tržiti izdelke [33].

S kratkimi roki za prevod navodil se soočajo tudi nekatere bolnišnične lekarne, ki morajo občasno same prevajati navodila za uporabo specifičnih zdravil, ki jih naročijo pri proizvajalcih zdravil znotraj EU. Zaradi njihove ozke namembnosti nimajo slovenskih navodil za uporabo, potrebujejo pa jih hitro, da lahko zdravilo dajo pacientu čim hitreje.

Glede na predstavljene izzive lahko prepoznamo potrebo po rešitvi, ki bi vpletenim omogočala hitro pomoč za probleme, ki sicer zahtevajo veliko človeškega časa.

Možen pristop k rešitvi je v smeri prevajanja z uporabo statističnega strojnega prevajanja, ki ga je mogoče relativno hitro naučiti prevajanja iz in v različne tuje jezike. Ko je ta enkrat naučen, lahko ponudi prevode dosti hitreje in predvsem ceneje kot strokovni prevajalci. O kakovosti in ostalih vidikih spregovorimo v nadaljevanju.

1.3. Predstavitev glavnih pristopov

Ta naloga obravnava pristope za prilagoditev strojnega prevajalnika za specifično domeno v slovenskem jeziku. Glede na posebnosti posameznih načinov, ki so na voljo za strojno prevajanje, se odločimo za pristop s statističnim strojnim prevajanjem, ker omogoča učenje na

podlagi že prevedenih dvojezičnih zbirk podatkov, te pa imamo bodisi že na voljo, bodisi jih izdelamo s poravnavo dokumentov v slovenskem in angleškem jeziku.

Z uporabo nekaterih javno dostopnih zbirk ter domensko specifičnih prilagodimo prevajalni sistem za prevajanje farmacevtskih besedil. Vsebino za učenje prevajalnega sistema črpamo iz zbirk Euparl [12], DGT [21], ccGigafida [29], zbirk, izdelanih iz dokumentov Evropske agencije za zdravila in Lekarne UKC v Ljubljani, javno dostopnih slovarjev s področja farmacije ter nekaterih manjših zbirk. Za eksperiment s transduktivnim prevajanjem uporabimo zbirko PubMed. Vpliv na kakovost prevodov farmacevtske domene raziščemo z različnimi pristopi prilagajanja jezikovnega in prevajalnega modela. Zanima nas, kako vpliva približevanje vsebine dvojezičnih zbirk ciljni domeni na kakovost prevoda. Tako izdelamo prevajalni model najprej iz splošnih zbirk in ga nato približujemo domeni z omejitvijo na zgolj farmacevtsko zbirko. Podoben pristop izberemo tudi za prilagajanje jezikovnega modela.

Za učenje prevajalnih in jezikovnih modelov, uglaševanje ter v končni fazi dekodiranje oz. prevajanje uporabimo sistem Moses [26]. Moses je zbirka odprtokodnih orodij za statistično strojno prevajanje. Sestavljen je iz vseh komponent, ki so potrebne za obdelavo podatkov in učenje prevajalnih in jezikovnih modelov. Nudi tudi orodja za optimizacijo modelov ter samodejno oceno prevodov z metriko BLEU (*angl. Bilingual Evaluation Understudy)*. Za boljši vpogled v kakovost izdelanih prevodov izvedemo tudi ročno oceno s pomočjo izkušenih prevajalcev in strokovnjakov.

1.4. Napoved vsebine po poglavjih

V prvem poglavju podamo motivacijo za delo, na splošno predstavimo strojno prevajanje, izpostavimo izzive, s katerimi se lahko srečamo pri prevajanju farmacevtskih besedil, ter navedemo glavne pristope, ki jih uporabimo v tem magistrskem delu. Vsebina v nadaljevanju je usmerjena na bolj specifična področja.

V drugem poglavju podrobneje predstavimo statistično strojno prevajanje, načine oz. modele statističnega prevajanja in pristope za evalvacijo oz. oceno prevodov, prevedenih z uporabo strojnega prevajanja. Predstavimo ročne in samodejne pristope.

V tretjem poglavju se poglobimo v možnosti za prilagajanje SSP za specifično domeno. Najprej osvežimo znanje o načinu delovanja SSP, ki obsega jezikovni in prevajalni model, v naslednjih podpoglavjih pa predstavimo pristope za prilagajanje posameznega modela za specifično

področje. Za boljše razumevanje pri vsakem modelu navedemo tudi nekaj primerov dobre prakse.

V četrtem poglavju uvodoma predstavimo problem prilagajanja SSP, v nadaljevanju pa še problem, s katerim se srečujejo v Lekarni UKC Ljubljana in Javni agenciji za zdravila in medicinske pripomočke (JAZMP) in ga je mogoče nasloviti s statističnim strojnim prevajalnikom, prilagojenim za farmacevtsko domeno. V drugem delu istega poglavja opišemo izzive pri zbiranju primernih zbirk ter predstavimo zbirke, ki jih uporabimo v naših eksperimentih.

Peto poglavje je namenjeno predstavitvi sistema Moses. Opišemo komponente tega odprtokodnega sistema ter glavne načine delovanja.

V šestem poglavju se posvetimo empiričnemu ovrednotenju strojnega prevajalnika. Predstavimo pripravo potrebnih zbirk, uporabljene programe, izvedbo prilagajanja prevajalnega modela ter izvedbo prilagajanja jezikovnega modela. Opišemo tudi uglaševanje uporabljenih parametrov.

V sedmem poglavju opišemo pristope k ocenjevanju testnih prevodov in dobljene rezultate. V začetku tega poglavja predstavimo samodejno in ročno ocenjevanje prevodov, v nadaljevanju pa rezultate prevajanja. Ob koncu sedmega poglavja naredimo še podrobnejšo analizo rezultatov.

V sklepnem poglavju povzamemo opravljeno delo in strnimo glavne rezultate empiričnega ovrednotenja.

2. Statistično strojno prevajanje

»Statistično strojno prevajanje je vrsta strojnega prevajanja, ki temelji na večji količini vzporednih besedil, iz katerih se s statističnimi algoritmi izračunavajo verjetnosti prevodne ekvivalence za posamezne jezikovne enote. Besedilo je prevedeno glede na verjetnostno porazdelitev. Na koncu je izbran prevod z najvišjo verjetnostjo, ta pa se običajno računa po posameznih povedih.« [49]

Pri statističnem strojnem prevajanju lahko modele obdelave besedila razdelimo na več vrst. Glavni modeli, ki se danes uporabljajo, so:

- modeli, ki temeljijo na besedah (*angl. word-based models),*
- modeli, ki temeljijo na besednih zvezah (*angl. phrase-based models*),
- modeli, ki temeljijo na drevesih (*angl. tree based models*) in
- faktorski modeli (*angl. factored models*) [11].

Model, ki ga glede na način obdelave lahko tudi uvrstimo med statistične modele, je model, ki temelji na principu globokih nevronskih mrež (*angl. neural machine translation models*).

Naštete modele predstavimo v nadaljevanju.

2.1. Model, ki temelji na besedah

Pri tem tipu prevajanja je osnovna prevodna enota beseda nekega naravnega jezika. Število besed v izhodiščni in ciljni povedi je običajno različno – zaradi sestavljenk, oblikoslovja in frazeologije. Razmerje med dolžinami prevedenih povedi se imenuje plodnost (*angl. fertility*). Ta nam pove, koliko besed v ciljnem jeziku proizvede vsaka beseda v izhodiščnem jeziku. Sistemi predpostavljajo, da med seboj ustrezajoči si leksemi pokrivajo isti pojem, resnica pa je pogosto drugačna. Tako se lahko slovensko besedo *kót* v španščino prevede z besedo *rincón* ali *esquina*; odvisno, ali gre za notranji ali zunanji kot [54].

Primer prevajalnega sistema za prevajanje po besedah je program GIZA++[37].

Ta preprosti način prevajanja ni ustrezen za prevajanje med jeziki z različno plodnostjo. Sicer je relativno preprosto ustvariti sistem strojnega prevajanja na osnovi besed, ki zna obravnavati visoko plodnost, saj ene besede ni težko prevesti z več besedami. Težava se pojavi v nasprotni smeri – prepoznavanju večbesednih enot in prevajanju le-teh z eno besedo.

Spodnji primer prikazuje pravilen prevod Googlovega Prevajalnika [15] iz slovenščine v angleščino ter napačnega iz angleščine v slovenščino, pri katerem prevajalnik kot povedek ni prepoznal fraznega oziroma sestavljenega glagola *call off = cancel = preklicati*, temveč le njegov del *call = poklicati*.

On je preklical poroko. **– He canceled the wedding.**

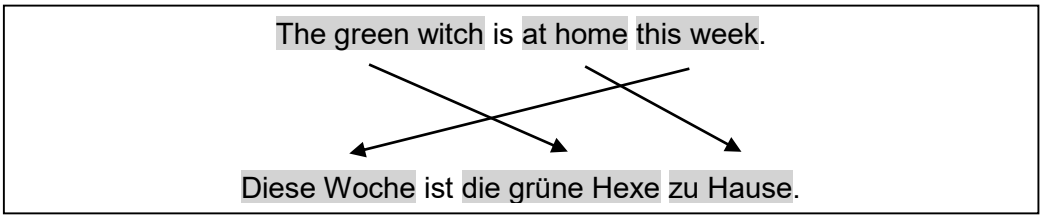
He called off the wedding. **– Poklical je off poroko.**

(Google Prevajalnik, 13. 4. 2016)

2.2. Model, ki temelji na besednih zvezah

Bolj kot po besedah se v zadnjem času prevaja po (različno dolgih) besednih nizih, in sicer s ciljem, da bi se zmanjšale omejitve prevajanja po besedah. Z »nizi« niso mišljeni stavki kot slovnične strukture, temveč nizi besed, ki jih v korpusu prepoznajo sistemi, ustvarjeni po statistični metodi [54].

Pri prevajanju po besednih zvezah kot osnovne prevodne enote uporabimo besedne zveze (zaporedje besed) kot tudi posamezne besede. Na sliki 2.1 lahko vidimo, da je včasih treba prevesti ali premakniti kot enote celotne besedne zveze [22].



Slika 2.1: Zamenjava vrstnega reda besednih zvez pri prevajanju iz angleščine v nemščino [22]

Ker je statistično strojno prevajanje po besednih zvezah zadnje čase razširjeno tudi pri strojnem prevajanju v slovenski jezik, ta model podrobneje opišemo.

Na trgu je sicer na voljo več modelov prevajanja po besednih zvezah, za potrebe predstavitve pa se omejimo na model Koehna in sod [27].

V tem modelu kot zgled za izračun verjetnosti, da računalnik izračuna prevod, vzamemo primer španske povedi *P (Maria no dió una bofetada a la bruja verde. | Mary did not slap the green witch.)* (Marija ni klofnila zelene čarovnice.).

Generativno gledano je postopek prevajanja razdeljen na tri dele. Najprej združimo angleški izvirnik v besedne zveze $\bar{e}_1$, $\bar{e}_2 \ldots$ $\bar{e}_I$. Nato prevedemo angleške besedne zveze $\bar{e}_i$ v španske besedne zveze f_i. Na koncu vsako od španskih besednih zvez (po potrebi) razporedimo v drugačen vrstni red. Verjetnostni model za prevajanje po besednih zvezah se opira na verjetnost prevoda in verjetnost prerazporeditve. Faktor Φ (f_i | $\bar{e}_i$) je verjetnost prevoda v špansko besedno zvezo f_i iz angleške besedne zveze $\bar{e}_i$. Prerazporeditev španskih besednih zvez se izvede na podlagi verjetnosti prerazporeditve d^0. Prerazporeditev (popačenje) se v statističnem strojnem prevajanju nanaša na dejstvo, da ima beseda drugačen (popačen) položaj v španskem stavku, kot ga je imela v angleškem stavku. Kot taka je mera za razdaljo med položaji besednih zvez v dveh jezikih. Verjetnost popačenja torej pomeni verjetnost, da sta dve zaporedni angleški besedni zvezi v španskem prevodu ločeni za določeno razdaljo (španskih besed).

Bolj formalno lahko popačenje zapišemo kot $d(a_i - b_{i-1})$, pri čemer je a_i začetni položaj tuje (španske) besedne zveze, ki je ustvarjena z i-to angleško besedno zvezo $\bar{e}_i$, in b_{i-1} je končni položaj tuje (španske) besedne zveze, ki je ustvarjena z $i - 1$ – to angleško besedno zvezo $\bar{e}_{i-1}$. Uporabimo lahko tudi preprosto verjetnost popačenja, pri kateri v popačenje vpeljemo konstanto α $d(a_i - b_{i-1}) = \alpha^{|ai - bi-1 - 1|}$. Ta model popačenja kaznuje velika popačenja tako, da daje nižje verjetnosti, če so popačenja (razdalje) večje.

Končni prevajalni model za strojno prevajanje na osnovi besednih zvez lahko opišemo s formulo:

$$P(F|E) = \prod_{i=1}^{I} \Phi(\overline{f_i}, \overline{e_i}) d(a_i - b_{i-1})$$

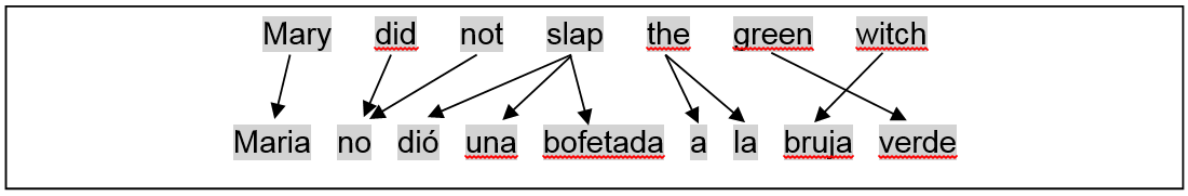
Če želimo uporabljati model na osnovi besednih zvez, potrebujemo še dve stvari. In sicer model za dekodiranje, ki bo v našem primeru znal povezati španske nize z angleškimi nizi v ozadju, ter model za učenje, da se prevajalnik nauči parametrov [27].

V praksi najprej poskrbimo za model za učenje, zato ga bomo tu predstavili najprej.

2.2.1. Model za učenje

Kako lahko prevajalnik naučimo preprost verjetnostni model prevajanja po besednih zvezah, ki je predstavljen v zgornji enačbi? Glavni niz parametrov, ki se jih je treba naučiti, je niz verjetnosti za prevajanje besedne zveze $\Phi(\overline{f_l}, \overline{e_l})$. Te parametre, kot tudi konstanto popačenja α, je mogoče nastaviti samo, če bi imeli na voljo veliko dvojezično učno zbirko, v kateri bi bil vsak španski stavek poravnan z angleškim stavkom, in bi tudi natančno vedeli, katera besedna zveza v španskem stavku je prevedena s katero besedno zvezo v angleškem stavku. Tako preslikavo imenujemo poravnava besednih zvez.

Žal pa v praksi nimamo velikih ročno označenih učnih zbirk s poravnanimi besednimi zvezami. Izkaže pa se, da lahko besedne zveze ekstrahiramo iz drugačne poravnave, ki se imenuje poravnava besed *(angl. word alignment)*. Poravnava besed se razlikuje od poravnave besednih zvez, ker natančno prikazuje, katera španska beseda se poravna s katero angleško besedo znotraj posamezne besedne zveze. Poravnavo besed lahko vizualno predstavimo na več načinov. Na sliki 2.2 je prikazan grafični model poravnave [22].



Slika 2.2: Grafična predstavitev poravnave besed med angleškim in španskim stavkom

Za poravnavo besed se v praksi uporablja več metod, ki se med seboj razlikujejo po tem, kako izračunavajo verjetnost in kateremu vidiku dajejo več prednosti. Najbolj znane med njimi so:

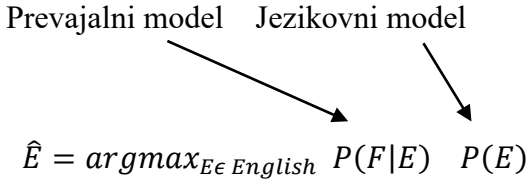
- prevajanje besed (word translation IBM-1),
- lokalna poravnava (local alignment IBM-2),
- plodnost (fertility IBM-3),
- poravnava na osnovi razredov (class-based alignment IBM-4),
- poravnava brez nesmislov (non-deficient alignment IBM-5),
- skriti Markovski model (Hidden Markov Model (HMM)),

- kontekstno odvisni modeli [4].

Ko je postopek poravnave besed končan, dobimo matriko poravnanih besed, ki pa jih je treba še dodatno obdelati, da dobimo matriko besednih zvez. Ko je učni model končan, lahko začnemo z dekodiranjem.

2.2.2. Dekodiranje pri statističnem strojnem prevajanju na osnovi besednih zvez

Končna komponenta v sistemu statističnega strojnega prevajanja je dekoder. Njegova naloga je, da iz tujega izvirnega stavka F ustvari najboljši možni prevod E v skladu s prevajalnim in jezikovnim modelom (Slika 2.3):



Slika 2.3: Prevajanje zahteva prevajalni in jezikovni model

Iskanje stavka, ki poišče največjo verjetnost prevajalnega in jezikovnega modela, je iskalni problem in tako je dekodiranje pravzaprav nekakšna oblika iskanja. Dekoderji v strojnem prevajanju temeljijo na najboljšem prvem iskanju, ki je oblika hevrističnega ali informiranega iskanja. Gre za iskalne algoritme, ki dobivajo znanje iz domene problema. Algoritmi najboljšega prvega iskanja izberejo vozlišče *n* v iskalnem prostoru na osnovi ocenjevalne funkcije *f* (*n*), ki poišče prevod z največjo verjetnostjo. Ob tem upošteva stroške za iskanje besednih zvez. Ker število možnih iskanj lahko preveč naraste, algoritem uporabi tudi rezanje, ki odstrani neobetajoče variante.

2.3. Model, ki temelji na drevesih

Prevajanje po besedah in besednih zvezah je v praksi zelo uspešno, vendar vprašljivo, ker ne upošteva dovolj slovničnih pravil in zajema skladnjo posredno prek uporabe n-gramskega[1] jezikovnega modela, ki pa ne more modelirati dolgih odvisnosti (priredij in podredij).

[1] n-gram je niz elementov (npr. besed) dolžine n

Prav tako se pri statističnem strojnem prevajanju na osnovi besednih zvez drevo možnih kandidatov zelo poveča in je treba rezanje in izbiranje najprimernejših kandidatov.

Zato se model na osnovi skladenjskih dreves osredotoči na skladenjske jezikovne modele. Na ta način so lahko upoštevane funkcijske besede (predlogi in določila), prevajanje glagola je lahko odvisno od osebka ali predmeta, možno je premikanje stavčnih členov po prevodu (npr. predmeta na konec stavka).

Za razliko od tradicionalnih modelov na osnovi besednih zvez, kjer je prevajanje izvedeno s preslikavo vhodne besedne zveze v izhodno besedno zvezo, poteka pri drevesnih modelih prevajanje s pomočjo t. i. slovničnih pravil. Pravila vključujejo spremenljivke v pravilih preslikovanja, ki so videti, kot prikazuje slika 2.4 za primer prevajanja med angleščino in nemščino ter francoščino [25].

```
ne X1 pas -> not X1          (French-English)
ate X1 -> habe X1 gegessen (English-German)
X1 of the X2 -> le X2 X1     (English-French)
```

Slika 2.4: Primer dela datoteke s slovničnimi pravili

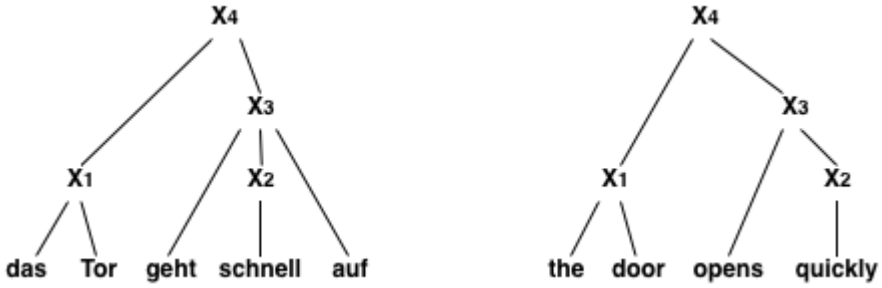
Spremenljivke v teh slovničnih pravilih se imenujejo ne-končne, ker njihova pojavitev nakazuje, da pot do končnih besed še ni končana. Poleg splošne spremenljivke `X` lahko uporabimo tudi slovnične ne-končne spremenljivke, kot sta `NP` (samostalnik – *noun phrase*) ali `VP` (glagol – *verb phrase*).

Med prevajanjem se zgradi podatkovna struktura v obliki drevesa, ki jo lahko nazorneje pokažemo z naslednjim zgledom. Vzemimo, da vsebina na sliki 2.5. predstavlja vhodne podatke in pravila prevajanja:

```
Input: Das Tor geht schnell auf
Rules: Das Tor -> The door
       schnell -> quickly
       geht X1 auf -> opens X1
```

Slika 2.5: Zgled vhodnih podatkov in pravil prevajanja za drevesni model

Če ta pravila uveljavimo v danem vrstnem redu, je prevod *The door opens quickly* zgrajen na naslednji način:



Slika 2.6: Drevo prevajanja stavka »Das Tor geht schnell auf«

Najprej je izvedena enostavna preslikava besednih zvez (1) *Das Tor* v *The door* in (2) *schnell* v *quickly*. Po končani preslikavi je mogoče uveljaviti kompleksnejše pravilo (3) *geht* X_1 *auf* v *opens* X_1. Na tej točki je X, ki zajema nemško besedo *schnell*, zamenjan s prevodom *quickly*. V zadnjem koraku pravilo lepljenja (4) X_1 X_2 v X_1 X_2 združi oba dela v en celoten stavek [25].

Pri prevajanju po skladnji snovalci najprej izdelajo skladenjsko drevo, samo dekodiranje pa je za razliko od prevajanja na osnovi besednih zvez, kjer gre za obliko iskanja, tukaj razčlenjevanje (*angl. parsing*) [24].

2.4. Faktorski model

Faktorski jezikovni modeli delujejo na kontekstu prejšnjih n besed, podobno kot modeli, ki temeljijo na n-gramskih besednih zvezah, vendar faktorski model zajema širši kontekst kot enako velik n-gramski model na osnovi besednih zvez.

Faktorski jezikovni modeli ne zmanjšujejo kompleksnosti n-gramskega modela, ampak ga nadgrajujejo z bogatejšim nizom možnosti postavljanja pogojev, dodajanjem dodatnih informacij o besedah in s tem izboljševanjem učinkovitosti prevajanja [5].

V faktorskem jezikovnem modelu je beseda obravnavana kot zbirka funkcij ali faktorjev, od katerih je lahko ena od njih tudi dejanska površinska oblika besede. Beseda w je skupek ali vektor K (vzporednih) faktorjev, za katere velja:

$$w \equiv \{f, f^1, f^2, \dots, f^K\} = f^{1:K}$$

Zapis verjetnosti faktorskega jezikovnega modela za stavek s T besedami, od katerih ima vsaka K faktorjev, je:

$$P(w_1, w_{2,\dots,} w_T) = P(f_1^{1:K}, f_2^{1:K}, \dots, f_T^{1:K}) = P(f_{1:T}^{1:K})$$

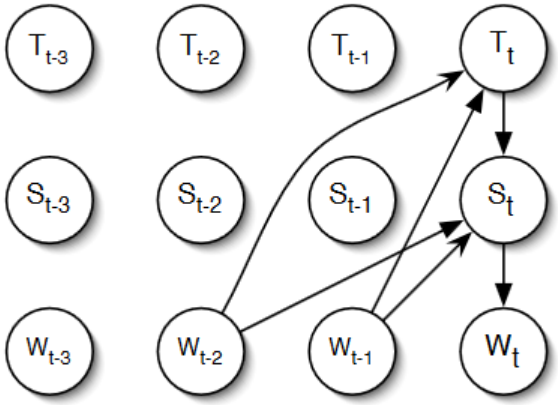
Faktorji besede so lahko karkoli, kot denimo besedna vrsta (samostalnik, glagol, pridevnik), lastnosti besed (spol, število, sklon), koren, lema ali katerakoli druga jezikoslovna zvrst, ki jih lahko srečamo v visoko pregibnih jezikih. Faktor je lahko denimo sama površinska oblika besede, tako da lahko verjetnostni jezikovni model pokriva besede ter tudi njene dekompozicijske faktorje. Axelrod [5] navaja primer besed iz angleškega jezika, pri katerih imamo na voljo informacije o stavčnih členih in jih želimo uporabiti kot dodatni faktor. Njihova faktorska predstavitev besed je videti takole:

$$the = (\text{'}the\text{'}, article)$$
$$black = (\text{'}black\text{'}, adjective)$$
$$cat = (\text{'}cat\text{'}, noun)$$

Faktorje lahko izberemo ročno, če poznamo jezikovna pravila določenega jezika, ali pa prek podatkovno usmerjenega pristopa. Na določanje najboljšega verjetnostnega modela za dani niz faktorjev lahko gledamo kot na problem učenja strukture za grafične modele, pri čemer je cilj zmanjšanje zmedenosti jezikovnega modela [8].

Na sliki 2.7. je prikazan grafični model faktorskega jezikovnega modela, ki ocenjuje verjetnost naslednje besede v besedilu z uporabo morfoloških faktorjev in korenov besed kot dodatek k običajnim površinskim oblikam besed. Ocena verjetnosti tega jezikovnega modela za besedo w_t je izračunana v skladu z naslednjim modelom:

$$P(w_t) = P(w_t | s_{t,} m_t) \cdot P(s_t | m_t, w_{t-1}, w_{t-2}) \cdot P(m_t, w_{t-1}, w_{t-2})$$



Slika 2.7: Grafični model faktorskega jezikovnega modela za besede W, korene S in morfološke faktorje M.

2.5. Model na principu globokih nevronskih mrež

Za razliko od tradicionalnih prevajalnih sistemov na osnovi besednih zvez, ki so sestavljeni iz več manjših podkomponent, ki jih uglašujemo ločeno, nevronsko strojno prevajanje poskuša zgraditi in naučiti eno samo globoko nevronsko mrežo. Ta prebere stavek in kot izhodne podatke vrne pravilen prevod [6]. Nevronsko strojno prevajanje je nevronska mreža, ki neposredno modelira pogojno verjetnost $p(y|x)$ za prevajanje stavka x v ciljni stavek y.

Model umetne nevronske mreže (*angl. artifical neural network*) se zgleduje po zasnovi in delovanju bioloških nevronskih mrež možganov. Z uporabo ANN je mogoče izvesti veliko število nalog, kot so klasifikacija, združevanje v gruče in napovedovanje. To dosežemo tako, da za učenje in prilagajanje povezav v mreži uporabimo tehnike strojnega učenja, kot je denimo nadzorovano ali spodbujevano (*angl. reinforced*) učenje. Če primerjamo umetne nevrone z biološkimi, lahko primerjamo vhodne podatke z dendriti, funkcijo aktivacije s proženjem nevrona, če je dosežen prag napetosti, izhodne podatke pa z aksoni.

Prevajanje poteka tako, da nevronska mreža s koderjem, ki je dvosmerna rekurzivna nevronska mreža (*angl. recurrent neural network*) (RNN), najprej kodira stavek v izvornem jeziku. Drugi RNN, ki ga imenujemo dekoder, je nato uporabljen za napovedovanje besed v ciljnem jeziku [6]. Z uporabo mehanizma pozornosti (*angl. attentional mechanism*) je mogoče nevronsko strojno prevajanje izboljšati, da se med prevajanjem selektivno osredotoči na določene dele izvornega stavka. Tako se denimo prevajalni sistem lahko osredotoči na določen samostalnik in gradi prevod na tem.

Ko je nova nevronska mreža izdelana, se uči za določeno področje ali uporabo. Ko je enkrat vzpostavljen mehanizem za učenje, nevronska mreža vadi in sčasoma začne delovati v skladu z lastno presojo.

Eno od izvedb nevronskega prevajanja v času pisanja tega magistrskega dela uporabljajo pri podjetju Google. Za razčlenjevanje uporabljajo SyntaxNET [40], ki ga uporabijo kot globalno prevajalno platformo za druge Googlove aplikacije. Ocenjujejo ga kot trenutno najnatančnejši razčlenjevalnik na svetu [40].

Googlovo odprtje razčlenjevalnika SyntaxNET za javnost odpira nove možnosti za razvoj strojnega prevajanja v smeri globokih mrež. Zadnje raziskave [43] kažejo, da s pristopom globokih mrež dobimo boljše rezultate od statističnega strojnega prevajalnika za isti testni prevod. Učenje modelov je za zdaj precej daljše od pristopa na osnovi besednih zvez, je pa dobljeni model manjši in prevajanje hitrejše.

2.6. Evalvacija strojnih prevodov

Ko je postopek strojnega prevajanja končan, nas zanima kakovost dobljenega prevoda. Jurafsky in Martin [22] ocenjujeta, da pri ocenjevanju strojnih prevodov na splošno poskušamo oceniti prevode v okviru dveh kriterijev, in sicer glede na natančnost in razumljivost. Ocenjevanje je glede na način pregleda mogoče izvesti ročno, s pomočjo strokovnjakov, ali strojno oz. samodejno s pomočjo ocenjevalnih algoritmov.

2.6.1. Ročno ocenjevanje strojnih prevodov

Pri ročnem ocenjevanju strojnih prevodov strokovnjaki ocenjujejo posamezen prevod glede na natančnost in razumljivost [22].

Tako lahko denimo z vidika razumljivosti ocenjevalce vprašamo kako razumljiv, kako jasen, kako berljiv ali kako naraven je strojno prevedeni prevod. Strokovnjaki (človeški ocenjevalci) lahko na ta vprašanja odgovorijo na dva glavna načina. Pri prvem načinu ocenjevalci prevod ocenjujejo z lestvico od denimo 1 (popolnoma nerazumljivo) do 5 (popolnoma razumljivo). Pri drugem načinu pa se manj zanašamo na zavestne odločitve sodelujočih. Merimo lahko na primer čas, potreben, da ocenjevalci preberejo vsak stavek ali odstavek strojnega prevoda. Jasnejše ali bolj tekoče stavke je lažje ali enostavneje prebrati.

Podoben nabor metrik lahko uporabimo za natančnost. Dva pogosto merjena vidika natančnosti sta ustreznost in informativnost. Ustreznost prevoda ocenjuje, koliko informacij iz izvirnika je ohranjenih v prevodu. Za informativnost prevoda lahko rečemo, da gre za oceno na osnovi naloge, ali je v strojnem prevodu dovolj informacij za izvedbo določene naloge [22].

Ročno ocenjevanje sega že daleč v preteklost. Leta 1966 je ALPAC (*Automatic Language Processing Advisory Committee*) [1] objavil raziskavo na temo ročne evalvacije, v kateri so ocenjevali človeške in strojne prevode. Ocenjevalci so bili usposobljeni posebej za raziskavo. Ocenjevali so prevode iz ruščine v angleščino, in sicer z dveh vidikov: z vidika razumljivosti (*angl. intelligibility*) in z vidika zvestobe (*angl. fidelity*).

Razumljivost so ocenjevali z ocenami od 1 do 9, brez reference izvirnika.

Zvestobo, torej kako informativen je prevod v primerjavi z izvirnikom, so vrednotili z ocenami od 0 do 9. Najprej so prebrali prevod, potem izvirnik, nato pa ocenili, koliko informacij prinaša izvirnik v primerjavi s prevodom – več novih informacij prinaša izvirnik, slabši je prevod.

Raziskava je pokazala, da so bile razlike med ocenjevalci majhne, kljub temu pa priporočajo, da pri evalvaciji sodelujejo vsaj trije ali štirje ocenjevalci. Ocenjevalci so zlahka ločili, ali gre za človeški ali za strojni prevod [49].

V podjetju ARPA (*Advanced Research Projects Agency*) so leta 1991 pod okriljem projekta Human Language Technologies Program vzpostavili evalvacijski program in izdelali metodologijo za evalvacijo strojnih prevajalnikov. Glavni izziv pri evalvaciji je bil zmanjšati subjektivnost. Ocenjevanje mora biti intuitivno in hkrati kar se da objektivno, kar se kaže v minimalnih odstopanjih med ocenjevalci. V okviru evalvacijskega programa so z različnimi metodami ovrednotili strojne prevode in najprimernejše metode obdržali [50].

Na podlagi večletnih izkušenj in opazovanj so leta 2013 v združenju TAUS (*Translation Automation User Society*) pripravili nove smernice za ocenjevanje strojnih prevodov [46]. Osredotočili so se na ocenjevanje kakovosti s stališča berljivosti in ustreznosti. Ta metoda se redno uporablja za ocenjevanje strojnih prevodov, uporabna pa je tudi za ocenjevanje človeških prevodov v določenih kontekstih. Metoda z ocenjevanjem berljivosti in ustreznosti je cenejša in manj časovno zahtevna kot pristop z ocenjevanjem več tipov napak in omogoča, da se osredotočimo na ocenjevanje atributov, ki so najbolj relevantni za določene vrste vsebine.

Glede ustreznosti ta metoda ocenjevalca sprašuje, koliko pomena, izraženega v referenčnem prevodu ali izvirnem besedilu, je izraženega tudi v ciljnem prevodu. Glede ustreznosti pa ocenjevalci ocenijo, v kolikšni meri je prevod slovnično pravilen, brez črkovalnih napak in govorcu maternega jezika zveni naravno.

Ta metoda narekuje, da pri ocenjevanju prevoda sodelujejo vsaj štirje ocenjevalci, da se izračuna raven strinjanja, ter da ocenjevalci ocenjujejo iste podatke [46].

Metodo TAUS uporabimo tudi za ocenjevanje prevodov v tem magistrskem delu. Več o tem načinu ocenjevanja opišemo v poglavju 6.5.

2.6.2. Povratni prevod

Povratni prevod (*angl. round-trip translation*) je eden od načinov, da ocenimo kakovost prevoda. »Povratni prevod je prevod, ki smo ga s pomočjo istega strojnega prevajalnika najprej prevedli v ciljni jezik, nato pa nazaj v izvirni jezik. Največja težava pri tem je, da ne moremo vedeti, ali je sistem naredil napako ob prevajanju v drug jezik ali ob prevajanju nazaj v izvirni jezik.« [49]

Spodnji primeri prikazujejo, kako je lahko povratno prevajanje za evalvacijo zavajajoče. V prvem primeru je prevod v italijanščino popolnoma sprejemljiv, medtem ko je v povratnem prevodu kar nekaj napak. V drugem primeru je povratni prevod identičen izvirniku, medtem ko je prevod v portugalščino brezpredmeten. Pri tretjem primeru (iz Google Prevajalnika) je prevod slovenskega frazema v angleščini popolnoma nesmiseln, povratni prevod pa je sicer slovnično pravilen, a nerelevanten, saj se ob zamenjani informaciji izgubi pomen frazema. [49]

Angleški izvirnik: **Select this link to look at our home page.**

Prevod v italijanščino: **Selezioni questo collegamento per guardare il nostro Home Page.**

Povratni prevod: **Selections this connection in order to watch our Home Page.**

Angleški izvirnik: **Tit for tat.**

Prevod v portugalščino: **Melharuco para o tat.**

Povratni prevod: **Tit for tat.**

Slovenski izvirnik: **Ne tič ne miš.**

Prevod v angleščino: **Do not cock Sun mouse.**

Povratni prevod: **Ne petelin ne miš.**

2.6.3. Samodejno ocenjevanje

Človeška ocena strojnega prevoda je sicer najboljša, vendar tudi zelo časovno zahtevna in lahko traja dneve ali celo tedne. Zato je zelo uporabno imeti metriko za samodejno ocenjevanje, ki jo je mogoče izvajati relativno pogosto z namenom hitre ocene možnih izboljšav sistema. V ta namen smo pripravljeni sprejeti veliko slabšo metriko v primerjavi s človeško oceno, le da vključuje nekaj sorodnosti s človeškim ocenjevalcem [22].

Obstaja več hevrističnih metod, s katerimi lahko samodejno izmerimo kakovost prevoda. Pri vsaki od teh metod predvidevamo, da imamo na voljo človeške prevode relevantnih stavkov. Na podlagi strojno prevedenega stavka izračunamo bližino prevoda izvirniku. Strojni prevod je obravnavan kot boljši, če je v povprečju bližje človeškemu prevodu. Različne metrike se razlikujejo v tem, kaj pomeni bližina izvirniku [22]. Ker v tem magistrskem delu za samodejno ocenjevanje kakovosti prevoda uporabimo matriko BLEU, jo v nadaljevanju podrobneje predstavimo.

BLEU (Bilingual Evaluation Understudy)

Pri metriki BLEU [39] rangiramo vsak strojni prevod z uteženim povprečjem števila N-gramov, ki se prekrivajo s človeškim prevodom. Na sliki 2.8 sta povzeta dva kandidatna prevoda izvirnika [39] skupaj s še tremi referenčnimi človeškimi prevodi izvirnega stavka.

Kand 1: It is a guide to action which ensures that the military always obeys the commands of the party

Kand 2: It is to ensure the troops forever hearing the activity guidebook that party direct

Ref 1: It is a guide to action that ensures that the military will forever heed Party commands

Ref 2: It is guiding principle which guarantees the military forces always being under the command of the Party

Ref 3: It is the practical guide fort he army always to heed the directions of the party

Slika 2.8: En prevod si deli več besed z referenčnimi človeškimi prevodi [22]

Opazimo, da si Kandidat 1 deli veliko več besed z referenčnim prevodom kot Kandidat 2. Oglejmo si, kako se izračuna rezultat BLEU. Za začetek samo na unigramih. BLEU temelji na natančnosti. Osnovna meritev natančnosti unigramov je kar število besed v kandidatnem prevodu (strojni prevod), ki se pojavijo tudi v katerem od referenčnih prevodov, deljeno s skupnim številom besed v kandidatnem prevodu.

Če ima kandidatni prevod 10 besed in se jih 6 od teh pojavi v vsaj enem od referenčnih prevodov, dobimo natančnost 6/10 = 0,6. Če ima kandidatni prevod 10 besed in se jih 6 od teh pojavi v vsaj enem od referenčnih prevodov, dobimo natančnost 6/10 = 0,6.

Na žalost pa ima uporaba zgolj besed za oceno natančnosti slabost: nagrajeni so kandidati, ki imajo ponovljene besede.

Kandidat:	the the the the the the the
Referenca 1:	the cat is on the mat
Referenca 2:	there is a cat on the mat

Slika 2.9: Primer unigramske in spremenjene unigramske natančnosti [22]

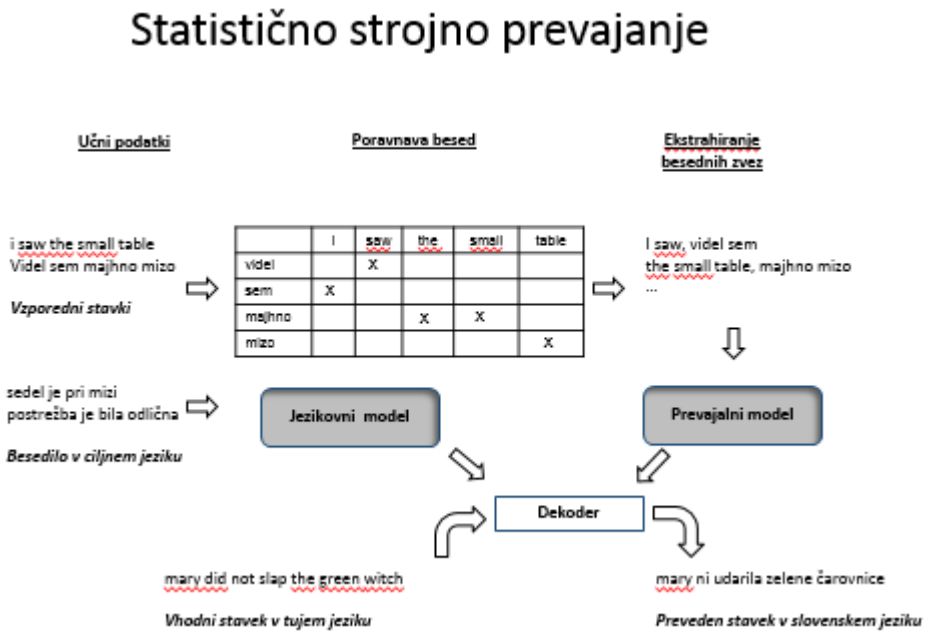
Slika 2.9 prikazuje sicer izmišljen primer kandidatnega stavka, sestavljenega iz več primerov iste besede *the*. Ker se vsaka od sedmih (identičnih) besed iz kandidatnega stavka pojavi v enem od referenčnih prevodov, bi bila unigramska natančnost 7/7. Temu problemu se v metodi BLEU izognemo z uporabo spremenjene n-gramske natančnosti. Najprej preštejemo največje število ponovitev besede v kateremkoli referenčnem prevodu. Število posamezne kandidatne besede je nato zmanjšana za to maksimalno referenčno število. Unigramska natančnost na sliki 2.9 bi bila 2/7, ker ima Referenca 1 maksimalno dve ponovitvi besede *the* [22].

Metoda BLEU je kljub temu, da je v uporabi že vrsto let, še vedno ena od najbolj uporabljenih metod za oceno strojnih prevodov.

3. Prilagajanje SSP za specifično domeno

Za učinkovito prevajanje potrebujejo statistični strojni prevajalniki zadostno količino vzporednih podatkov za učenje. Statistični strojni prevajalniki dajejo dobre rezultate v primerih, ko imajo na voljo dovolj učnih podatkov. Nieheus in Waibel glede obstoja podatkov s specifičnega področja ugotavljata, da podatki vedno niso na voljo, so na voljo samo v izvornem ali samo ciljnem jeziku, je na voljo samo slovar s tega področja ali pa je na voljo omejena količina dvojezičnih podatkov [35]. To velja še posebej za jezike z manj jezikovnimi viri, kot je denimo slovenščina.

Preden predstavimo možnosti za prilagajanje statističnega strojnega prevajanja za specifično domeno, si oglejmo glavni princip delovanja tega načina strojnega prevajanja. Na sliki 3.1 je prikazano, da dekoder v času prevajanja stavka iz izvornega jezika v ciljni jezik dobi podatke iz jezikovnega in prevajalnega modela ter jih uporabi za prevajanje poljubnega besedila, ki mu ga predložimo.



Slika 3.1: Princip delovanja statističnega strojnega prevajanja

Na kakšen način bomo strojni prevajalnik prilagodili za specifično domeno, se odločimo glede na obstoj virov v našem primeru. Pri prilagajanju se lahko osredotočimo na prilagajanje prevajalnega modela ali prilagajanje jezikovnega modela.

3.1. Prilagajanje jezikovnega modela

V tem podpoglavju opišemo pristope k prilagajanju jezikovnega modela, v nadaljevanju pa je navedenih še nekaj primerov dobrih praks, ki so jih predstavili različni raziskovalci na področju prilagajanja SSP.

3.1.1. Pristopi k prilagajanju jezikovnega modela

Z jezikovnim modelom na osnovi besed je določena verjetnost *p(e)*, da beseda *(e)* v besedilu, ki ga prevajamo, nastopa v zbirki ciljnega jezika. Izboljšava modela na osnovi besed je n-gramski jezikovni model, ki dodeli verjetnosti hipotezam v ciljnem jeziku na podlagi zgodovine konteksta predhodnih n-1-gramskih besednih zvez [56].

Za statistični jezikovni model lahko poenostavljeno rečemo, da izraža verjetnost, da je zaporedje besed tekoče zaporedje besed v določenem jeziku. Verjetna zaporedja besed dobijo višje verjetnosti, medtem ko nesmiselna zaporedja dobijo nizke verjetnosti [5].

Na sliki 3.2 je prikazan primer jezikovnega modela v obliki zapisa ARPA, ki smo ga zgradili z orodjem KenLM med postopkom gradnje prevajalnega sistema.

```
\data\
ngram 1=97539
ngram 2=828958
ngram 3=1829701

\1-grams:
-4.700301      svojim -0.24061029
-4.3441725     zdravnikom     -0.335519
-4.987664      farmacevtom    -0.3012766
. . .
\2-grams:
-1.8700966     ponavadi </s>  0
-1.392477      ščitnice </s>  0
-1.5096847     hipofize </s>  0
. . .
\3-grams:
-1.1149234     agencije EMEA </s>
-0.8040243     zahtevo EMEA </s>
-0.18791237    postopka EMEA </s>
```

Slika 3.2: Primer dela jezikovnega modela v obliki ARPA

Slika prikazuje trigramski jezikovni model, ki je sestavljen iz treh delov. V glavi je navedeno število vnosov za vsakega od delov. Tako je v datoteki skupaj prisotnih 97.539 unigramov. Nato so v vsakem od treh delov zapisane definicije n-gramov. Definicija vsakega n-grama se začne z verjetnostno vrednostjo v obliki desetiškega logaritma (log_{10}), tej sledi zaporedje n besed, ki opisuje dejanski n-gram. Kot zadnji podatek je pri dveh razdelkih zapisana vračalna utež (*angl. back-off weight*), ki je prav tako v obliki (log_{10}). Vračalni postopki se uporabljajo, če jezikovni model odkrije zaporedje besed, ki ga predhodno še ni videl. V tem primeru se vrne nazaj na zanesljivejšo, vendar mogoče manj natančno oceno verjetnosti.

Ko govorimo o prilagajanju statističnega strojnega prevajanja za specifično domeno, se za prilagajanje jezikovnega modela odločimo, ko imamo na voljo dovolj podatkov v ciljnem jeziku za določeno domeno oz. področje.

Prilagajanje jezikovnega modela je v praksi osredotočeno na dva načina:

- usmerjeno v količino podatkov,
- usmerjeno v globino podatkov.

Za način, usmerjen v količino podatkov, velja, da več kot imamo podatkov, boljše je. Pri tem načinu za nekatere jezike zbirke obsegajo milijardo ali celo bilijon besed, jezikovni modeli pa so porazdeljeni v več delov. Tovrstni jezikovni modeli dosegajo boljše rezultate pri prevajanju, vendar zahtevajo veliko pomnilniških kapacitet in procesorskih moči [56].

Bulyko in sod. [9] v svoji raziskavi glede prilaganja jezikovnega modela opozarjajo, da se lahko pri gradnji domenskega jezikovnega modela v primeru ogromnih zbirk zgodi, da izgubimo slogovne značilnosti posameznih n-gramov, če jih združimo z n-grami vseh korpusnih zbirk. V ta namen predlagajo gradnjo ločenih jezikovnih modelov, ki jih nato medsebojno kombiniramo z interpolacijo verjetnosti n-gramov.

Način, usmerjen v globino podatkov, se osredotoča na globino enojezičnih podatkov, da zgradi slovnično informirane jezikovne modele. Nekateri avtorji predstavijo jezikovne modele na osnovi skladnje, ki so zgrajeni s pomočjo skladenjskih razčlenjevalnih dreves [10]. Na področju, ki ga povezujemo z globino jezikovnega modela, lahko gradnjo običajnega jezikovnega modela razširimo na kombinacijo predhodnega n-gramskega jezikovnega modela in vnaprejšnjega n-gramskega jezikovnega modela ter s tem zajamemo medsebojne povezave besed. Pri običajnih n-gramskih jezikovnih modelih izračunavamo verjetnost trenutne besede

glede na predhodnih n-1 besed. Pri tej izboljšavi pa jezikovni model zgradimo s kombinacijo predhodniškega in naslednjiškega jezikovnega modela. V predhodniškem modelu je pri izračunu verjetnosti upoštevanih prejšnjih n-1 besed ter trenutna, medtem ko je v naslednjiškem upoštevana trenutna in naslednjih n-1 besed. Na ta način zajamemo tudi odvisnosti na daljši razdalji [56].

3.1.2. Primeri dobrih praks

Na področju izboljševanja jezikovnega modela Bulyko in sod. [9] v svoji raziskavi preverijo predvidevanje, da se v primeru zelo velikih jezikovnih zbirk zaradi same velikosti lahko izgubijo slogovne značilnosti posameznih stavkov oz. n-gramov, ki bi se ohranili, če bi izdelali jezikovni model na podlagi testnega besedila, ki ga je treba prevesti. Dokažejo, da lahko s prilagajanjem jezikovnega modela in vnovičnim določanjem n-najboljših vrednosti izboljšamo učinkovitost strojnega prevajanja od 0,3 do 0,4 BLEU točke v primerjavi z neprilagojenim jezikovnim modelom.

Xiong in sod. [56] se osredotočijo na izboljšanje jezikovnega modela z vidika prepoznavanja daljnosežnih odvisnosti besed. V statistični strojni dekoder vgradijo dva jezikovna modela. Predhodniški jezikovni model zajema predhodni kontekst trenutne besede, medtem ko naslednjiški model zajema prihodnji kontekst trenutne besede. Na ta način pri prevajanju dosežejo izboljšanje do ene BLEU točke v primerjavi z običajnim jezikovnim modelom.

V primerih, ko za točno določeno področje ni na voljo dovolj podatkov, na voljo pa so podatki z zelo podobnih področij, lahko jezikovni model zgradimo s pomočjo teh dokumentov. Kako to naredimo opišejo Snover in sod. [48] v svoji raziskavi, kjer za vsak dokument, ki ga je treba prevesti, v veliki zbirki enojezičnih podatkov poiščejo dokumente oz. podatke, ki bi bili lahko podobni vsebini izvornega dokumenta ter z njimi prilagodijo jezikovni model ter s tem prevajalni sistem.

3.2. Prilagajanje prevajalnega modela

Podobno kot za prilagajanje jezikovnih modelov predstavimo tudi prilagajanje prevajalnih modelov. Najprej predstavimo kaj so prevajalni modeli, v nadaljevanju pa kako jih lahko prilagodimo za specifično domeno in kaj so različni avtorji na tem področju že naredili.

3.2.1. Pristopi k prilagajanju prevajalnega modela

Preden se osredotočimo na prilagajanje prevajalnega modela, poglejmo, kako je opredeljen prevajalni model. Dokument je pri statističnem strojnem prevajanju preveden v skladu z verjetnostno porazdelitvijo $p(e|f)$, da je niz e v ciljnem jeziku prevod niza f v izvornem jeziku. Eden od pristopov modeliranja verjetnostne porazdelitve $p(e|f)$, ki se je uveljavila v računalniški izvedbi, je uporaba Bayesovega teorema, da je $p(e|f) \propto p(f|e)p(e)$, pri čemer je prevajalni model $p(f|e)$ verjetnost, da je izvorni niz prevod ciljnega niza, jezikovni model $p(e)$ pa verjetnost, da niz e nastopa v ciljnem jeziku [53].

Na sliki 3.3 je prikazan del datoteke s prevajalno tabelo.

```
agency ||| agencija ||| 0.3 ||| |||
pill   ||| tableta ||| 0.4 ||| |||
this is ||| to je ||| 0.8 ||| |||
```

Slika 3.3: Primer datoteke prevajalne tabele

Če pogledamo prvo vrstico primera datoteke na sliki 3.3, vidimo vnos, ki pomeni, da je verjetnost prevajanja angleške besede *agency* v slovensko besedo *agencija* enaka 0,3.

Za prevajalne modele je značilno, da so zgrajeni iz dvojezičnih zbirk podatkov. Te so precej obsežne za določena splošnejša področja, katerih zbirke izvirajo iz javnih institucij ali so zbirke same po sebi javnega značaja. Če želimo prevajalni model prilagoditi za določeno področje, si moramo priskrbeti tudi ustrezne podatke za to področje. Tu pa lahko pridemo do več izzivov, ki jih na grobo lahko razdelimo na dve skupini:

- prva skupina izzivov se sooča s pomanjkanjem ali majhno količino podatkov za določeno področje;
- druga skupina pa se posveča izboljšanju kakovosti prevoda z boljšim izkoriščanjem obstoječega prevajalnega modela.

V primerih, ko dvojezične vzporedne učne zbirke niso na voljo, lahko zgradimo sintetične vzporedne zbirke, tako da enojezični domenski korpus v izvornem jeziku prevedemo z osnovnim, splošnim sistemom ter nato prevedeni korpus uporabimo za učenje novega sistema ali prilagodimo osnovni sistem. Ta postopek prevajanja lahko iterativno večkrat ponovimo, da

z vsakim naslednjim prehodom dobimo boljši sistem [35]. Slabost tega sistema je vprašljiva kakovost dobljenega prevoda korpusa, ker lahko pridemo do problema manjkajočih izrazov. V primeru, da imamo za določeno področje na voljo slovar izrazov, lahko tega uporabimo pri prevajanju korpusa iz izvornega jezika v ciljni jezik in tako izboljšamo kakovost prevoda. Wu in sod. [55] v svoji raziskavi vidijo možnost za izboljšanje sintetičnega ciljnega korpusa v dopolnitvi prevajanja izvornega korpusa z dodatkom slovarja, ki ga izdelajo ročno. Več o tem opišemo v podpoglavju 3.2.2.

V praksi imamo zelo pogosto situacijo, ko je na voljo velika splošna zbirka dvojezičnih podatkov, medtem ko je zbirka podatkov s specifičnega področja precej manjša ali celo majhna. V tem primeru se osredotočimo na boljše izkoriščanje obstoječih virov. V prvi vrsti lahko na različne načine kombiniramo domensko in splošno zbirko. Zbirki lahko uporabimo sami po sebi, obe združimo v eno ali pa uporabimo kombinacijo obeh ali celo več zbirk, tako da med dekodiranjem uporabljamo dve ali več zbirk. Relevantnost podatkov v vseh zbirkah ni nujno enaka in statistično gledano se bo stavek, ki bi bil v očeh človeškega prevajalca najprimernejši za izvirni stavek, izgubil v množici ostalih podatkov, ki imajo enako ali celo boljšo statistiko. Dekoder daje vsem modelom enako prednost, želimo pa si, da bi v takih primerih dal prednost zbirkam s specifičnega področja. Dajanje prednosti lahko izvedemo na več načinov in eden od njih je uporaba faktorskega modela, ki v prevajalni sistem vpelje dodatne parametre. Z dodatnimi parametri lahko damo prednost posameznim zbirkam, v osnovi pa je faktorski model predstavljen predvsem z namenom izboljšanja prevodov v visoko pregibne jezike, kot so denimo slovenščina, nemščina ipd. S tem načinom namreč vsaki besedi dodamo dodatne informacije o njenih lastnostih [35].

Terminologija je ena od možnosti, s katero lahko tudi izboljšamo kakovost prevoda z določenega področja. Ne spada sicer v skupino prilagajanja jezikovnega modela, je pa z njim tesno povezana. Značilnost prevajanja besedil s specifičnega področja je med drugim tudi ta, da se uporablja poleg posebnega sloga tudi specifična terminologija. Zato želimo v prevajalni sistem vpeljati tudi dvojezične slovarje. Ti so včasih že na voljo, če pa niso, jih lahko zgradimo s pomočjo majhne področne dvojezične zbirke. Obstoječe ali pridobljene izraze nato vdelamo v prevajalni sistem [3].

Med prevajanjem večjega terminološko bogatega projekta, ki ga prevaja več prevajalcev, lahko pride do primera, ko so določeni izrazi prevedeni na različne načine s strani različnih prevajalcev. Zato imamo na voljo možnost, da pred začetkom prevajanja velikega projekta

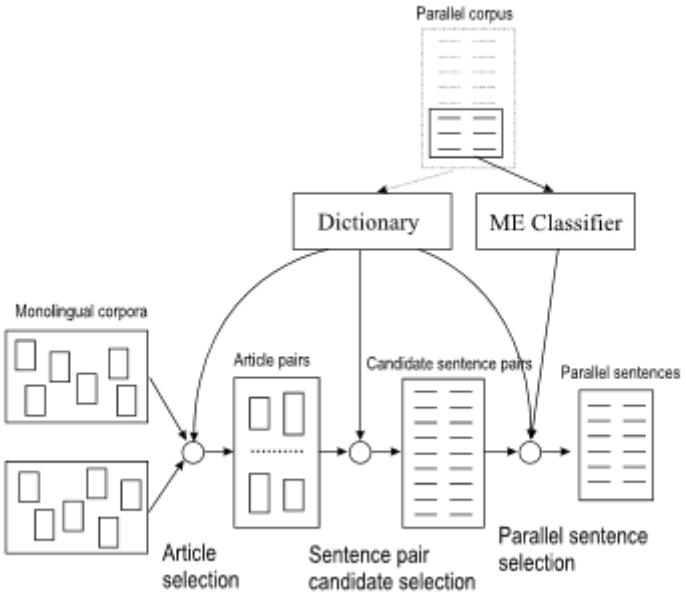
pripravimo seznam vseh izrazov, ki se v projektu ponovijo n-krat. Če želimo doseči večjo skladnost prevoda, bomo pripravili seznam vseh izrazov, ki se ponovijo trikrat ali večkrat. Ta seznam lahko nato med prevajanjem projektnih datotek prevedemo z uporabo strojnega prevajalnika, ki ga učimo z zbirko, ki jo prevajalci tekom projekta stalno dopolnjujejo. Prevedene izraze strokovnjaki pregledajo, nato pa so po metodi dopolnjevanja (*angl. fill-in*) dodani v sistem strojnega prevajalnika. Z dodajanjem novih izrazov ter na novo prevedenih stavkov strojnemu prevajalniku se izboljša kakovost prevodov, ki jih ta nudi projektnim prevajalcem. V tem primeru imamo eno temeljno zbirko in predpomnilniško zbirko (*angl. cache based*), ki jo stalno posodabljamo [3].

3.2.2. Primerih dobrih praks

Naj znova omenimo, da imamo v določenih primerih za statistične strojne prevajalne sisteme na voljo samo velike splošne zbirke, ki pa ne dajejo dobrih rezultatov pri prevajanju v specifični domeni. S problemom pomanjkljivih dvojezičnih domenskih podatkov se v svoji raziskavi ukvarjajo Wu in sod. [55]. Za izhodišče vzamejo dejstvo, da so poleg splošnih zbirk pogosteje kot dvojezični domenski korpusi na voljo domenski enojezični korpusi v izvornem ali ciljnem jeziku. Za določene domene so na voljo tudi domenski slovarji ali pa jih je mogoče relativno hitro zgraditi. V raziskavi med drugim izdelajo sintetični domenski dvojezični korpus, tako da prevedejo domenski korpus iz izvornega jezika v ciljni z uporabo domenskega dvojezičnega slovarja. Prevajanje nato po metodi transduktivnega učenja ponavljajo, dokler ni več zaznati izboljšav rezultatov. V njihovem primeru dosežejo 2,39 točke boljši rezultat v primerjavi z uporabo samo splošnega modela.

Transduktivno učenje oz. pol-nadzorovano učenje prvi predstavijo Ueffing in sod. [48] kot način za prilagoditev sistema statističnega strojnega prevajanja za novo domeno oz. novo vrsto besedila. V svoji študiji sistem, naučen za domeno novic, prilagodijo za prevajanje besedil spletnih dnevnikov. S predlagano metodo prilagodijo model slogu in domeni novega izvornega besedila, ne da bi pridobili dvojezične podatke za novo domeno.

Povečevanju velikosti domenskih dvojezičnih korpusov se posvetita tudi Munteanu in Marcu [34]. Domenske dvojezične jezikovne pare ekstrahirata iz primerljivih korpusov. Z dodajanjem ekstrahiranega dvojezičnega korpusa k učnim podatkom empirično dokažeta izboljšanje kakovosti statističnega strojnega prevajalnika.



Slika 3.4: Sistem za ekstrahiranje vzporednih stavkov [34]

Na sliki 3.4 je prikazan njun pristop k ekstrahiranju vzporednih stavkov. Iz dveh velikih enojezičnih korpusov, ki sta razdeljena po člankih, najprej izbereta pare iz podobnih člankov. Iz vsakega od parov vzameta vse možne pare stavkov in jih posredujeta skozi enostaven filter, ki išče medsebojno prekrivanje besed in tako dobita kandidatne pare. Za kandidatne pare klasifikator maksimalne entropije oceni, ali sta stavka v vsakem paru medsebojna prevoda in jih doda v zbirni korpus

V primerjavi z avtorjema Munteanu in Marcu [34], ki zbirata dvojezične podatke iz več virov v enega, Niehues in Waibel [35] predstavita pristop z uporabo več manjših ločenih domenskih zbirk. V prevajalni model uvedeta identifikator korpusa kot dodaten ciljni faktor. V tovrsten faktorski model dodata funkcije za modeliranje ustvarjanja oznak in funkcije za ocenjevanje zaporedja oznak. Pri prevajanju iz nemščine v angleščino s svojim pristopom uspeta izboljšati učinkovitost prevoda do ene BLEU točke.

Prednost jezikovnemu modelu dasta v enem od poskusov znotraj svoje raziskave tudi Koehn in Schroeder [28]. Z uporabo faktorskega modela v Mosesu razširita predstavitev besed z vektorjem faktorjev. Besedam tako dodata dodatne lastnosti, kot so besedna vrsta (samostalnik, glagol, pridevnik itd), njihove lastnosti (spol, število, sklon, koren, lema) itn. Preslikavo vhodne besedne zveze v izhodno besedno zvezo razstavita v več korakov prevajanja in generiranja, pri

čemer je pri vsakem koraku uporabljena druga prevajalna ali generativna tabela. Tako razstavljanje poimenujeta pot dekodiranja.

Ker novejši modeli poti dekodiranja omogočajo dve ali več tovrstnih poti, uporabita dve poti dekodiranja. Eno za domensko prevajalno tabelo in drugo za izven-domensko prevajalno tabelo. Z uporabo faktorskega modela dosežeta BLEU oceno 27,64, medtem ko v isti raziskavi dosežeta oceno 25,88, če uporabita zgolj domensko zbirko.

Arčan in sod. [3] domensko statistično prevajanje prilagodijo za okolje CAT s prirastnim dopolnjevanjem dvojezičnega slovarja z ekstrahiranimi dvojezičnimi izrazi iz vzporednih podatkov. Osredotočijo se na realen praktični primer, v katerem je velik prevajalski projekt razdeljen med različne prevajalce, od katerih vsak prevajalec dnevno pregleda in popravi manjšo količino stavkov, prevedenih s strojnim prevajalnikom. Te dnevno pregledane stavke nato uporabijo za generiranje domenskih slovarjev, s katerimi skupaj s pregledanimi stavki izboljšujejo splošni strojni prevajalni sistem, tako da samodejno dodajajo dvojezične izraze in optimizirajo logaritemsko-linearne uteži za te specifične podatke.

Ker zaradi stalne uporabe CAT orodja s strani prevajalcev ne morejo prekiniti prevajalskega procesa, da bi nove učne množice (slovarje in stavke) dodali med učne podatke ali na konec datoteke besednih zvez, prvič vgradijo predpomnilniški jezikovni in prevajalni model, ki ga predstavijo Bertoldi in sod. [7]. Predpomnilniški model omogoča periodično dodajanje dvojezičnih izrazov v sistem SSP v realnem času, ne da bi ga morali ustaviti.

4. Opis problema in podatkovne množice

V tem poglavju uvodoma predstavimo problem, s katerim se ukvarjamo v tem magistrskem delu, ter tudi problem, s katerim se soočajo v Lekarni Univerzitetnega kliničnega centra Ljubljana in Javni agenciji za zdravila in medicinske pripomočke in bi ga lahko naslovili z ustrezno prilagoditvijo statističnega strojnega prevajalnika. V nadaljevanju opišemo zbirke, ki ji uporabimo v našem eksperimentu.

4.1. Opis problema

V okviru te magistrske naloge se posvečamo problemu prilagajanja statističnega strojnega prevajalnika za specifično domeno. Osredotočimo se na področje farmacije, ker se nam zdi, da se to področje sooča z izzivi, pri katerih si je mogoče v veliki meri pomagati s statističnim strojnim prevajalnikom. Te probleme podrobneje opišemo v poglavju 1.2. Tu naj omenimo samo glavne, ki so:

- potreba po hitrih prevodih,
- potreba po uniformnih prevodih,
- potreba po povratnem prevodu.

Pri razvoju sodelujemo z dvema institucijama, ki sta tesno povezani s področjem farmacije. To sta Lekarna Univerzitetnega kliničnega centra v Ljubljani in Javna agencija za zdravila in medicinske pripomočke (JAZMP). V nadaljevanju ju na kratko predstavimo in opišemo del težav, s katerimi se soočajo.

4.1.1. Lekarna Univerzitetnega kliničnega centra Ljubljana

V Lekarni UKC Ljubljana si prizadevajo za zagotavljanje kakovostne preskrbe z zdravili in farmacevtskimi storitvami. Prva skrb lekarne je, da pacienti Univerzitetnega kliničnega centra Ljubljana dobijo zdravila, ki jih potrebujejo. Lekarna zdravila priskrbi po ustaljeni poti preko veledrogerij, včasih zdravilo izdela v lekarni, občasno pa zdravilo uvozi po posebnem postopku uvoza za potrebe posameznega bolnika na podlagi zahteve zdravnika.[2]

V primeru uvoza po posebnem postopku mora Lekarna občasno sama prevajati navodila za uporabo specifičnih zdravil, ki jih naroči pri proizvajalcih zdravil znotraj EU, vendar zaradi

[2] Vir: spletna stran Lekarne UKC Ljubljana (http://www.kclj.si/)

njihove ozke namembnosti nimajo slovenskih navodil za uporabo. Ko zdravilo dobijo, morajo biti navodila za uporabo, povzetek značilnosti zdravila (SmPC) in spremljajoči dokumenti hitro prevedeni. Medicinsko osebje mora zdravilo predpisati in priskrbeti pacientom, zato pa potrebujejo navodila v obliki, ki jo razumejo.

Prevajanje z zdravilom povezanih dokumentov za lekarno predstavlja dodatno delo, za katerega pa vedno nimajo na voljo dovolj virov.

4.1.2. Javna agencija za zdravila in medicinske pripomočke (JAZMP)

Kratka predstavitev

Poslanstvo Javne agencije Republike Slovenije za zdravila in medicinske pripomočke je varovanje javnega zdravja z reguliranjem in nadzorom zdravil, medicinskih pripomočkov, krvi, tkiv in celic ter z njimi povezanih dejavnosti v zasebnem in javnem sektorju.[3]

Ker so v JAZMP odgovorni za pridobitev dovoljenja za promet z zdravili, jim morajo vsi proizvajalci in trgovci zdravil predložiti povzetek glavnih značilnosti zdravila (SmPC), če ga želijo prodajati na območju Republike Slovenije. SmPC-ji so pogosto v izvirniku napisani v angleškem ali drugem tujem jeziku, nato pa prevedeni v slovenski jezik. Njihova naloga je, da te povzetke pregledajo in med drugim ugotovijo, ali so zapisani v skladu z zakonodajo in je uporabljena ustrezna terminologija. V dokumentih je lahko pogosto uporabljeno različno izrazoslovje in to povzroča slogovno in terminološko neskladnost med dokumenti.

4.2. Pristop k pridobivanju zbirk

Kot opišemo že v 2. poglavju, je statistično strojno prevajanje vrsta strojnega prevajanja, ki temelji na večji količini vzporednih besedil, iz katerih se s statističnimi algoritmi izračunavajo verjetnosti najboljših prevodov. Če želimo vzpostaviti sistem za statistično strojno prevajanje, potrebujemo ustrezne dvojezične zbirke besedil. Poleg dvojezičnih zbirk potrebujemo tudi enojezične zbirke v slovenskem jeziku ter domenske slovarje.

Kje dobiti ustrezne zbirke? Po zaslugi financiranja projektov statističnega strojnega prevajanja s strani Evropske komisije je tekom različnih projektov postalo na voljo več javno dostopnih zbirk. Veliko od njih je objavljenih na spletni strani sistema Moses [26] ter spletni strani OPUS

[3] Vir: spletna stran Javne agencije za zdravila in medicinske pripomočke (https://www.jazmp.si/)

[47], ki je namenjena zbiranju odprtih vzporednih korpusov. Obširnejše zbirke so na voljo predvsem za večje svetovne jezike, medtem ko je količina zbirk za manjše jezike ustrezno manjša.

Kot zgled pridobivanja podatkov naj navedemo raziskavo, ki se je usmerila v pridobivanje zbirk podatkov za jezike z manj viri (*angl. under-resourced languages*). Eden od načinov, ki ga izberejo, je zbiranje dokumentov iz Wikipedije. Za kombinacijo angleški-slovenski jezik zberejo 20.351 dokumentov, vendar izvorni dokumenti vsebujejo 5 milijonov besed, medtem ko je v slovenskih dokumentih samo 2,6 milijona besed. Težave z neporavnanostjo besedil se lotijo tako, da najprej izdelajo dvojezični leksikon naslovov, ki so med seboj povezani z Wikipedijinimi medjezikovnimi povezavami. Nato z uporabo koeficienta podobnosti dokumentov pripravijo zbirko primerljivih dokumentov [44].

Kljub projektom z namenom pridobivanja zbirk je še vedno velik primanjkljaj predvsem na področju specifičnih domen. Morda bi lahko dosegli napredek, če bi krovne organizacije ali institucije uspele zbrati dvojezične zbirke od svojih članic, vendar za zdaj je treba za nekaterimi podatki še vedno rudariti.

V tabeli 4.1 je prikazan pregledni seznam zbirk, ki jih uporabimo v tem magistrskem delu.

Tabela 4.1: Pregledni seznam zbirk, uporabljenih za učenje SSP

	Št. dokumentov	**Št. segmentov**	**Ang. besed**	**Slo. besed**
Korpus EMA	1.872	1,1 mio	11,7 mio	14,6 mio
Korpus DGT	26.595	3,2 mio	71,9 mio	55,9 mio
Korpus Euparl	8.031	635.000	16,9 mio	14,6 mio
Centralna baza zdravil	2.135	1,3 mio		15,9 mio
Korpus ccGigafida	31.722	~14 mio		100 mio
Korpus ccKres	9.376	~1,4 mio		
Dokumenti UKC Ljubljana	10	~ 2000		13.981

Uporabljene zbirke pridobimo na več različnih načinov. Nekatere zbirke pridobimo z javno dostopnih mest, nekatere nam priskrbijo iz Lekarne UKC Ljubljana in JAZMP, bolj specifične pa moramo tudi izdelati. Ob imenu zbirk v tabeli 4.1 so zapisani še podatki o velikosti zbirke.

V nadaljevanju so opisane uporabljene zbirke in postopki priprave v primeru zbirk, ki jih pripravimo sami.

4.3. Dvojezične zbirke

Za statistično strojno prevajanje potrebujemo primerno količino dvojezičnih zbirk prevodov, ki služijo za učenje strojnega prevajalnika. V ta namen je za slovenski jezik na voljo kar nekaj splošnih zbirk, vendar se v primeru prilagajanja prevajalnika za specifično domeno izkaže, da gradnja izjemno velike splošne zbirke znatno ne pripomore h kakovosti prevodov s farmacevtskega področja. Pri iskanju splošnih zbirk se osredotočimo na zbirke, ki se nam zdijo najbolj kakovostne glede na njihovo vsebino in poravnavo stavkov.

Večji izziv je pridobivanje zbirk s področja farmacije. Dvojezičnih zbirk ni veliko, zato jih je treba tudi zgraditi iz datotek, ki jih dobimo na internetu ali nam jih priskrbijo pri JAZMP in Lekarni UKC Ljubljana. Prav tako z različnih virov dobimo strokovne slovarje in jih spremenimo v ustrezno obliko. V nadaljevanju so zbirke in slovarji podrobneje opisani.

4.3.1. Zbirka Evropske agencije za zdravila (European medicines agency – EMA)

Podobno kot JAZMP spremlja katera zdravila se tržijo na področju Republike Slovenije, Evropska agencija za zdravila (EMA) nadzira evropski trg in ji morajo proizvajalci zdravil predložiti ustrezne dokumente. Vsi dokumenti, ki so javne narave, so na njihovi spletni strani na voljo v obliki PDF.

Na spletni strani OPUS je na voljo dvojezična zbirka EMA iz leta 2009. Od datuma objave te zbirke so na spletni strani EMA v več jezikih objavljeni že tudi novejši dokumenti. Ti dokumenti so predvsem t. i. SmPC-ji in študije. Zato obstoječo zbirko dopolnimo z zbirko, ki jo izdelamo sami iz datotek PDF.

1.195 datotek s programom *pdftotext* pretvorimo v besedilne datoteke (.txt) in iz njih najprej odstranimo odvečne kode in kode za formatiranje. Vse dokumente nato združimo v dva dokumenta (angleški in slovenski) zaradi lažje obdelave pri nadaljnjem pretvarjanju. Dokumenta nato s pomočjo programa Open Translation Manager med seboj poravnamo na

ravni segmenta. Zaradi različnih oblik stavkov in ločil ostane določen manjši odstotek segmentov neporavnan. Poravnana dokumenta pretvorimo najprej v dvojezično zbirko (.tmx). S programom Apsic Xbench nato ločimo vzporedno zbirko na po segmentih poravnan angleški del in slovenski del. V tabeli 4.2 so prikazane podrobnosti zbirke, ki jo uporabimo.

Tabela 4.2: Podrobnosti zbirke EMA

	Št. dokumentov	**Št. segmentov**	**Ang. besed**	**Slo. besed**
Korpus EMA	1.872	1,1 mio	11,7 mio	14,6 mio

Pri pretvarjanju datotek se najprej soočimo z velikim izzivom, kako pretvoriti veliko število datotek oblike PDF, pozneje pa kako poravnati angleške in slovenske datoteke na ravni segmenta, ker se stavki med seboj ne ujemajo v razmerju 1 : 1.

4.3.2. Zbirka Generalnega direktorata za prevajanje pri Evropski komisiji (DGT)

Ta zbirka v našem prevajalnem sistemu služi kot del splošne učne zbirke. Gre za zbirko, ki jo je javno objavil Generalni direktorat za prevajanje pri Evropski komisiji iz enega od svojih največjih prevajalnih pomnilnikov EURAMIS (*European advanced multilingual information system*) [21].

V tej zbirki je združenih 26.595 dokumentov. V vseh dokumentih je 3,2 mio segmentov, sestavljenih iz skupaj 55,9 mio besed. Podrobnejši podatki so navedeni v tabeli 4.3.

Tabela 4.3: Podrobnosti zbirke DGT

	Št. dokumentov	**Št. segmentov**	**Ang. besed**	**Slo. besed**
Korpus DGT	26.595	3,2 mio	71,9 mio	55,9 mio

4.3.3. Zbirka Evropskega parlamenta (Euparl)

Tudi to zbirko izberemo kot dopolnilno splošno zbirko za naš prevajalni sistem. Sestavljena je iz zapisnikov razprav Evropskega parlamenta [12]. Na voljo je za 21 evropskih jezikov.

Izdelana je z namenom ustvariti po segmentih poravnano besedilo za statistične strojne prevajalne sisteme.

Vsebuje zapisnike Evropskega parlamenta od leta 1996 do leta 2011. Podrobnosti o vsebini zbirke so prikazane v tabeli 4.4.

Tabela 4.4: Podrobnosti zbirke Europarl

	Št. dokumentov	**Št. segmentov**	**Ang. besed**	**Slo. besed**
Korpus Euparl	8.031	635.000	16,9 mio	14,6 mio

4.4. Enojezične zbirke

Enojezične zbirke v prevajalnem sistemu potrebujemo za gradnjo jezikovnega modela. Na podlagi zbirke v ciljnem jeziku se sistem uči zaporedja izrazov v ciljnem jeziku. Večja kot je zbirka, boljše statistične rezultate lahko dosežemo.

Pri iskanju enojezičnih zbirk se v prvi vrsti odločimo za zbirke iz farmacevtske domene, dodamo pa tudi splošne korpuse v slovenskem jeziku.

4.4.1. Dokumenti, objavljeni v Centralni bazi zdravil (CBZ)

Spletno stran Centralne baze zdravil[4] urejajo Ministrstvo za zdravje, Javna agencija za zdravila in medicinske pripomočke, Zavod za zdravstveno zavarovanje Slovenije in Nacionalni inštitut za javno zdravje. Na tem spletnem mestu je objavljen seznami vseh zdravil, ki se tržijo na območju Republike Slovenije. Seznam zdravil redno posodabljajo in na spletni strani je na voljo preglednica vseh zdravil v obliki .xls. V preglednici so za nekatera zdravila objavljene povezave na dokumente PDF z opisi posameznih zdravil.

Vse dokumente s povezavami do navodil ali SmPC-jev, ki jih je 14. 4. 2016 na voljo 2.135 od skupaj 17.735 vseh oblik zdravil, s pomočjo programa wget prenesemo ter nato pretvorimo v besedilno obliko. Zbirka vsebuje 1,3 milijona segmentov in skupaj 15,9 milijona slovenskih besed.

[4] http://www.cbz.si

4.4.2. ccGigafida

Korpus ccGigafida vsebuje približno 9 % korpusa Gigafida oz. 100 milijonov besed. Njena struktura je enaka strukturi korpusa, iz katerega je nastala. To je korpus Gigafida, ki je obsežna zbirka slovenskih besedil najrazličnejših zvrsti, od dnevnih časopisov, revij do knjižnih publikacij vseh vrst (leposlovje, učbeniki, stvarna literatura), spletnih besedil, prepisov parlamentarnih govorov in podobno, vsebuje pa skoraj 1,2 milijarde besed oz. natančneje 1.187.002.502 besedi.

Gigafida je namenjena raziskovanju sodobnega slovenskega jezika na več ravneh. Tako po eni strani daje odgovore na posamezne sprotne poizvedbe, še pomembneje pa je, da daje podatke o celotni podobi slovenščine. Na ta način je danes skoraj edini razmeroma zanesljiv vir za izdelavo sodobnih slovarjev, slovnic in različnih jezikovnih priročnikov za slovenščino, uporablja pa se tudi v jezikovnih tehnologijah.

Tabela 4.5: Podrobnosti enojezične zbirke ccGigafida

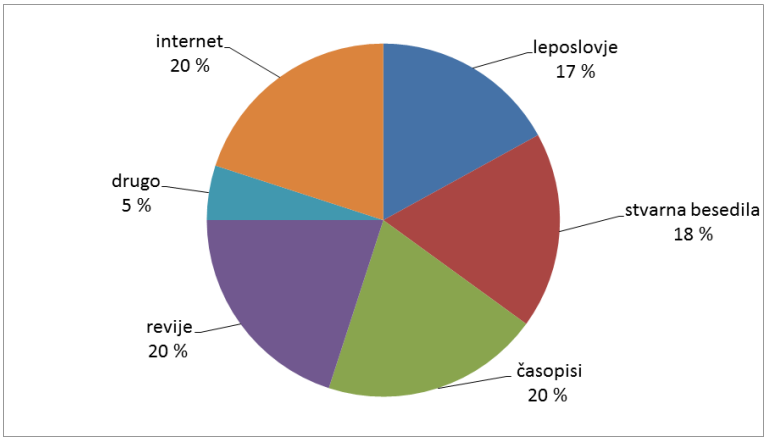
	Št. dokumentov	**Št. segmentov**	**Slo. besed**
Korpus ccGigafida	31.722	~14 mio	100 mio

Korpus vsebuje tudi druge vrste informacij. Vsak posamezni dokument, ki jih je v korpusu ccGigafida 31.722, vsebuje informacijo o viru (npr. Mladina, Delo, Dnevnik), letu nastanka, vrsti besedila (npr. leposlovje, revija), naslovu in avtorju, če je ta znan. Poleg tega sta korpusa jezikoslovno označena, kar pomeni, da sta prav vsaki besedi v korpusu pripisana še dva podatka. Prvi je osnovna oblika besede ali lema (npr. jagode, jagodi, jagodam = jagoda), drugi je t. i. oblikoskladenjska oznaka. Ta oznaka opisuje, v katero besedno vrsto spada beseda (samostalnik, glagol, pridevnik itd.) in kakšne so njene lastnosti (npr. spol, število, sklon). Ker gre za ogromne količine besedil, je označevanje potekalo povsem avtomatsko s pomočjo statističnega označevalnika Obeliks [16].

4.4.3. ccKres

V korpus ccKres je vključenih 9.376 dokumentov oz. 9% celotne zbirke Kres. Kres je iz Gigafide vzorčeni uravnoteženi podkorpus. Za korpuse, ki predstavljajo celovito podobo

nekega jezika, je ključno, da so veliki in besedilnovrstno pestri. Gigafida je tak, referenčni korpus, težko pa bi mu pripisali uravnoteženost, saj je v njem 77 % besed iz periodike (časopisi, revije) in npr. le dobrih 6 % besed iz knjig (leposlovje, stvarna besedila). Kot Gigafidin uravnoteženi podkorpus je bil izdelan 10-milijonski Kres, katerega sestavo prikazuje slika 4.1, podatke o količini podatkov pa povzema tabela 4.6.



Slika 4.1: Sestava zbirke Kreis

Podobno kot ccGigafida tudi korpus ccKres vsebuje poleg samih besedil tudi dodatne informacije o viru in slovnične značilnosti.

Tabela 4.6: Podrobnosti enojezične zbirke ccKres

	Št. dokumentov	**Št. stavkov**	**Slo. besed**
Korpus ccKres	9.376	~1,4 mio	10 mio

4.4.4. Prevedeni dokumenti Lekarne UKC Ljubljana

Lekarna UKC Ljubljana mora občasno za zdravila, ki jih naroči po decentralizirani poti, sama prevesti navodila za uporabo zdravila. Dokumente z navodili oblikuje na način, ki ustreza slovenskim predpisom glede oblike navodil za uporabo.

Za potrebe priprave sistema za strojno prevajanje nam zaupajo 10 dokumentov v obliki .doc. Te dokumente za potrebe prevajalnega modela pretvorimo v obliko čistega besedila in segmente poravnamo na ravni vrstice. Kot prikazuje tabela 4.7 lahko v domensko učno zbirko dodamo skoraj 14.000 besed.

Tabela 4.7: Podrobnosti zbirke iz dokumentov Lekarne UKC Ljubljana

	Št. dokumentov	**Št. segmentov**	**Slo. besed**
Dokumenti UKC Ljubljana	10	~ 2000	13.981

4.5. Slovarji

Slovarji igrajo v sistemu statističnega strojnega prevajanja pomembno vlogo predvsem v primeru, ko želimo, da so izrazi iz izvirnega jezika v ciljni jezik prevedeni, kot je določeno v slovarju. Več o tem opišemo v poglavju 6.

V našem delu poiščemo predvsem slovarje s področja farmacije, kemije in medicine, za katere ocenimo, da lahko znatno doprinesejo h kakovosti našega prevajalnega sistema. Poiščemo in na primeren način obdelamo naslednje slovarje:

- biokemijski slovar slovenskega biokemijskega društva,
- trojezični terminološki slovar kemijskih pojmov,
- medicinski slovar lekarne UKCLJ.

Ko vse slovarje združimo in uredimo, nastane slovar s 3.286 vnosi. V nadaljevanju te slovarje podrobneje opišemo.

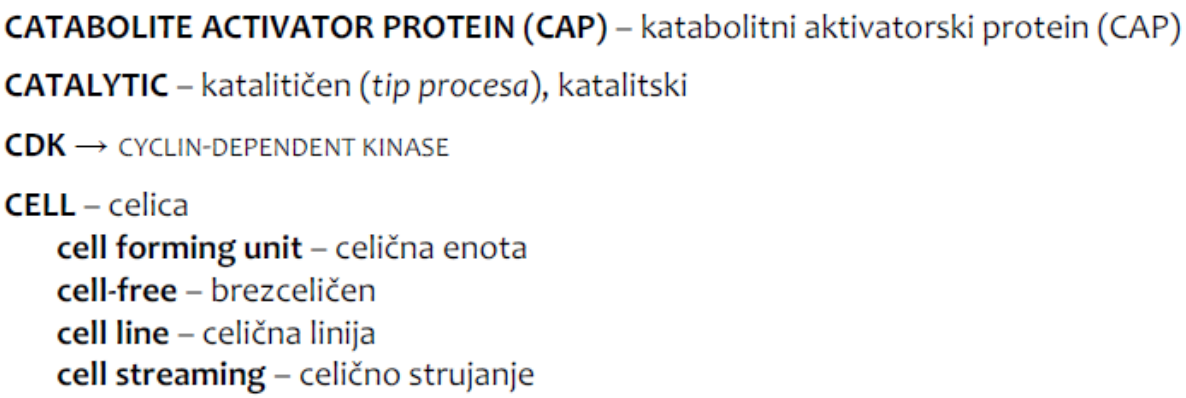
4.5.1. Biokemijski slovar slovenskega biokemijskega društva

Angleško-slovenski slovar izbranih izrazov iz biokemije in molekularne biologije je izdala terminološka komisija Slovenskega biokemijskega društva leta 2012. Delo obsega prevode angleških gesel, ki jih je vsebovala spletna različica slovarja konec leta 2009. Te so člani Terminološke komisije Biokemijskega društva znova pretehtali in uredili za tiskano različico slovarja. Ponovno so pregledali tudi uvodni splošni del, ga popravili in dopolnili. Pri tem so bili posebej pozorni na pravilno rabo slovenskega jezika, na kar jih je še posebej opozarjal jezikovni svetovalec Tomaž Sajovic. Slovar je namenjen v prvi vrsti članom Slovenskega biokemijskega društva in študentom naravoslovnih ved, ki se srečujejo z biokemijskimi in molekularnobiološkimi izrazi, v pomoč pa bo lahko tudi drugim zainteresiranim posameznikom. Razdeljen je na splošni del in slovarski del [58].

Na sliki 4.2 je prikazan izsek iz slovarja v obliki PDF, iz katerega vzamemo slovarski del in ga s pomočjo postopka OCR pretvorimo v obliko preglednice, nato pa s pomočjo razčlenjevalnika in urejevalnika preglednic uredimo v obliko, primerno za vključitev v prevajalni sistem.

Pri urejanju slovarja naletimo na težavo, ker nekateri izrazi vsebujejo samo sklic na drug izraz. Te iz slovarja odstranimo, ker ne najdemo načina, da bi sklice zamenjali z dejanskimi prevodi. Obenem ocenimo, da manjkajoči izrazi znatno ne doprinesejo k izboljšanju prevajalnega sistema.

Po končani obdelavi slovar vsebuje 1.600 izrazov.



Slika 4.2: Oblika PDF biokemijskega slovarja

4.5.2. Trojezični terminološki slovar kemijskih pojmov

Trojezični terminološki slovar kemijskih pojmov so pripravili in uredili na Oddelku za kemijo Fakultete za naravoslovje in matematiko Univerze v Mariboru. Na spletu je javno dostopen v obliki PDF in vsebuje 560 izrazov [31].

Tudi ta slovar smo s pomočjo postopka OCR pretvorili v besedilno obliko in ga s pomočjo urejevalnika preglednic uredili v primerno obliko.

4.5.3. Usklajeni izrazi Lekarne UKC Ljubljana

V Lekarni UKC Ljubljana so v toku prevajanja dokumentov za zdravila pripravili usklajen seznam pogostih besednih zvez, za katere želijo, da se uporabljajo v vseh prevodih. Ta seznam 40 besednih zvez nam posredujejo, da ga vključimo v prevajalni sistem.

Slovar je v obliki preglednice xls, ki ga priredimo za vključitev v prevajalni sistem.

4.5.4. Farmacevtski terminološki slovar

Pri pripravi farmacevtskega terminološkega slovarja, ki so ga izdali pri založbi ZRC SAZU leta 2013, je sodelovala tudi Javna agencija za zdravila in medicinske pripomočke (JAZMP). Popolna zbirka vsebuje 5.600 vnosov in poleg angleških izrazov vključuje tudi latinske izraze ter slovenski pojmovnik. JAZMP nam zaupa svoj del angleško-slovenskega slovarja s skupaj 650 izrazi. Slovar je v obliki preglednice, ki jo lahko enostavno obdelamo in prilagodimo za vključitev v prevajalni sistem.

5. Moses

V tem poglavju predstavimo sistem Moses, ki ga uporabimo za izvedbo našega eksperimenta. V prvem podpoglavju podamo njegov opis in osnovne značilnosti, v nadaljevanju pa predstavimo glavne načine delovanja, ki jih podpira.

5.1. Opis

Moses [26] je sistem za statistično strojno prevajanje, ki omogoča samodejno učenje prevajalnih modelov za poljubni jezikovni par. Za zgrajene prevajalne modele nudi učinkovit algoritem, ki med eksponentnim številom izbir hitro poišče prevod z najvišjo verjetnostjo.

Glavni sestavini Mosesa sta dekoder in komponente za učni proces.

Komponente učnega procesa

Komponente učnega procesa predstavlja zbirka orodij, ki za vhodne podatke sprejmejo surove podatke (dvojezične in enojezične) in jih spremenijo v model za strojno prevajanje. Dekoder je ena sama aplikacija, napisana v programskem jeziku C++, ki z uporabo naučenega prevajalnega modela prevede stavek v izvirnem jeziku v ciljni jezik.

Preden lahko podatke uporabimo za učenje, jih je treba pripraviti (tokenizacija, odstranjevanje ločil in spreminjanje velikih začetnic v male). Za pripravo podatkov ima Moses vgrajene skripte, ki vključujejo tudi hevristiko za odstranjevanje jezikovnih parov, ki so videti neporavnani in skrajševanje stavkov na določeno dolžino. Za poravnavo besed iz parov stavkov sta na voljo orodji GIZA++ [37] in Mgiza [14].

Pomemben del prevajalnega sistema je jezikovni model, s katerim dekoder skuša zagotoviti gladkost prevoda. Za gradnjo jezikovnega modela Moses uporablja zunanja orodja SRILM [45], IRSTLM [13] in KENLM [19].

Zadnji korak pri pripravi strojnega prevajalnega sistema je optimizacija (*angl. tuning*) modela. V tem koraku so različni statistični modeli medsebojno primerjani in uteženi, da proizvedejo najboljši možni prevod. Moses za optimizacijo modelov podpira orodja z uporabo algoritmov MERT (*Minimum Error Rate Training*) [36] in MIRA (*Margin Infused Relaxed Algorithm*) [17].

Za ocenjevanje prevodov uporablja metriko BLEU (*Bilingual Evaluation Understudy*) [39].

Dekoder

Naloga Mosesovega dekoderja je, da v ciljnem jeziku poišče stavek z najvišjim rezultatom (v skladu s prevajalnim modelom), ki ustreza danemu izvornemu stavku. Dekoder lahko izpiše tudi rangirani seznam kandidatov in različne informacije o tem, kako je prišel do svoje odločitve. Mosesov dekoder podpira tudi večnitno dekodiranje ter ima na voljo skripte za večprocesno dekodiranje, če imamo dostop do gruče (*angl. cluster*) strežnikov [26].

5.2. Glavni načini delovanja

Moses omogoča prevajanje na več različnih načinov glede na naše potrebe. Glavne načine lahko razvrstimo v tri glavne:

- model na osnovi besednih zvez,
- model na osnovi skladnje,
- faktorski model.

5.2.1. Model na osnovi besedni zvez

Za prevajanje na osnovi besednih zvez je model sestavljen iz dveh datotek. To sta prevajalna tabela besednih zvez in konfiguracijska datoteka za dekoder (moses.ini).

Prva vrstica prevajalne tabele besednih zvez (datoteka phrase-table) je videti, kot prikazuje slika 5.1.

```
pes ||| dog ||| 0.3 ||| |||
```

Slika 5.1: Prva vrstica prevajalne tabele besednih zvez

Vnos na sliki 5.1 pomeni, da je verjetnost prevajanja angleške besede dog v slovensko besedo pes enaka 0,3. Matematično zapisano to pomeni $P(dog|pes) = 0{,}3$.

V prevajalnih tabelah niso zapisane samo besede, ampak tudi besedne zveze, vendar brez dodatnih slovničnih informacij. Prevajalne tabele so izvor glavnega znanja za dekoder strojnega prevajalnika, ki jih uporabi pri prevajanju. Med izvajanjem dekoderja lahko z dodatnima stikaloma `-report-segmentation` in `-verbose` izpišemo dodatne informacije o prevajanju dekoderja.

Prevajanje v tem načinu omogoča tudi optimizacijo za hitrost, ki jo lahko dosežemo z omejevanjem iskalnega prostora dekoderja ter optimizacijo za kakovost. Ključnega pomena za kakovosten prevod je, da imamo dobro prevajalno tabelo besednih zvez, nekaj pa je mogoče doseči tudi optimizacijo parametrov modela. Verjetnost, ki je dodeljena prevodu, je produkt verjetnosti štirih modelov, od katerih vsak prispeva informacije glede posameznih značilnosti dobrega prevoda:

- prevajalna tabela besednih zvez zagotavlja, da so besede v izvornem in ciljnem jeziku dobri medsebojni prevodi;
- jezikovni model zagotavlja, da se prevod bere tekoče v ciljnem jeziku;
- model popačenja omogoča preurejanje vhodnega stavka;
- kazen za besede omogoča, da prevodi niso predolgi ali prekratki.

Za vsako komponento je mogoče dodeliti utež, ki določa njeno pomembnost. Matematično zapisano je cena enaka:

$$p(e|f) = \Phi(f|e)^{w_\Phi} * \mathrm{LM}(e)^{w_{\mathrm{LM}}} * \mathrm{D}(e|f)^{w_\mathrm{D}} * \mathrm{W}(f|e)^{w_\mathrm{W}}$$

Verjetnost $p(e|f)$ prevoda v ciljnem jeziku *e*, če je *f* stavek v izvornem jeziku, je razdeljena med štiri modele, ki so prevod besednih zvez $\Phi(f|e)$, jezikovni model LM(*e*), model popačenja $\mathrm{D}(e|f)$ in kazen za besede W(*e*). Vsak od štirih modelov je utežen z utežjo.

Uteži lahko dekoderju podamo s štirimi parametri: `weight-t, weight-l, weight-d` in `weight-w`. Privzeta nastavitev za te uteži je 1,1,1 in 0. To so tudi vrednosti v konfiguracijski datoteki moses.ini. Z nastavljanjem teh uteži na prave vrednosti lahko izboljšamo kakovost prevoda.

Kakšne so prave vrednosti uteži, je odvisno od zbirke in jezikovnega para. Najlažji pristop je, da poskušamo več uteži in vidimo, kaj nam prinese najboljše rezultate. Dobre vrednosti uteži za prevajalno tabelo besednih zvez (`weight-t`, kratko `tm`), prevajalni model (`weight-l`, kratko `lm`) in model popačenja – preurejanja (`weight-d`, kratko `d`) so 0.1-1, dobri vrednosti za kazen za besede (`weight-w`, kratko `w`) sta -3-3. Negativne vrednosti za kazen za besede preferirajo daljše prevode, medtem ko pozitivne vrednosti preferirajo krajše prevode.

5.2.2. Model na osnovi skladnje

Moses podpira tudi modele, ki jih različni avtorji imenujejo hierarhični modeli ali modeli na osnovi skladnje. Ti modeli uporabljajo slovnico, ki je sestavljena iz pravil sinhrone brezkontekstne slovnice (*SCFG – Synchronous Context-Free Grammar*). V Mosesu ta model zaradi svoje strukture imenujejo drevesni modeli (*angl. tree-based models*).

Za razliko od tradicionalnih modelov na osnovi besednih zvez, kjer je prevajanje izvedeno s preslikavo vhodne besedne zveze v izhodno besedno zvezo, poteka pri drevesnih modelih prevajanje s pomočjo t. i. slovničnih pravil. Pravila vključujejo spremenljivke v pravilih preslikovanja. Pri dekodiranju z modeli na osnovi besednih zvez poteka generiranje stavka od leve proti desni z dodajanjem besednih zvez na konec delnega prevoda. Med dekodiranjem se pri drevesnih modelih zgradi diagram, ki je sestavljen iz delnih prevodov za vse možnosti vhodnega stavka [26].

Za zagon dekodiranja potrebujemo datoteko `moses.ini`, ki kaže na lokacijo jezikovnega modela in datoteko s tabelo pravil `rule-table`. Vsaka vrstica v tabeli pravil opisuje eno prevajalno pravilo. Med dekodiranjem lahko tudi v tem načinu omogočimo sledenje in dobimo vpogled v to, kako je bil prevod sestavljen.

5.2.3. Faktorski model

Kot omenimo v prejšnjih poglavjih, faktorski model omogoča vgradnjo dodatnih parametrov, s katerimi lahko v prevod vnesemo dodatne informacije o besednih zvezah.

Za gradnjo faktorskega modela v Mosesovem vodiču [25] predstavijo vzorec učnih podatkov. Na sliki 5.2 lahko vidimo, da je vsaka beseda predstavljena s svojo površinsko obliko in tudi dodatnimi faktorji.

Faktorji za nemščino so:

- površinska oblika,
- lema,
- besedna vrsta,
- besedna vrsta z dodatnimi oblikoskladenjskimi oznakami.

```
% tail -n 1 factored-corpus/proj-syndicate.??
 ==> factored-corpus/proj-syndicate.de <==
 korruption|korruption|nn|nn.fem.cas.sg floriert|florieren|vvfin|vvfin .|.|per|per

 ==> factored-corpus/proj-syndicate.en <==
 corruption|corruption|nn flourishes|flourish|nns .|.|.
        X1 X2 -> X1 X2
```

Slika 5.2: Primer učne zbirke za faktorski model

Faktorji za angleščino so:

- površinska oblika,
- lema,
- besedna vrsta.

Faktorski prevajalni modeli omogočajo izdelavo modelov, ki izvedejo morfološko analizo in dekompozicijo med postopkom prevajanja. Tak model naučimo v štirih korakih:

- prevajalni korak, ki preslika leme;
- generativni korak, ki nastavi možne besedne vrste za lemo;
- prevajalni korak, ki preslika morfološke informacije v oznake besednih vrst in
- generativni korak, ki preslika oznako besedne vrste v površinsko obliko.

Moses podpira več načinov uporabe faktorskega modela, od naših podatkovnih zbirk pa je odvisno, kateri pristop bomo izbrali.

6. Empirično ovrednotenje strojnega prevajalnika

Z empiričnim ovrednotenjem želimo spoznati kakovost prevodov in rezultate, ki jih dobimo pri prilagajanju statističnega strojnega prevajanja za farmacevtsko domeno v slovenskem jeziku. V tem poglavju opišemo ovrednotenje z različnimi modeli prevajanja. V prvih dveh poglavjih predstavimo programe, uporabljene pri delu, in pripravo zbirk za učenje, v nadaljevanju pa različne pristope k prilagajanju domeni. V okviru modela na osnovi besednih zvez najprej naredimo preizkus prilagajanja prevajalnega modela z uporabo različnih zbirk. Predvidevamo, da ima vpliv na kakovost prevoda tudi strokovni slovar, zato naredimo preskus, ki vključuje slovar. V načinu na osnovi besednih zvez predstavimo še pristop s transduktivnim prevajanjem angleške domenske zbirke in opazujemo vpliv na kakovost prevoda testnega dokumenta. V naslednjem koraku preizkusimo vpliv prilagajanja jezikovnega modela. V zadnjem delu opišemo še pristop k prevajanju s faktorskim modelom in uglaševanje parametrov za optimizacijo.

6.1. Uporabljeni programi

Za eksperimente z modelom na osnovi besednih zvez uporabimo ogrodje statističnega strojnega prevajalnega sistema Moses, ki ga podrobneje opišemo v poglavju 5. Spodaj so na kratko opisani glavni uporabljeni Mosesovi programi ter ostali programi, ki jih uporabimo.

Dekoder Moses: testne dokumente prevedemo z dekoderjem Moses [26], ki z uporabo prevajalne tabele večbesednih zvez in jezikovnim modelom prevaja stavke in pri tem upošteva nastavitve v inicializacijski datoteki.

GIZA++: prevajalno tabelo izdelamo z orodjem GIZA++ [37], ki za učenje prevajalnih modelov na ravni besede uporablja poravnane vzporedne korpuse. GIZA++ je izvedba IBM-ovih modelov in daje prednost poravnavi na ravni besed med stavkom v izvirnem in stavkom v ciljnem jeziku.

KenLM: jezikovni model zgradimo s programom KenLM. Odlikuje ga hitrost in nizka poraba pomnilnika, zato omogoča gradnjo velikih jezikovnih modelov. Po potrebi lahko programu s parametrom omejimo porabo pomnilnika, ki v tem primeru izvede sortiranje z združevanjem vsebine na disku in v pomnilniku.

MERT: uglaševanje prevajalnega sistema izvedemo s programom MERT (*Minimum Error Rate Training*) [36]. MERT uporablja vnaprej izračunan jezikovni model in niz verjetnostnih poravnav, s katerimi optimizira uteži posameznih funkcij, da maksimira rezultat BLEU za celoten sistem glede na referenčni niz.

TreeTagger: označevalnik TreeTagger [41] je orodje za označevanje besedil z informacijami o besedni vrstah in lemah. Uporabimo ga za označevanje angleških in slovenskih zbirk.

Označevalnik Obeliks [16]**:** je orodje za oblikoslovno označevanje slovenskega jezika, izdelan v okviru projekta Sporazumevanje v slovenskem jeziku. Sestavljen je iz treh komponent: iz segmentacijskega in tokenizacijskega modula, ki besedilo razdeli na stavke in besede, samega oblikoslovnega označevalnika, ki besedam pripiše besedno vrsto in njene lastnosti, ter lematizatorja, ki jim pripiše njihovo osnovno obliko. Uporabimo ga za označevanje slovenskih zbirk.

Pdftotext: je odprtokodni program za pretvarjanje datotek PDF v datoteke z golim besedilom. Je del več distribucij operacijskega sistema Linux. Omogoča več parametrov, s katerimi nastavimo način pretvarjanja datotek. S programom lahko naenkrat pretvorimo eno datoteko, s pomočjo skripta pa lahko v ukaznem pozivu pretvorimo tudi več tisoč datotek naenkrat [52].

Wget: odprtokodni program za prenos datotek z oddaljenega mesta. Je del distribucije Linux.

Awk: odprtokodni program (ime je okrajšava za avtorje Aho, Weinberger in Kernighan), ki je namenjen obdelavi besedila. Zasnovan je bil za iskanje vzorcev v besedilnih podatkih in je zelo uporaben za razčlenjevanje sistemskih podatkov. Uporabimo ga za naključno krajšanje domenske zbirke podatkov.

Open Translation Manager (OTM): v programu Open Translation Manager [38] je na voljo pripomoček za poravnavo izvornih in prevedenih datotek na ravni segmenta oziroma stavka, ki ga uporabimo za poravnavo datotek in pripravo korpusov.

Apsic Xbench [2]: ta program uporabimo za pretvarjanje jezikovnih zbirk iz oblike zapisa TMX v obliko zapisa, primerno za učenje jezikovnih modelov.

6.2. Priprava zbirk podatkov

Za učenje strojnega prevajalnika moramo zbirke ustrezno pripraviti. Za vse zbirke velja, da je treba izvesti naslednje postopke:

- **Tokenizacijo**, s čimer ločimo besedilo na besede in ločila.
- **Prilagoditev velikih in malih začetnic**, s katero začetna beseda v vsakem stavku dobi naverjetnejšo začetnico, v večini primerov so začetnice vseh besed spremenjene v male črke.
- **Čiščenje,** s katerim so dolgi stavki skrajšani, prazni stavki pa odstranjeni, saj sicer povzročajo težave med učenjem. Prav tako so odstranjeni neporavnani stavki.

Posebno pozornost potrebujejo zbirke za faktorski model prevajanja. Če zanj nimamo na voljo primerno označene zbirke, jo moramo v času priprave zbirke pripraviti z uporabo ustreznega označevalnika.

Za uspešno učenje prevajalnika moramo označiti tako angleško kot tudi slovensko učno zbirko. Za označevanje angleške zbirke uporabimo program TreeTagger, ki angleško zbirko zelo hitro označi, površinski obliki pa doda oblikoskladenjsko oznako in osnovno obliko (lema). Izhodni podatki so v obliki zapisa, ločenega s tabulatorji, ki še ni primerna za učenje prevajalnika v Mosesu. Slabost tega programa je, da se v izhodnih podatkih izgubijo podatki o koncu vrstic. Zato pred obdelavo v zbirko na konec vsake vrstice dodamo oznako <*novavrst*>. Zbirko po koncu označevanja pretvorimo v obliko, primerno za Moses, z nizom ukazov, ki jih združimo v skripto *ttpos-to-moses.sh*. Skripta je prikazana v prilogi 3.

Slovenska zbirka ccGigafida je označena z označevalnikom Obeliks (*PosTaggerTag*), zato ta program uporabimo tudi mi za označevanje slovenske domenske zbirke. Obeliks vsem besedam v slovenski zbirki poleg površinske oblike doda oblikoskladenjsko oznako in osnovno obliko. Oblika zapisa izhodnih podatkov je XML, ki jo pretvorimo v obliko, primerno za Moses. To naredimo s skripto *xml-to-moses.pl*, ki je prikazana v prilogi 4.

Označevanje besedil v slovenskem jeziku podpira tudi označevalnik TreeTagger, ki je hitrejši od Obeliksa, ni pa tako natančen, saj je bil zasnovan v poznih 90-ih letih. Osnovna namestitev TreeTaggerja za slovenščino sprva ne deluje, zato modelne datoteke zamenjamo z novejšimi, ki jih dobimo na spletni strani SketchEngine [23]. Po tej zamenjavi program uspešno označi zbirko v slovenskem jeziku.

6.3. Prilagajanje prevajalnega modela

V tretjem poglavju predstavimo več možnosti za prilagajanje jezikovnega modela za specifično domeno, od katerih smo glede na razpoložljivost virov nekatere prilagodili tudi v tem delu. Na voljo imamo splošne zbirke in manjšo količino domenskih dvojezičnih podatkov. V eksperimentu prilagajanja prevajalnega modela za prvi eksperiment uporabimo smernice iz raziskave avtorjev Koehn in Schroeder [28], ki sta analizirala vpliv uporabe različnih zbirk na kakovost prevoda znotraj določene domene. V nadaljevanju nas po vzoru raziskave, ki so jo izvedli Wu in sod. [55], zanima, ali se kakovost prevoda izboljša, če s pristopom transduktivnega prevajanja strojno prevedemo domensko zbirko iz angleškega v slovenski jezik in tako povečamo dvojezično domensko zbirko in s tem tudi prevajalni model. Je velikost učne zbirke res pomembna? V zadnjem delu predstavimo način, s katerim iščemo odgovor na to vprašanje.

6.3.1. Prilagajanje s kombinacijo različnih učnih zbirk

V tem razdelku opišemo gradnjo prevajalnih modelov z uporabo različnih učnih zbirk. S postopnim približevanjem domeni želimo ugotoviti vpliv količine podatkov in specifičnosti na kakovost prevoda. V ta namen vzpostavimo naslednje tri modele:

- samo splošni učni podatki,
- samo domenski učni podatki,
- združeni splošni in domenski podatki.

Spodaj opišemo primere prilagajanja prevajalnega modela z različnimi zbirkami. Jezikovni model je vseh primerih zgrajen z uporabo slovenskega dela posamezne dvojezične zbirke.

6.3.1.1. Samo splošni učni podatki

Za učenje prvega prevajalnega sistema uporabimo samo splošni učni korpus, sestavljen iz zbirk Euparl in DGT. V tabeli 6.1. je prikazana količina podatkov, ki jih zbirki vsebujeta. S tem želimo ugotoviti razliko v kakovosti in ocenah v primerjavi z modeli, ki so bližje ciljni domeni.

Tabela 6.1: Povzetek količine podatkov za splošni model

	Št. segmentov	Ang. besed	Slo. besed
Korpus EUParl	0,6 mio	16,9 mio	14,6 mio
Korpus DGT	3,2 mio	71,9 mio	55,9 mio
Skupaj:	3,8 mio	88,8 mio	70,5 mio

6.3.1.2. Samo domenski učni podatki

Za učenje drugega prevajalnega sistema uporabimo domenski učni korpus. Zgrajen je iz dokumentov Evropske agencije za zdravila (EMA), ki jih pretvorimo po postopku, opisanem v poglavju 4.3.1. Ta zbirka je manjša in vsebuje 1,1 milijona stavkov oz. 11,7 milijona angleških ter 14,6 milijona slovenskih besed.

Predvidevamo, da z domensko učno zbirko dosežemo boljše rezultate kot s splošno, vendar nismo prepričani koliko, glede na to, da je ta učna zbirka precej manjša od naše splošne.

6.3.1.3. Združeni splošni in domenski podatki

Da čimbolj izkoristimo učne podatke, združimo vse razpoložljive podatke v eno samo učno zbirko, iz katere izdelamo prevajalni model. Združena zbirka vsebuje 3,9 milijonov stavkov, 100,5 milijona angleških in 85,1 milijona slovenskih besed.

6.3.2. Prilagajanje z dodajanjem slovarja

Pri tem pristopu želimo izvedeti, kakšen vpliv na kakovost prevoda ima dodajanje slovarja v učno zbirko in v prevajalno tabelo. V zmanjšan model, ki uporablja domensko zbirko, dodamo še združen slovar terminoloških izrazov, ki vsebuje 3.300 izrazov.

6.3.3. Prilagajanje s transduktivnim prevajanjem

Predpostavljamo, da z večjo domensko zbirko lahko dosežemo boljše BLEU rezultate pri prevajanju testnega dokumenta. S pristopom transduktivnega prevajanja nameravamo prevesti zbirko PubMed in jo dodati obstoječi domenski učni zbirki.

Celotna zbirka PubMed vsebuje več kot 26 milijonov citatov iz biomedicinske literature. Vsak citat je shranjen v svoji datoteki, ki je poimenovana z imenom revije in izdaje, v kateri je bil objavljen. Zbirka je razdeljena v več delov, od katerih vsak del vsebuje datoteke dveh zaporednih začetnih črk datotek.

Ker je zbirka res velika, se odločimo, da za ta eksperiment uporabimo petino dokumentov, ki se začnejo na črko A. Združeni dokument je dolg približno 500 MB in vsebuje več kot 8,3 milijona vrstic.

6.4. Prilagajanje jezikovnega modela

Eden od načinov, da prilagodimo prevajalni sistem ciljni domeni, je z uporabo jezikovnega modela. Bistvenega pomena pri prilagajanju jezikovnega modela je, da za učenje uporabimo dovolj veliko enojezično učno zbirko, ki vsebuje besedila s ciljnega področja. Ker imamo v Sloveniji na voljo več splošnih enojezičnih zbirk slovenskih besedil, želimo preveriti vpliv ene od teh zbirk na kakovost prevoda v kombinaciji z domensko zbirko.

Različne izvedbe jezikovnega modela pri dekodiranju uporabimo v kombinaciji z vsemi tremi prevajalnimi modeli. Jezikovni model zgradimo s programom KenLM. Za potrebe našega eksperimenta zgradimo 3-gramski jezikovni model, za katerega pozneje med pregledom ugotovimo, da ima v besedne zveze vključene tudi oblikovne kode, kot denimo <b>, kar pa bistveno ne vpliva na gradnjo jezikovnega modela.

6.4.1. Jezikovni model, izdelan z zbirko ccGigafida

Pri tem poskusu jezikovni model zgradimo iz besedilnega korpusa ccGigafida. Ugotoviti želimo, kakšen vpliv ima na kakovost prevoda jezikovni model, zgrajen iz velikega korpusa slovenskega jezika. Ker je korpus Gigafida največji slovenski korpus, želimo ta jezikovni model ponuditi tudi drugim raziskovalcem slovenskega jezika.

Priprava korpusa traja 75 minut, medtem ko gradnja modela in pretvarjanje v dvojiški format zahteva še 150 minut. Korpus ccGigafida vsebuje približno 14 milijonov stavkov in 100 milijonov besed.

6.4.2. Jezikovni model, izdelan z domensko zbirko

Za gradnjo domenskega jezikovnega modela združimo vse slovenske dokumente, povezane s farmacijo, ki jih lahko dobimo. Kot prikazuje tabela 6.2, v gradnjo jezikovnega modela vključimo slovenski del korpusa EMA, dokumente, do katerih smo prišli prek povezav na spletni strani centralne baze zdravil[5], ter dokumente, ki smo jih dobili od Lekarne UKC Ljubljana.

Tabela 6.2: Povzetek količine podatkov v domenski enojezični zbirki

	Št. segmentov	**Slo. besed**
Korpus EMA	1,1 mio	14,6 mio
Centralna baza zdravil	1,3 mio	15,9 mio
Dokumenti UKC Ljubljana	~ 2000	13.981
Skupaj:	1,4, mio	30,6 mio

6.4.3. Kombiniran jezikovni model

Kombinirani jezikovni model izdelamo z uporabo slovenskega dela zbirke Euparl, domenske zbirke in dokumentov UKC Ljubljana. Nastane zbirka, ki vsebuje 4,5 milijona stavkov oz. 44,6 milijona slovenskih besed.

6.5. Faktorski model

V okviru prilagajanja strojnega prevajalnika za specifično domeno raziščemo tudi vpliv označenih učnih zbirk z uporabo faktorskega modela. Pomembna razlika od modela na temelju besednih zvez je, da moramo pri faktorskem modelu za učenje prevajalnega modela uporabiti zbirke, ki so označene z dodatnimi parametri, ki jih imenujemo faktorji. Za slovenščino sta na voljo označeni zbirki ccKres in ccGigafida, ostale pa moramo označiti sami z enim od programov za označevanje. Označevanje podrobneje opišemo v poglavju 6.2.

[5] http://www.cbz.si

Podobno kot pri prilagajanju prevajalnega modela tudi pri prilagajanju faktorskega modela eksperiment začnemo s splošno zbirko, nadaljujemo z domensko zbirko, domenski zbirki dodamo jezikovni model ccGigafida in na koncu kombinirano zbirko.

Za gradnjo faktorskega prevajalnega modela podamo Mosesu poleg dveh označenih učnih zbirk tudi dva jezikovna modela. Prvi je enak jezikovnemu modelu pri načinu na osnovi besednih zvez. Vsebuje 3-gramski model verjetnosti površinskih oblik besed. Tega zgradimo s programom GIZA++ iz neoznačene zbirke.

Drugi jezikovni model zgradimo iz oblikoskladenjskih oznak, ki jih iz označene slovenske zbirke prestavimo v ločeno zbirko. Če zbirko označimo s programom Obeliks, skripto *xml-to-moses.pl* spremenimo tako, da v izhodno datoteko izpisuje samo oblikoskladenjske oznake. Iz te datoteke nato s programom GIZA++ izdelamo jezikovni model. Zbirke, označene z označevalnikom TreeTagger, pretvorimo s skripto *convert-to-lm.sh*, ki iz označene datoteke izdela oblikoskladenjski jezikovni model.

Ko imamo na voljo označeni zbirki in oba jezikovna modela, sprožimo gradnjo faktorskega prevajalnega modela. Pri gradnji lahko podamo več korakov prevajanja in generiranja. V našem primeru se odločimo za prevajanje angleške površinske oblike v slovensko površinsko obliko in oblikoskladenjsko oznako ter nato dekodiranje v površinsko obliko. Za tak postopek se odločimo, da dobi prevajalnik dodatne informacije o besedi, ki jo obdeluje. Rezultate prevajanja s faktorskim modelom predstavimo v poglavju 7.2.5.

6.6. Uglaševanje parametrov in optimizacija

Kakovost prevoda pri statističnem strojnem prevajanju je v veliki meri odvisna od ustrezno nastavljenih parametrov. Pomembnejših parametrov je več. Če sledimo zaporedju korakov pri gradnji prevajalnega modela, se s parametri srečamo pri pripravi učnih zbirk, ko omejujemo dolžino stavkov, pri gradnji jezikovnega modela z nastavitvijo n-gramov in pri optimizaciji prevajalnega modela, ki jo izvedemo po končani gradnji prevajalnega modela. Pri faktorskem modelu so parametri izbrani faktorji. O nastavljanju parametrov nekaj več zapišemo v nadaljevanju.

6.6.1. Nastavljanje dolžine stavkov v zbirki

V vsaki zbirki, ki jo želimo uporabiti kot učno zbirko za sistem strojnega prevajanja, so stavki oz. vnosi v vsaki vrstici zbirke različne dolžine. Nekatere vrstice so zelo kratke, medtem ko so nekateri stavki lahko tudi zelo dolgi. Teoretično so dolgi stavki zelo koristni, saj vsebujejo večjo količino informacij, težava pa nastopi, da zelo dolgi stavki podaljšajo poravnavo besedil s programom GIZA++ in si zato zelo dolgih stavkov v učnih zbirkah ne želimo, če želimo dobiti prevajalni model v razumnem času.

Še vedno pa ostane vprašanje, kakšna dolžina stavkov je optimalna. Koehn in sod. [28] so ugotovili, da povečanje dolžine stavkov s 40 na 80 besed daje boljše rezultate, ne da bi to preveč podaljšalo čas obdelave. Po njihovi raziskavi smo se zgledovali tudi v našem eksperimentu.

6.6.2. Nastavljanje n-gramov pri gradnji jezikovnega modela

Pri gradnji jezikovnega modela se odločimo, kako velik jezikovni model želimo zgraditi s tem, da podamo do koliko besed naj bo vključenih v besedno zvezo. 3-gramski model bo v jezikovni model dodal besedne zveze z enim, dvema in tremi besedami.

Z vidika razmerja med časom gradnje modela ter časom prevajanja in pokritja jezikovnih pomenov lahko rečemo, da je 3-gramski model optimalna izbira. S poznejšim eksperimentom ugotovimo, da testni prevod, preveden s 4-gramskim jezikovnim modelom, doseže za 1 boljšo oceno BLEU od 3-gramskega, vendar na račun daljše gradnje.

6.6.3. Uglaševanje parametrov prevajalnega modela

Po končani gradnji prevajalnega modela se v inicializacijsko datoteko *moses.ini* zapišejo privzete vrednosti parametrov modela, ki so razdeljeni v naslednje skupine:

- kazen za neznane besede,
- kazen za besede,
- jezikovni model,
- prevajalni model,
- model popačenja.

Ti parametri so v sistemu nastavljeni glede na izkušnje iz gradnje več prevajalnih modelov, kar pa ni nujno dobro za naš primer. Z uglaševanjem parametrov za prevajanje v slovenski jezik se

v svojem magistrskem delu ukvarja Dugonik [11]. Primerja uglaševanje z algoritmi MERT [36], MIRA [17], PRO [11] in lastnim algoritmom. Rezultati pokažejo, da lastni algoritem z diferencialno evolucijo prinese večje izboljšanje ocene BLEU, vendar za obdelavo potrebuje 200 ur časa v primerjavi z drugimi tremi, ki za sicer manjše izboljšanje prevoda potrebujejo od 1 do 3 ur. Izmed ostalih treh najboljši rezultat doseže algoritem MERT, ki oceno BLEU 92,40 % dvigne za 1,45 % na 93,85 % v času ene ure.

Na podlagi teh rezultatov v našem eksperimentu uporabimo optimizacijo z algoritmom MERT.

MERT poišče uteži, ki minimizirajo napako ali maksimirajo določeno metriko. Glavna značilnost tega algoritma je, da izhaja iz n-najboljših prevodov za podan stavek. S tem dosežemo hitro konvergenco optimizacijskega procesa. Najpogosteje se uporablja metrika BLEU, lahko pa se uporabi poljubna metrika. MERT se lahko ujame v lokalni maksimum in ker uporablja sezname n-najboljših kot približek za izhod dekodirnika, ne more raziskati dejanskega prostora parametrov. Kljub tem omejitvam daje MERT dobre rezultate [11].

V prilogi 3 je prikazan parametrski del datoteke *moses.ini* pred uglaševanjem in po njem.

7. Ocenjevanje prevodov in rezultati

Različni modeli prevajanja generirajo vsak svoj prevod testnega dokumenta. V tem poglavju najprej predstavimo testni dokument in pristope k ocenjevanju prevodov, v drugem delu pa rezultate ocenjevanja in naše ugotovitve.

7.1. Ocenjevanje prevodov

Za ocenjevanje prevodov pripravimo testni dokument, ki nam pokaže več lastnosti pristopov k prilagajanju za specifično domeno. V nadaljevanju najprej opišemo testni dokument, nato pa še uporabljene pristope k ocenjevanju testnega dokumenta. Slednje izvedemo z uporabo dveh metod – samodejne in človeških ocenjevalcev.

7.1.1. Testni dokument

Testni dokument izdelamo iz delov treh, v preteklosti že prevedenih farmacevtskih besedil. Za vnaprej prevedene dokumente se odločimo, da lahko izračunamo samodejno oceno kakovosti prevoda z algoritmom BLEU. Od treh dokumentov je eden vsebovan v učni zbirki, s čimer želimo preveriti, kako statistični strojni prevajalnik prevede besedila, ki jih že ima v svoji zbirki. Dela ostalih dveh dokumentov nista vsebovana v učni zbirki in predstavljata vsebino povzetkov zdravil, izdanih v preteklem letu. Vsebino celotnega testnega dokumenta si je mogoče ogledati v prilogi 1.

7.1.2. Samodejno ocenjevanje prevodov

Ob koncu prevajanja vsakega dokumenta želimo dobiti oceno, kako kakovosten je dobljeni prevod. Prvi način ocenjevanja je samodejen in hitrejši, saj oceno lahko dobimo takoj po koncu prevajanja, tako da zaženemo program `multi-bleu.perl`. S tem programom zaženemo algoritem BLEU, ki primerja vsebino dobljenega prevoda z vsebino referenčnega, vnaprej prevedenega dokumenta. Ta način podrobneje opišemo v poglavju 2.5.3.

7.1.3. Ročno ocenjevanje prevodov

Ker samodejna ocena daje zgolj referenčne ocene in je lahko včasih prevod s sicer nižjo oceno BLEU v praksi bolj berljiv ali natančnejši, se odločimo, da izvedemo tudi ocenjevanje prevodov s pomočjo strokovnjakov.

Za pripravo vprašalnika za subjektivno oceno se opremo na smernice združenja TAUS [46], ki so podrobneje opisane v poglavju 2.5.1. Mersko lestvico iz štiristopenjske razširimo na petstopenjsko, da po Likertovi lestvici dobimo še srednjo vrednost, ki je v TAUS-ovi ni, in na ta način ocenjevalcem omogočimo lažje odločanje.

Pri ročnem ocenjevanju sodeluje 6 ocenjevalcev z vsaj 4 leti izkušenj na področju farmacevtskih besedil. Dva ocenjevalca sta predstavnika sodelujočih institucij, Lekarne UKC Ljubljana in JAZMP, ki sta strokovnjaka na področju farmacije in podata svoji oceni z vidika končnega uporabnika strojnega prevoda. Oba sta bila že prej vključena v prevajanje različnih farmacevtskih dokumentov. Ostali štirje ocenjevalci so prevajalci z večletnimi izkušnjami s področja prevajanja farmacevtskih besedil.

Za ročno oceno strojnega prevoda iz 6 testnih prevodov, ki so bili prevedeni vsak s svojim prilagoditvenim modelom, vzamemo vzorec šestnajstih povedi. Teh šestnajst povedi z uporabo spletne ankete razdelimo na šestnajst vprašanj, pod vsako vprašanje pa združimo enake povedi, prevedene z različnimi modeli, da imajo ocenjevalci boljši pregled in možnost primerjave prevodov. Vzorci testnih prevodov in celoten testni dokument so predstavljeni v poglavju 7.1.1. Ocenjevalci za vsako poved podajo dve oceni. Najprej z ocenami od 1 do 5 vrednotijo berljivost prevoda. Pri berljivosti nas ne zanima semantična pravilnost prevoda, temveč izključno slovnična, npr. ujemanje spola, sklona in števila med besednimi vrstami, kohezivnost povedi oziroma segmenta ipd. Ocenjevalce prosimo, da na lestvici od 1 do 5 ocenijo, v kolikšni meri govorec maternega jezika prepozna strojni prevod kot naravnega in je zanj razumljiv, v kolikšni meri je prevod slovnično pravilen brez črkovalnih napak in koliko upošteva splošno sprejeto rabo izrazov.

Ko ocenjevalci ocenijo berljivost pri vseh povedih, nadaljujejo z ocenjevanjem pomenske ustreznosti. Tokrat imajo ocenjevalci poleg že ocenjenih povedi vpogled še v njihov izvirnik. Z ocenami od 1 do 5 ocenjujejo ustreznost, tj. semantično pravilnost prevoda v primerjavi z izvirnikom. Pri ocenjevanju ustreznosti ocenjevalce prosimo, naj primerjajo strojni prevod z angleškim izvirnikom in ocenijo, koliko pomena, izraženega v izvornem besedilu, je izraženega

tudi v ciljnem strojnem prevodu. Pri ocenjevanju ustreznosti nas torej zanima predvsem pomenska zvestoba.

Ocenjevalci berljivost in ustreznost ocenijo z oceno od 1 do 5. V tabelah 7.1 in 7.2 sta prikazani lestvici, po katerih so ocenjevalci vrednotili prevode.

Tabela 7.1: Orientacijska lestvica za ocenjevanje berljivosti

5	Brezhiben
4	Dober
3	Sprejemljivo berljiv
2	Slabo berljiv
1	Nerazumljiv

Tabela 7.2: Orientacijska lestvica za ocenjevanje ustreznosti

5	Ves
4	Večina
3	Veliko
2	Malo
1	Nič

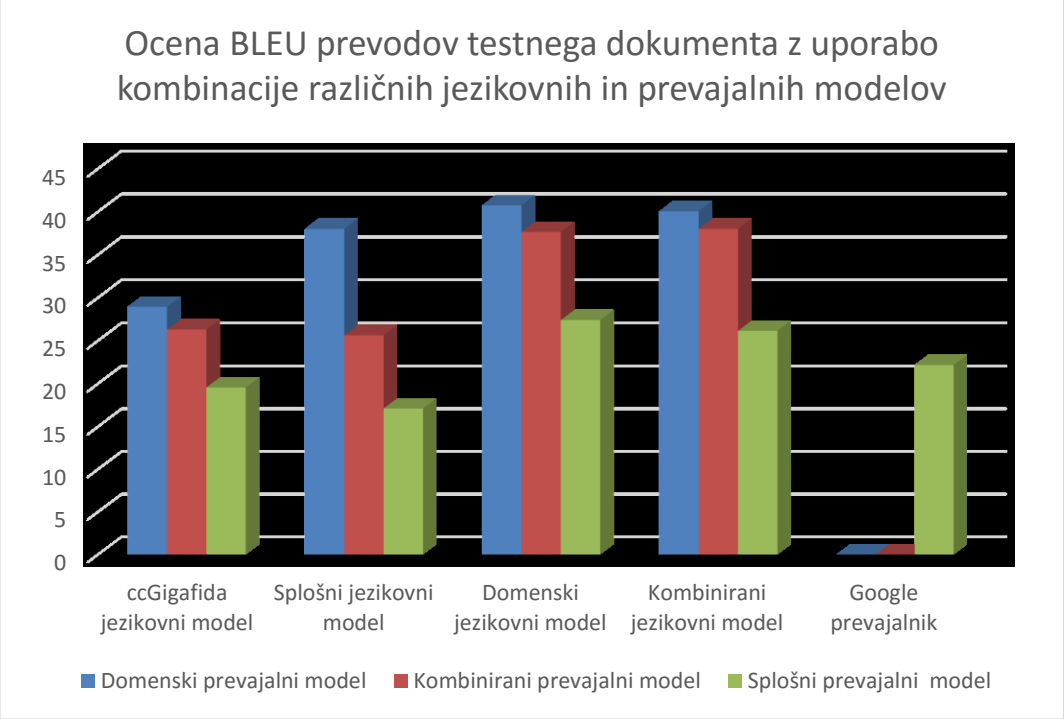
7.2. Rezultati prevajanja

Prevajanje testnega dokumenta izvedemo z več kombinacijami jezikovnega in prevajalnega modela. Zaradi lažjega pregleda dobljenih ocen BLEU zapišemo povzetek rezultatov samodejnega ocenjevanja v obliki preglednice, ki je prikazana v tabeli 7.3.

Tabela 7.3: Ocene BLEU po prevajanju testnega dokumenta z več kombinacijami jezikovnega in prevajalnega modela

	Splošni PM	**Domenski PM**	**Kombinirani PM**
ccGigafida jezikovni model	19,56	28,94	26,29
Splošni jezikovni model	17,11	38	25,62
Domenski jezikovni model	27,40	40,74	37,65
Kombinirani jezikovni model	26,15	40,05	38
Google Prevajalnik	22,17	X	X

Na sliki 7.1 so grafično prikazani rezultati samodejnih ocen BLEU. Ocene so združene v skupine po jezikovnih modelih. Za posamezen jezikovni model so prikazane vrednosti za različne prevajalne modele. Tako lahko iz tabele razberemo, da so najvišje ocene BLEU dosegli modeli, ki vključujejo domenski prevajalni sistem. Ti so na grafikonu predstavljeni z modrim stolpičem.



Slika 7.1: Grafikon ocen BLEU, razdeljen po jezikovnih modelih glede na različne prevajalne modele

V nadaljevanju natančneje opišemo posamezne kombinacije prilagajanja.

7.2.1. Rezultati prevajanja s prilagojenim prevajalnim modelom

V tem podpoglavju predstavimo rezultate prilagajanja prevajalnega modela z uporabo različnih zbirk in za ročno ocenjene prevode tudi izseke iz testnih prevodov. Pri tem eksperimentu za jezikovni model uporabimo slovenske dele posameznih jezikovnih zbirk.

Prilagajanje začnemo s splošno zbirko, ki je sicer zelo velika, ni pa neposredno s področja farmacije, nato uporabimo samo domensko zbirko in na koncu kombinacijo obeh.

7.2.1.1. Splošna učna zbirka

S splošno zbirko želimo izvedeti, kako dober prevod dobimo, če strojni prevajalnik nima znanja s področja farmacije, razen toliko, kot to omenjajo s farmacijo povezani zakoni, sprejeti v Evropskem parlamentu.

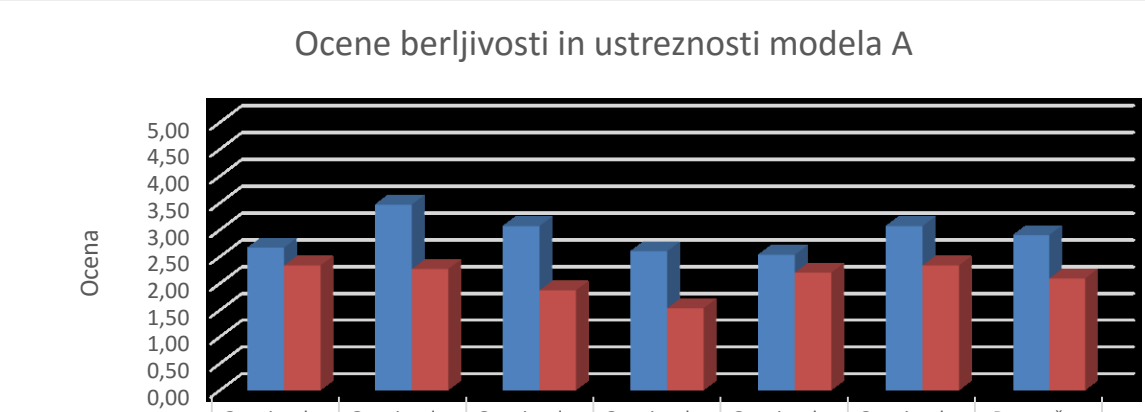
Na sliki 7.2 je prikazan izsek iz testnega dokumenta, ki je bil ocenjen z oceno BLEU 17,11.

```
Kaj koristi Deltyba, so pokazali v študijah?

Posledice Deltyba so na v eden od glavnih študija, ki vključuje 481 odrasli
s tuberkulozo odporen na standardno zdravljenje.

Patients in študije, ki so dobile Deltyba ali slepih placebom (a) za 2
mesecih poleg svojih drugih obravnav.

Glavna ukrep učinkovitosti je bil delež bolnikov, ki ne bo več imela
bakterije v njihovih sputum (phlegm).

Po 2 mesecih zdravljenja ostalo več kot 40 % bolnikov, ki so bili ob Deltyba
ni več bakterij v njihovo sputum v primerjavi s 30 % bolnikov, ki so bili
ob žegen.

Kaj je Mekinist in kaj je mogoče uporabiti za?

Mekinist je rak zdravilo, ki se uporabljajo za zdravljenje odrasle z melanom
(vrsta kožnega raka), ki se je razširil ali ne more biti surgically
odstrani.

Treba je uporabljen samostojno ali v kombinaciji z drugo zdravilo za raka,
dabrafenib.

Mekinist je le za bolnike, katerih melanom celice so bile preskušene in
pokazala, da imajo specifično genske mutacije (sprememba) v svojem genov,
ki se imenuje &quot; BRAF V600 &quot;.

Mekinist vsebuje aktivna snov trametinib.

Kako se Mekinist uporablja?

Zdravljenje z Mekinist in nadzorovano je treba začeti izvajati, ki ga
opravi zdravnik doživeli pri uporabi raka zdravil.

Kako zdravilo mogoče pridobiti samo z recepta.

Mekinist je na voljo kot tablet (0,5 mg, 1 mg in 2 mg).

Kako odmerek Mekinist bodisi, uporabljen sam ali v kombinaciji z dabrafenib
je 2 mg enkrat na dan, na podobnem za vprašanja vsak dan.
```

Slika 7.2: Izsek iz testnega dokumenta, preveden s splošnim jezikovnim modelom in splošnim prevajalnim modelom (BLEU 17,11)

Poleg tega, da ta prevod doseže najnižjo oceno BLEU v primerjavi z drugimi testnimi prevodi, ocenjevalci ocenijo kot najslabšo tudi berljivost tega testnega dokumenta. Na lestvici od 1 do 5 jo ocenijo s povprečno vrednostjo 2,09, kar ustreza opisni oceni »slabo berljiv«. Berljivost je

v veliki meri pogojena z uporabo ustreznih besednih zvez, količina teh pa je pri tem modelu omejena, ker nima veliko informacij s področja ciljne domene. Zato je ustreznost tega testnega prevoda skoraj za eno oceno nižja od vseh ostalih testnih prevodov, ki imajo vsaj nekaj znanja s področja ciljne domene. Pomensko ustreznost ocenjevalci ocenijo s povprečno oceno 2,90. Najslabša ocena, ki jo dobi, je 2,53, najboljša pa 3,47. Na sliki 7.3 je za lažjo predstavo prikazan grafikon vseh ocen, ki jih ocenjevalci dodelijo za berljivost in ustreznost. Standardni odklon od povprečja za berljivost je 0,32, za ustreznost pa 0,19.



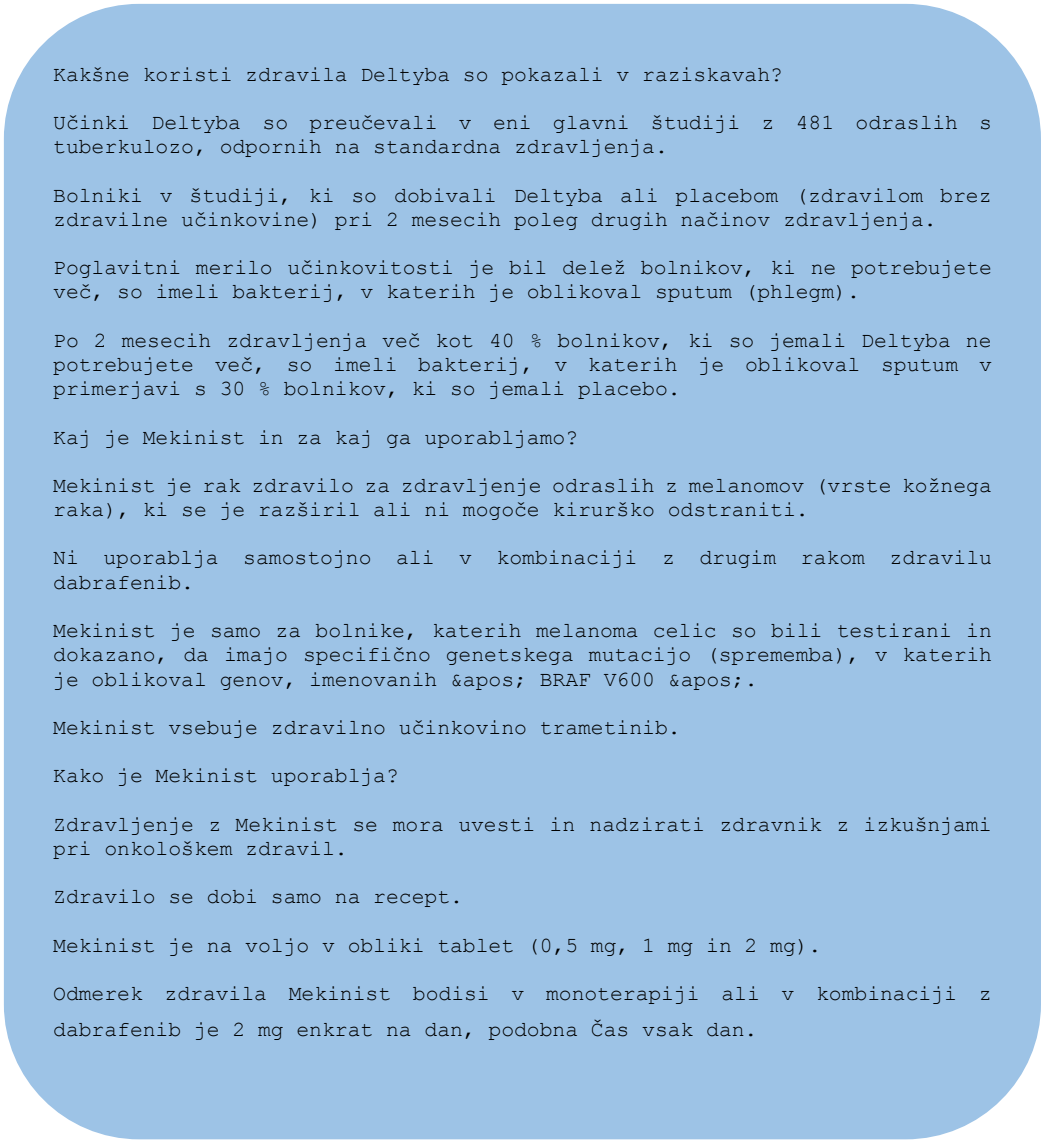
	Ocenjevalec 1	Ocenjevalec 2	Ocenjevalec 3	Ocenjevalec 4	Ocenjevalec 5	Ocenjevalec 6	Povprečna ocena
Ustreznost	2,67	3,47	3,07	2,60	2,53	3,07	2,90
Berljivost	2,33	2,27	1,87	1,53	2,20	2,33	2,09

Slika 7.3: Grafikon ocen berljivosti in ustreznosti za dokument, preveden s splošnim prevajalnim modelom

7.2.1.2. Domenska učna zbirka

V domensko učno zbirko vključimo vse razpoložljive vire s področja farmacije, ki nam jih uspe pridobiti, zato pričakujemo, da mora prevod, preveden z domenskim modelom, dobiti boljšo oceno od splošnega.

Na sliki 7.3 je prikazan izsek testnega prevoda, ki je preveden z domenskim prevajalnim modelom. Tudi jezikovni model je v tem primeru izdelan iz slovenskega dela učne zbirke.



Slika 7.4: Prevod, izdelan z domenskim jezikovnim modelom in domenskim prevajalnim modelom (BLEU 40,74)

Testni prevod doseže oceno BLEU 40,74, berljivost tega testnega prevoda pa ocenjevalci ocenijo s povprečno oceno 3,17, pri čemer je standardni odklon 0,27. V primerjavi z našimi ostalimi modeli doseže ta domenski prevajalni model najvišjo oceno berljivosti kot tudi pomenske ustreznosti, ki je 4,00. Glede na mersko lestvico to pomeni, da so ocenjevalci v povprečju mnenja, da ta model prenese večino pomena glede na izvirnik. Standardni odklon od povprečne ocene ustreznosti je 0,22.

Ta model glede na ostale naše modele doseže najvišje vrednosti berljivosti in ustreznosti, ki pa niso višje od ocen testnega prevoda, prevedenega z Google Prevajalnikom. Slednji sicer doseže oceno BLEU samo 22,17, ki je za skoraj polovico nižja od našega domenskega modela.

Sumimo, da je vzrok za tak rezultat v tem, da je uglaševanje s postopkom MERT naravnano na čim boljši rezultat BLEU. V tem primeru prevajalnik lahko doseže tako visoko vrednost BLEU, ker je učna zbirka terminološko najbližje testnim dokumentom.

7.2.1.3. Kombiniran prevajalni model

V eksperimentu, v katerem za učenje prevajalnega modela uporabimo združeno domensko in splošno zbirko, predvidevamo, da se s povečanjem zbirke izboljša tudi ocena testnega prevoda. Pri tem eksperimentu za jezikovni model uporabimo slovenski del splošne zbirke, domenske zbirke in kombinirane učne zbirke. Po končanem prevajanju testni dokument dobi najboljšo oceno BLEU v kombinaciji s kombiniranim jezikovnim modelom, ki znaša 38. Ročne ocene za ta model ne izvedemo, ker je vseh kombinacij jezikovnega in prevajalnega modela preveč, da bi ocenjevalci ocenili vse.

7.2.2. Rezultati prevajanja s prilagojenim jezikovnim modelom

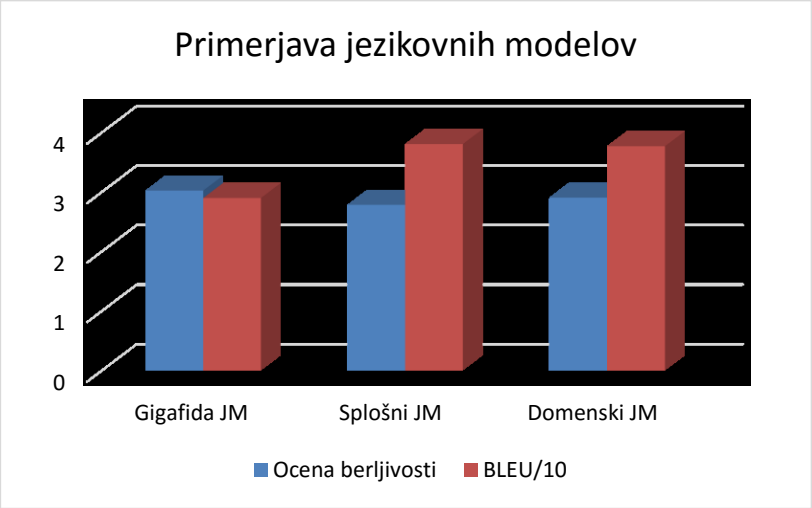
V teoretičnem delu tega magistrskega dela predvidevamo, da jezikovni model v veliki meri vpliva na berljivost prevoda. V našem eksperimentu želimo na praktičnih primerih preveriti vpliv različnih jezikovnih modelov tako na berljivost prevoda kot tudi na pomensko ustreznost. Ročno ocenjevanje izvedemo za tri testne dokumente, ki so prevedeni z uporabo treh različnih jezikovnih modelov. Za pripadajoči prevajalni model v dveh primerih izberemo isti domenski model, v enem primeru pa kombinirani model.

Prilagajanje začnemo s slovenskim korpusom ccGigafida, nato z istim prevajalnim modelom uporabimo še splošni jezikovni model, za tretji ocenjevalni dokument pa izberemo kombinacijo domenskega jezikovnega modela in kombiniranega prevajalnega modela. V nadaljevanju predstavimo svoje izsledke in zglede testnih dokumentov.

7.2.2.1. ccGigafida jezikovni model

Za jezikovni model, zgrajen iz slovenskega korpusa ccGigafida, pričakujemo, da bi lahko zaradi svoje velikosti in pestrosti vsebovanih besedil znatno prispeval k berljivosti prevodov. Testni prevod, preveden z uporabo jezikovnega modela ccGigafida in domenskega prevajalnega modela, doseže oceno BLEU 28,94. S povprečno oceno berljivosti 3,02 in standardnim odklonom 0,38 je ta model glede na berljivost boljši od drugih dveh testnih prevodov, pri katerih tudi opazujemo vpliv jezikovnega modela na prevod. Druga dva imata sicer višjo oceno

BLEU. Na sliki 7.5. je z grafikonom ponazorjena primerjava vseh treh jezikovnih modelov glede na berljivost.



Slika 7.5: Grafikon primerjave treh jezikovnih modelov glede na oceno berljivosti

Ocenjevalci temu testnemu prevodu dajo povprečno oceno 3,79 za pomensko ustreznost. Ker je v modelu uporabljen domenski prevajalni model, je v testni prevod, glede na ocene, prenesena večina pomena izvirnika.

Skupna slabost vseh testnih prevodov, prevedenih s statističnim strojnim prevajalnikom, je slovnična pravilnost, ki jo je brez uvedbe dodatnih pravil težko doseči.

Na sliki 7.6 prikažemo vsebino testnega dokumenta, prevedenega z jezikovnim modelom ccGigafida in domenskim prevajalnim.

Kakšne koristi Deltyba niso pokazale v raziskavah?

Učinki Deltyba so preverili v študiji 481 odraslih s tuberkulozo, odpornih na standardna zdravljenja.

Bolniki v študiji, so prejeli Deltyba ali placebo (navidezno) za 2 meseca poleg drugih načinov zdravljenja.

Poglavitni merilo učinkovitosti je bil delež bolnikov, ki ne potrebujete imeli bakterije v njihovo izločanje sluzi (phlegm).

Po 2 mesecih zdravljenja več kot 40 % bolnikov, ki so jemali Deltyba ne potrebujete imeli bakterije v njihovo izločanje sluzi v primerjavi s 30 % bolnikov, ki so jemali placebo.

Kaj je Mekinist in kaj ga uporabljamo?

Mekinist je rak zdravilo za zdravljenje odraslih z melanomom (vrste kožnega raka), ki se je razširil ali jih ni mogoče kirurško odstraniti.

Kot samostojno ali v kombinaciji z drugim rakom zdravilo, dabrafenib.

Mekinist je le za bolnike z melanomom celice so preizkusili in ugotovili, da imajo specifično genetske mutacije (sprememba) v genov, ki se imenuje ' BRAF V600 '.

Mekinist vsebuje učinkovino trametinib.

Kako je Mekinist uporablja?

Zdravljenje z Mekinist se mora uvesti in nadzirati zdravnik z izkušnjami pri onkološkem zdravil.

Pršilo se dobi samo na recept.

Mekinist je na voljo v obliki tablet (0,5 mg, 1 mg in 2 mg).

Odmerek Mekinist bodisi samo ali v kombinaciji z dabrafenib je 2 mg enkrat na dan, ob podobno Čas vsak dan.

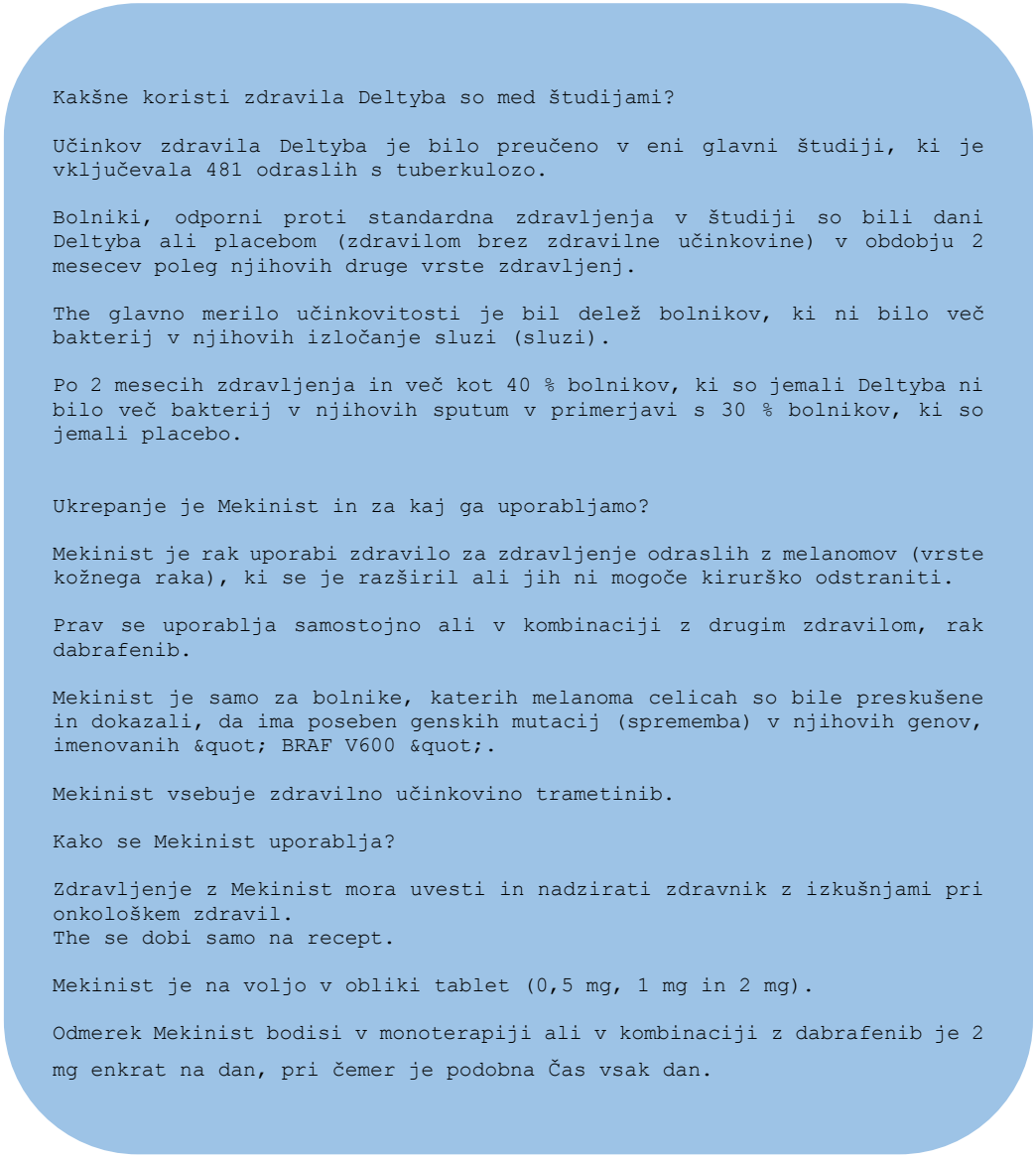
Slika 7.6: Prevod, izdelan z jezikovnim modelom Gigafida in domenskim prevajalnim modelom (BLEU 28,94)

7.2.2.2. Splošni jezikovni model

V našem naslednjem eksperimentu ohranimo domenski prevajalni model, za jezikovni model pa uporabimo jezikovni model, zgrajen iz splošnih zbirk Euparl in DGT. Ta testni prevod doseže oceno BLEU 38. Iz ocene BLEU lahko razberemo, da s tem modelom prenese kar velik del terminologije glede na izvirnik, kar dokazuje tudi povprečna ocena pomenske ustreznosti 3,73 (standardni odklon 0,26), ki je višja od splošnega in kombiniranega prevajalnega modela.

Predvidevamo, da zaradi slabše domenske pokritosti jezikovnega modela ta testni dokument dobi najslabšo oceno berljivosti v primerjavi z ostalimi dokumenti, pri katerih opazujemo vpliv jezikovnega modela. Povprečna ocena berljivosti tega testnega dokumenta je 2,78 in je za 0,12

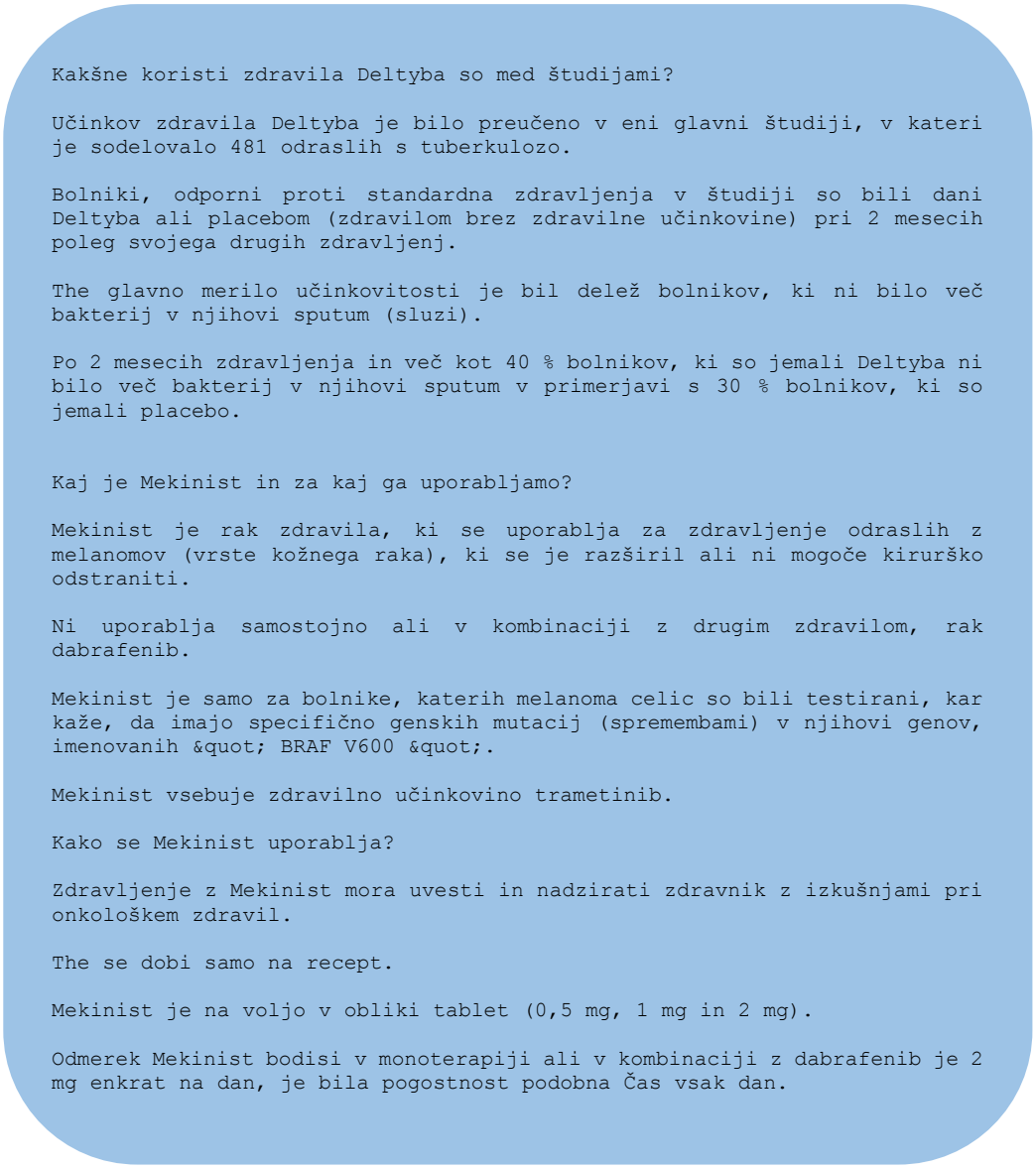
nižja od enakega modela, v katerem je namesto splošnega jezikovnega modela uporabljen jezikovni model, zgrajen iz zbirke ccGigafida.



Slika 7.7: Prevod, izdelan s splošnim jezikovnim modelom in domenskim prevajalnim modelom (BLEU 38,00)

7.2.2.3. Domenski jezikovni model

Z uporabo domenskega jezikovnega modela in kombiniranega prevajalnega modela lahko primerjamo vpliv jezikovnega modela na berljivost tudi z vidika, če je namesto domenskega prevajalnega modela uporabljen nekoliko splošnejši prevajalni model.



Slika 7.8: Prevod, izdelan z domenskim jezikovnim modelom in kombiniranim prevajalnim modelom (BLEU 37,65)

Testni dokument, preveden s to kombinacijo jezikovnega in prevajalnega modela, je ocenjen s samodejno oceno BLEU 37,65. Ta je nižja od ocene BLEU 38, ki jo doseže model s splošnim jezikovnim modelom in domenskim prevajalnim modelom, vendar pa ocenjevalci berljivost tega dokumenta ocenijo z oceno 2,90 (standardni odklon 0,32), kar je boljše od splošno-domenskega modela, ki doseže oceno 2,78. Slika 7.8 prikazuje izsek testnega prevoda z domenskim prevajalnim modelom, ki ga lahko primerjamo s sliko 7.7, ki prikazuje izsek testnega prevoda, kjer je uporabljen splošni jezikovni model. Ker je bil v tem primeru uporabljen kombiniran prevajalni model, je verjetno pričakovana ocena ustreznosti 3,62 s

standardnim odklonom 0,18, ki je nižja od drugih dveh modelov prilagajanja jezikovnega modela, pri katerih pa je uporabljen domenski prevajalni model.

7.2.2.4. Rezultati prevajanja z Google Prevajalnikom

Na sliki 7.9 je prikazan izsek testnega dokumenta, ki je preveden z Google Prevajalnikom. Algoritem BLEU mu da oceno 22,17, ki je nižja od vseh naših testnih prevodov, ki so prevedeni z enim od domensko prilagojenih modelov.

Na naše presenečenje pa dobi ta testni dokument pri ročnem ocenjevanju oceno 4,41 za pomensko ustreznost ter 3,48 za berljivost, kar je boljše od vseh naših prevajalnih modelov.

Predvidevamo, da je vzrok za veliko uspešnost dejstvo, da so v Googlov prevajalnik vključene vse javno dostopne zbirke in je zbirka EMA le en del med množico različnih, ki prevajalniku doprinese veliko znanja o farmacevtski domeni. Prav tako ugotovimo, da Google Prevajalnik v zadnjem obdobju uporablja pristop z nevronskim strojnim prevajanjem, ki je nadgradnja prevajanja na osnovi besednih zvez.

Kakšne koristi od Deltyba je bilo dokazano v študijah?

Učinki Deltyba so pogledal v eni glavni študiji, ki vključuje 481 odraslih z tuberkuloze, odporne na standardno zdravljenje.

Bolniki v študiji so prejemale Deltyba ali placeba (zdravljenje) za 2 meseca poleg svojih drugih oblikah zdravljenja.

Glavno merilo učinkovitosti je bil delež bolnikov, ki ni imela več bakterij v njihovem izpljunku (sluz).

Po 2 mesecih zdravljenja je imela bakterije v svojem izpljunku v primerjavi s 30% bolnikov, ki so jemali placebo, več kot 40% bolnikov, ki so jemali Deltyba ni več.

Kaj je Mekinist in za kaj ga uporabljamo?

Mekinist je zdravilo rak se uporablja za zdravljenje odraslih bolnikov z melanomom (vrsto kožnega raka), ki se je razširil ali ga ni mogoče kirurško odstraniti.

To se uporablja samostojno ali skupaj z drugimi zdravili raka, dabrafenib.

Mekinist je samo za bolnike, pri katerih melanoma celice so bile testirane in dokazano, da imajo posebne genetske mutacije (spremembe) v svojih genih imenovanih ', ki BRAF V600-l';.

Mekinist vsebuje zdravilno učinkovino trametinib snovi.

Kako se Mekinist uporablja?

Zdravljenje z Mekinist mora uvesti in nadzorovati zdravnik, ki ima izkušnje z uporabo zdravil raka.

Zdravilo se dobi samo na recept.

Mekinist je na voljo v obliki tablet (0,5 mg, 1 mg in 2 mg).

Doza Mekinist bodisi uporabljen sam ali pa v kombinaciji z dabrafenib je 2 mg enkrat na dan, po podobnem času vsak dan.

Slika 7.9: Prevod, izdelan z Google Prevajalnikom (BLEU 22,17)

7.2.3. Rezultati prevajanja z dodanim slovarjem

Uporaba slovarja je ključnega pomena, če nimamo dovolj velike domenske učne zbirke in želimo kakovost prevoda izboljšati tako, da strojnemu prevajalniku dodamo domenski slovar, iz katerega črpa izraze, ki jih sicer ne najde v prevajalni tabeli. Pri našem poskusu za učenje uporabimo zmanjšano domensko zbirko, ki vsebuje 260.000 vrstic. Brez dodane vsebine slovarja za testni prevod dosežemo rezultat BLEU 26,40, medtem ko z dodanim slovarjem dosežemo rezultat 27,51.

Dobljeni rezultat kaže, da uporaba slovarja pripomore k izboljšanju ocene BLEU za testni prevod. Menimo, da je uporaba slovarja še posebej koristna, kadar želimo za projekt z več dokumenti zagotoviti uporabo skladne terminologije.

7.2.4. Rezultati prevajanja s transduktivno učno zbirko

Za ta eksperiment uporabimo samo petino dokumentov zbirke PubMed, ki se začnejo na črko A. Združeni dokument je dolg približno 500 MB in vsebuje kot 8,3 milijona vrstic povzetkov. Po končani pripravi zbirke, ki traja slabe tri ure, za prevajanje uporabimo domenski jezikovni model in domenski prevajalni model. Prevajanje traja 75 ur.

Na sliki 7.10 sta prikazana izsek iz prevedene zbirke in pripadajoči izvirnik v angleškem jeziku. Ker je raven kakovosti prevoda na zelo nizki ravni z veliko manjkajočimi prevodi, se odločimo, da z učenjem prevajalnega modela po metodi transduktivnega prevajanja ne nadaljujemo.

Angleški izvirnik:
Experimental
Sodium hydride (4.8 g, 120 mmol) was suspended in dimethylformamide (DMF, 30 ml). Triazole (8.28 g, 120 mmol) dissolved in DMF (30 ml) was slowly added dropwise at 273 K, and reacted at room temperature for 30 min.
2-Chloro-1-(4-fluorophenyl)ethanone (15.48 g, 90 mmol) dissolved in DMF (30 ml) was then slowly added dropwise, and reacted at room temperature for 4 h.
The mixture was placed in ice-water (300 ml), and 1 mol hydrochloric acid (50 ml) was then added. After filtration, the filtrate was neutralized with sodium bicarbonate to pH = 6, and a yellow deposit was obtained. Recrystallization with ethanol yielded a white deposit (m.p. 397–400 K). Crystals suitable for a X-ray analysis study were obtained by dissolving the crude product (1.0 g) in 95% ethanol (30 ml) and then allowing the solution to evaporate slowly at room temperature for about 7 days.

Slovenski prevod:
Poskusnih
Natrijev hydride (4, 8 g, 120 mmol) so suspendirane v dimethylformamide (DMF, 30 ml). Triazolski (8.28 g, 120 mmol) raztopljenega v DMF (30 ml) je bila počasi dodano dropwise ob 273 K ter reagiralo na sobni temperaturi 30 min.
2-Chloro-1- (4-fluorophenyl) ethanone (15.48 g, 90 mmol) raztopljenega v DMF (30 ml) je bila potem počasi dodano dropwise, in reagiralo na sobni temperaturi 4 ure.
Mešanico je bila dati v ice-water (300 ml), in 1 mol klorovodikova kislina (50 ml) je bilo nato dodamo. Po filtracijo, filtrate je bila neutralized z natrijevim bikarbonat v pH = 6, in rumene usedlino so bili pridobljeni. Recrystallization z etanol dale pojavi bela usedlina (m.p. 397 – 400 K). Kristalov primerno za z rentgenskim slikanjem analiza študiji so bili pridobljeni z raztapljanjem neprilagojenimi stopnjami zdravila (1, 0 g) pri 95 % etanola (30 ml) in nato si raztopino do evaporate počasi – s sobni temperaturi približno 7 dni.

Slika 7.10: Prevod, izdelan z domenskim prevajalnikom z namenom transduktivnega prevajanja

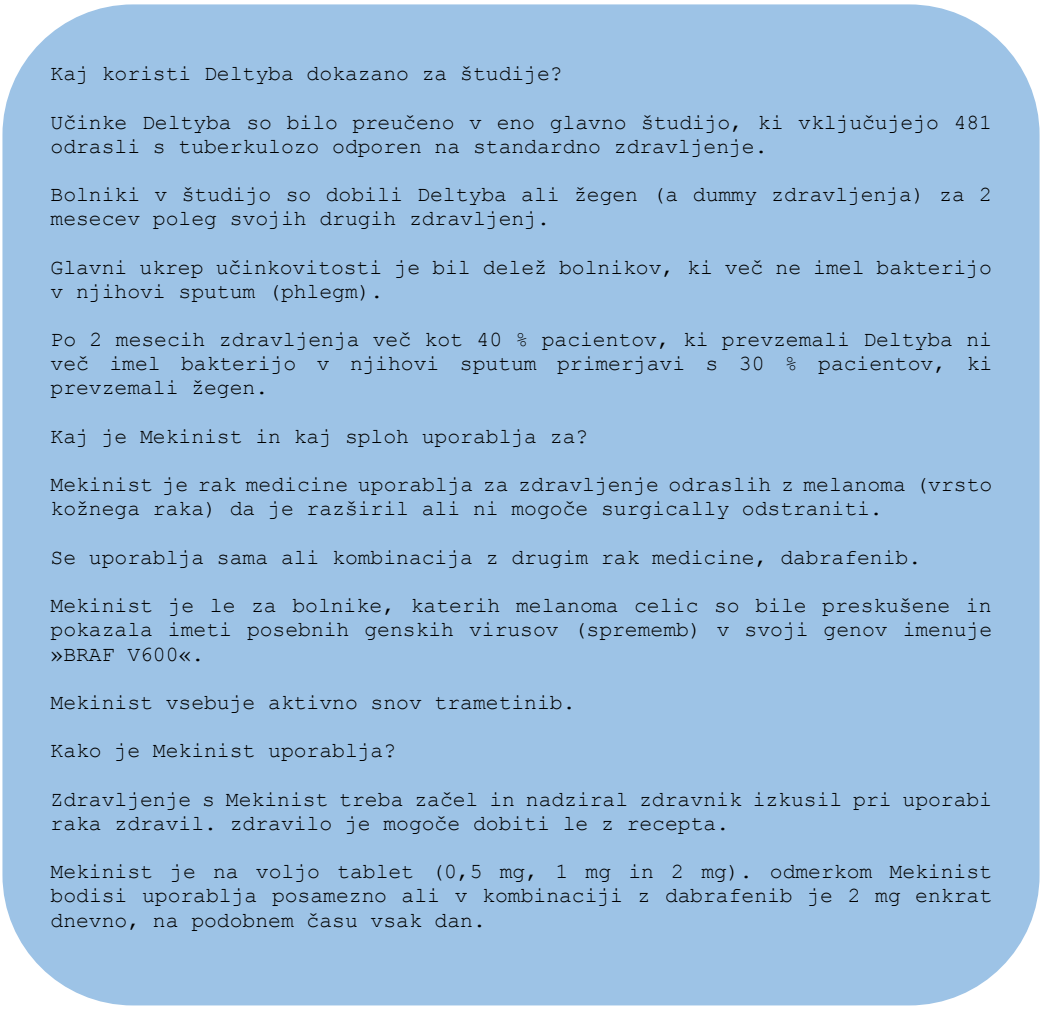
Zbirka PubMed vsebuje veliko besedil, tesneje povezanih z medicinskim področjem, ki pa v našem modelu ni dovolj dobro zastopano. Opazimo lahko, da so primerno prevedena imena kemijskih snovi in postopkov, slaba pa je berljivost in splošno izrazoslovje. Predvidevamo, da lahko dosežemo boljše rezultate, če bi imeli na voljo medicinski slovar, ki bi ga uporabili z našim kombiniranim modelom. Tega nam medicinska fakulteta za potrebe raziskave ne omogoči.

7.2.5. Rezultati prevajanja s faktorskim modelom

V tem podpoglavju predstavimo rezultate prevajanja testnega dokumenta s faktorskim modelom prevajanja. Najprej prevedemo testni dokument s splošno zbirko, nato z domensko in na koncu s kombinirano zbirko. V nadaljevanju je za vsak model prikazan vzorec testnega prevoda, ocena BLEU in kratek opis.

7.2.5.1. Splošna učna zbirka

Za učenje prvega faktorskega modela uporabimo splošno učno zbirko, s čimer želimo predvsem primerjati kakovost testnega prevoda s testnim prevodom, prevedenim z modelom na osnovi besednih zvez. Na sliki 7.11 je prikazan izsek iz testnega prevoda, ki je dosegel oceno 13,99. Če ga primerjamo s splošnim modelom na osnovi besednih zvez, lahko rečemo, da sta berljivost in ustreznost podobni, le da je v primeru faktorskega modela uporabljenih manj enakih izrazov in je ocena BLEU zato toliko nižja. V obeh primerih lahko vidimo, da moramo za dosego boljše kakovosti prevoda uporabiti učno zbirko, ki je bližja ciljnemu področju.



Slika 7.11: Prevod, izdelan s faktorskim splošnim modelom (BLEU 13,99)

7.2.5.2. Domenska učna zbirka

Za gradnjo faktorskega modela z uporabo domenske učne zbirke uporabimo enako zbirko kot za model na osnovi besednih zvez, le da so tokrat dodane še oblikoskladenjske oznake in osnovne oblike. Zanima nas vpliv dodatnih oznak na testni prevod in njegovo oceno. Testni prevod doseže oceno BLEU 39,99, ki je od domenskega modela na osnovi besednih zvez nižja, vendar zgolj za 0,75 ocene.

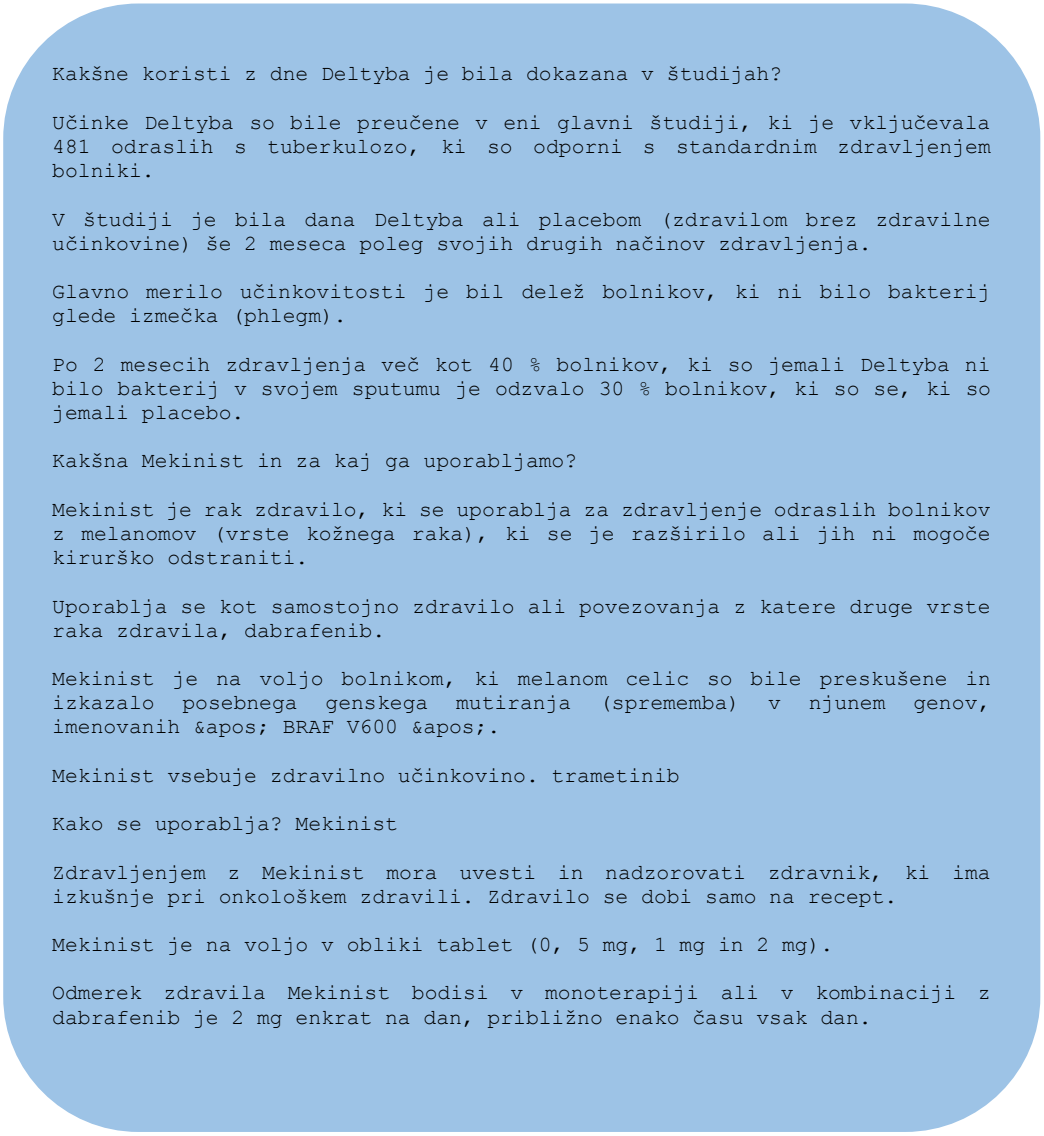
Na sliki 7.12 je prikazan izsek testnega prevoda, prevedenega z domenskim faktorskim modelom. V primerjavi z domenskim modelom na osnovi besednih zvez ne opazimo bistvenih razlik v kakovosti, le da so redke besedne vrste v pravilnejši obliki, kot denimo v naslednjih primerih:

- *Zdravljenje z Mekinist se mora uvesti in nadzirati zdravnik z izkušnjami pri onkološkem zdravil. (Model na osnovi besednih zvez)*
- *Zdravljenje z Mekinist mora uvesti in nadzirati zdravnik z izkušnjami pri onkološkem zdravil. (Faktorski model)*

Slika 7.12: Prevod, izdelan s faktorskim domenskim modelom (BLEU 39,99)

7.2.5.3. Kombinirana učna zbirka

V kombinirani zbirki sta vključeni obe, splošna in domenska zbirka. Zaradi velikosti združene učne zbirke gradnja modela traja dlje časa. Z zgrajenim faktorskim modelom prevedemo testni dokument, ki doseže oceno BLEU 39,29. Ocena je za 1,29 višja od ocene BLEU, ki jo je dobil testni dokument, preveden s kombiniranim modelom na osnovi besednih zvez.



Slika 7.13: Prevod, izdelan s faktorskim kombniniranim modelom (BLEU 39,29)

Ocena faktorskega kombiniranega modela je sicer visoka, vendar če si ogledamo sliko 7.13, vidimo, da je tako visoka ocena posledica optimizacije MERT. Prevod vsebuje veliko pravilnih izrazov, vendar berljivost ni primerljiva z domenskim faktorskim modelom. V prevodu opazimo tudi več pojavitev splošnejših izrazov.

7.2.5.4. Domenska zbirka in jezikovni model ccGigafida

Ta faktorski model je zgrajen iz domenske učne zbirke ter dveh jezikovnih modelov. Oba jezikovna modela sta zgrajena iz označene zbirke ccGigafida. Prvi vsebuje verjetnostno porazdelitev površinskih oblik, medtem ko drugi vsebuje verjetnostno porazdelitev oblikoskladenjskih oznak.

Testni prevod, preveden s tem modelom, doseže oceno BLEU 28,85. Ocena BLEU je za 11,14 nižja od ocene, ki jo doseže domenski faktorski model. Tak rezultat je pričakovan, vzrok za to pa vidimo v pomanjkanju domenskih izrazov v korpusu ccGigafida. Z uporabo zbirke ccGigafida ne pričakujemo izboljšanja ocene BLEU ali ustreznosti prevoda, želimo pa preveriti vpliv korpusa slovenskega jezika na berljivost testnega prevoda v primerjavi s splošnim modelom.

```
Kakšne koristi Deltyba niso pokazale v raziskavah?

Učinki Deltyba so preverili pri eni glavni študiji, ki 481 odraslih s
tuberkulozo, odpornih na standardna zdravljenja.

Bolnikov v študiji so dobili Deltyba ali placeba za 2 meseca, poleg katerih
drugih načinov zdravljenja.

Glavno merilo učinkovitosti je bil delež bolnikov, ki niso več imeli
bakterije v svojem sputumu (sluzi).

Po 2 mesecih zdravljenja, več kot 40 % bolnikov, ki so jemali Deltyba niso
več imeli bakterije v njihovem sputumu v primerjavi s 30 % bolnikov, ki so
jemali placebo.

Kaj je Mekinist in za kaj ga uporabljamo?

Mekinist je rak zdravilo za zdravljenje odraslih z melanomom (vrsta kožnega
raka), ki se je razširil ali ni mogoče kirurško odstraniti.

Ga uporabljamo samostojno ali v kombinaciji z drugim rakom zdravila,
dabrafenib.

Mekinist je samo za bolnike z melanomom celic so preizkusili in ugotovili,
da imajo specifično genetsko mutacijo (sprememba) glede genov, imenovanih
&apos; BRAF V600 &apos;.

Mekinist vsebuje učinkovino trametinib.

Kako je Mekinist uporablja?

Zdravljenje z Mekinist mora začeti in nadzirati zdravnik z izkušnjami pri
onkološkem zdravil.

Zdravilo se dobi samo na recept.

Mekinist je na voljo v obliki tablet (0, 5 mg, 1 mg in 2 mg).

Odmerek Mekinist bodisi uporablja samostojno ali v kombinaciji z dabrafenib
je 2 mg enkrat na dan, ob podobnem času vsak dan.
```

Slika 7.14: Prevod, izdelan z domenskim modelom in ccGigafido (BLEU 28,85)

Če primerjamo izsek tesnega prevoda na sliki 7.14 z izsekom splošnega faktorskega modela na sliki 7.11, lahko vidimo, da je berljivost več stavkov iz modela ccGigafida boljša od splošnega modela, oba modela pa zahtevata še precej popravljanja, da bi dosegla primerno raven

ustreznosti. Zaključujemo, da je korpus slovenskega jezika ccGigafida smiselno uporabiti kot dopolnilno zbirko pri gradnji jezikovnih modelov.

7.2.6. Rezultati ročnega ocenjevanja testnih prevodov

Z ročnim ocenjevanjem testnih dokumentov na podlagi ocen posameznih vprašanj predstavimo tudi povzetek ocen berljivosti in ustreznosti po modelih, skupne ocene po modelih, na koncu pa se dotaknemo še ocen po posameznih vprašanjih in strinjanja med prevajalci. Pogledamo tudi, ali dolžina prevoda vpliva na njegovo oceno.

7.2.6.1. Ocene berljivosti in ustreznosti po modelih

V prejšnjih podpoglavjih predstavimo ocene posameznih prevajalnih modelov in njihovo vsebino. V tem podpoglavju pa podamo povzetek vseh ocen berljivosti in ustreznosti po modelih.

Tabela 7.4: Povzetek vseh ocen berljivosti po posameznih modelih

Ocena berljivosti prevodov										
Mod	**Opis modela**	**Ocena BLEU**	**Ocen. 1**	**Ocen. 2**	**Ocen. 3**	**Ocen. 4**	**Ocen. 5**	**Ocen. 6**	**Povpreč. ocena**	**Std. odklon**
C	Google Prevajalnik	22,17	3,93	3,60	3,53	3,00	3,53	3,27	**3,48**	**0,27**
D	Domenski JM in domenski PM	40,74	3,47	3,27	3,20	2,60	3,33	3,13	**3,17**	**0,25**
B	Gigafida JM in domenski PM	28,94	3,27	3,07	3,07	2,20	3,33	3,20	**3,02**	**0,35**
A	Domenski JM in kombiniran PM	37,65	3,40	3,13	2,87	2,40	2,93	2,67	**2,90**	**0,30**
E	Splošni JM in domenski PM	38	3,07	3,07	2,73	2,33	2,87	2,60	**2,78**	**0,24**
F	Splošni JM in splošni PM	17,11	2,33	2,27	1,87	1,53	2,20	2,33	**2,09**	**0,27**

V tabeli 7.4 so prikazane ocene, ki jih dajo ocenjevalci posameznim modelom za berljivost, v tabeli 7.5 pa za ustreznost s povprečno oceno v zadnjem stolpcu obeh tabel. Za lažje razumevanje tabel so posamezni modeli označeni z zaporednimi črkami abecede v prvem stolpcu in na kratko opisani v drugem stolpcu. V opis sta zajeta podatka o uporabljenem

jezikovnem in prevajalnem modelu. V tabeli je dodana tudi vrednost BLEU, ki jo doseže vsak od ročno ocenjenih prevodov.

Tabela 7.5: Povzetek vseh ocen ustreznosti po posameznih modelih

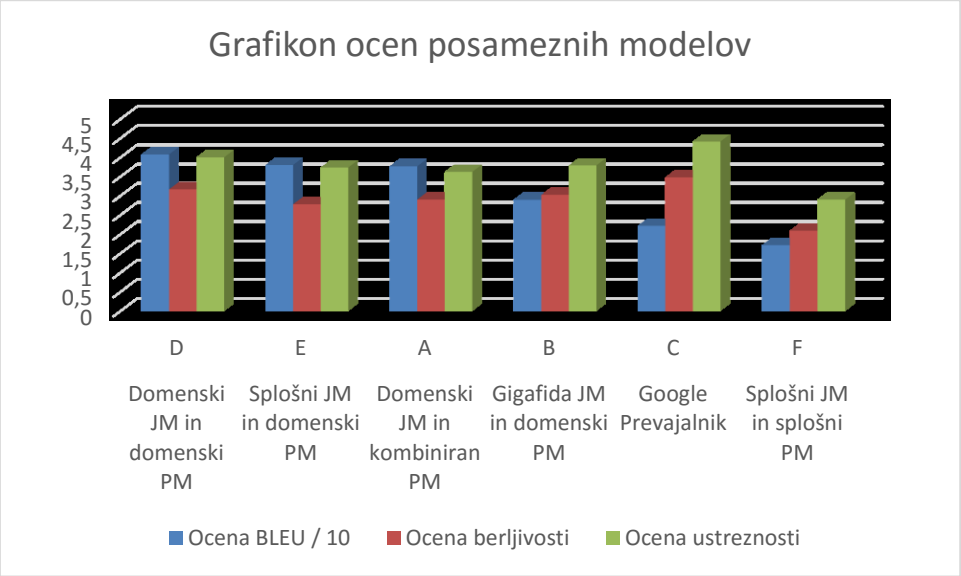
Ocena ustreznosti prevodov										
Mod	**Opis modela**	**Ocena BLEU**	**Ocen. 1**	**Ocen. 2**	**Ocen. 3**	**Ocen. 4**	**Ocen. 5**	**Ocen. 6**	**Povpreč. ocena**	**Std. odklon**
C	Google Prevajalnik	22,17	4,20	4,73	4,20	4,40	4,33	4,60	**4,41**	**0,18**
D	Domenski JM in domenski PM	40,74	3,67	4,13	4,13	3,73	4,07	4,27	**4,00**	**0,20**
B	Gigafida JM in domenski PM	28,94	3,67	4,20	3,53	3,60	3,67	4,07	**3,79**	**0,23**
E	Splošni JM in domenski PM	38	3,53	4,20	3,73	3,47	3,53	3,93	**3,73**	**0,24**
A	Domenski JM in kombiniran PM	37,65	3,60	3,93	3,73	3,40	3,40	3,67	**3,62**	**0,17**
F	Splošni JM in splošni PM	17,11	2,67	3,47	3,07	2,60	2,53	3,07	**2,90**	**0,31**

Tabeli sta razvrščeni po doseženi povprečni oceni posameznega modela. Če smo pozorni na zaporedje oznak modelov, lahko vidimo, da zaporedje oznak pri oceni berljivosti ni enako zaporedju oznak pri ustreznosti. Zaporedje prvih treh modelov je pri obeh kategorijah enako, medtem ko je model E, ki je sestavljen iz splošnega jezikovnega modela in domenskega prevajalnega modela, pri oceni ustreznosti uvrščen višje kot pri oceni berljivosti. Predvidevamo, da je višja ocena ustreznosti posledica domenske zbirke, ki je uporabljena za učenje prevajalnega modela.

Če se osredotočimo na opise posameznih modelov, lahko razberemo, da najboljše ocene pomenske ustreznosti dosežejo modeli, ki imajo domenski jezikovni model, sledijo pa mu modeli s kombiniranim in splošnim jezikovnim modelom. Pri ocenah berljivosti so v prednosti modeli, katerih jezikovni modeli so terminološko bližje slovenskemu jeziku.

V grafikonu na sliki 7.14 so posamezni modeli prikazani glede na ocene BLEU ter ocene berljivosti in ustreznosti, ki jih dobijo. Vidimo, da najvišje ocene BLEU dobijo modeli, katerih učne zbirke so najtesneje povezane z domeno besedila, ki je zajeta v testni prevod. Z grafa lahko

razberemo tudi, da ocenjevalci ne dajo najboljših ocen za berljivost in ustreznost modelom z najvišjo vrednostjo BLEU.



Slika 7.15: Grafikon ocen BLEU ter ocen za berljivost in ustreznost posameznih modelov

Za oceno BLEU in oceni berljivosti in ustreznosti za prve tri modele z najvišjimi vrednostmi BLEU lahko rečemo, da padajo približno simetrično, medtem ko so modeli, ki poleg prevajalnega modela uporabljajo splošni model ali ne uporabljajo domenskega modela, dobijo precej nizko oceno BLEU v primerjavi z oceno berljivosti in ustreznosti. Kot omenimo v prejšnjih poglavjih, vzrok za to vidimo v uglaševanju MERT za čim boljšo oceno BLEU. Ocenjujemo, da je pristop z uglaševanjem MERT primeren, če se želimo s prevajalnim sistemom čim bolj približati obliki in izrazoslovju v učni zbirki.

7.2.6.2. Medsebojno strinjanje ocenjevalcev

V tem podpoglavju predstavimo medsebojno strinjanje ocenjevalcev. Najprej pokažemo strinjanje glede ocen berljivosti in ustreznosti med vsemi ocenjevalci, v drugem delu pa še medsebojno strinjanje po parih ocenjevalcev.

Medsebojno strinjanje vseh ocenjevalcev

Ocenjevalci ocenjujejo berljivost in ustreznost šestih različnih dokumentov in dajo vsaki povedi svojo oceno, ne da bi vedeli, kakšne ocene dajo drugi ocenjevalci. Po ogledu tabele ocen na ravni posameznega stavka ugotovimo, da so ocenjevalci enotni glede ocene 5 pri 14 stavkih od skupaj 180 stavkov, ki jih ocenijo. Po enkrat se strinjajo še glede ocen 4, 3, 2 in 1. V tabeli 7.7

in 7.8 je prikazano strinjanje ocenjevalcev glede berljivosti in ustreznosti. Strinjanje izračunamo s pomočjo posplošene kappe oz. Fleisove kappe [51].

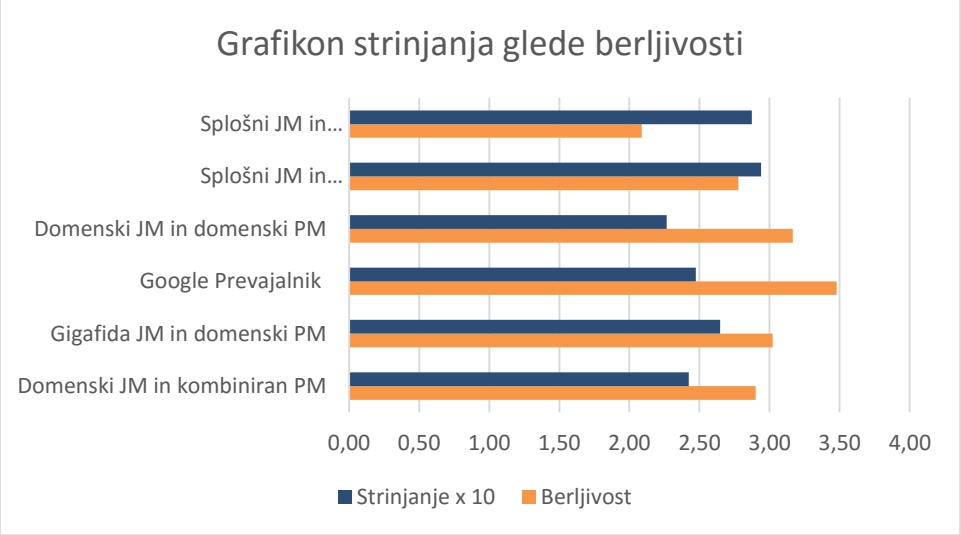
Tabela 7.6: Strinjanje ocenjevalcev glede ocen berljivosti posameznih modelov

Berljivost								
Model	**Opis modela**	**Ocena 1**	**Ocena 2**	**Ocena 3**	**Ocena 4**	**Ocena 5**	**Berljivost**	**Strinjanje**
A	Domenski JM in kombiniran PM	11	35	15	10	19	**2,90**	**0,24**
B	Gigafida JM in domenski PM	3	30	31	14	12	**3,02**	**0,26**
C	Google Prevajalnik	5	10	31	25	19	**3,48**	**0,25**
D	Domenski JM in domenski PM	6	26	25	13	20	**3,17**	**0,23**
E	Splošni JM in domenski PM	8	42	18	6	16	**2,78**	**0,29**
F	Splošni JM in splošni PM	35	31	11	7	6	**2,09**	**0,29**
Skupaj:		68	174	131	75	92		

Tabela 7.7: Strinjanje ocenjevalcev glede ocen ustreznosti posameznih modelov

Ustreznost								
Model	**Opis modela**	**Ocena 1**	**Ocena 2**	**Ocena 3**	**Ocena 4**	**Ocena 5**	**Ustreznost**	**Strinjanje**
A	Domenski JM in kombiniran PM	3	15	23	21	28	**3,62**	**0,24**
B	Gigafida JM in domenski PM	1	13	16	34	26	**3,79**	**0,27**
C	Google Prevajalnik	0	2	11	25	52	**4,41**	**0,42**
D	Domenski JM in domenski PM	2	10	14	24	40	**4,00**	**0,30**
E	Splošni JM in domenski PM	1	9	29	25	26	**3,73**	**0,27**
F	Splošni JM in splošni PM	10	31	19	18	12	**2,90**	**0,22**
Skupaj:		17	80	112	147	184		

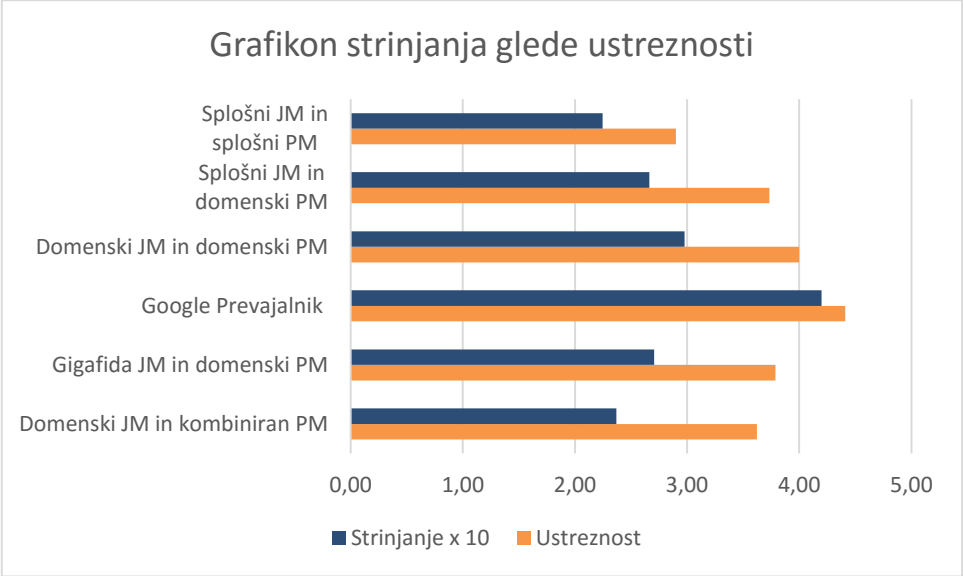
Najprej strinjanje izračunamo na ravni posameznega stavka glede na to, koliko istih ocen na lestvici od 1 do 5 ocenjevalci dodelijo berljivosti ali ustreznosti posameznega stavka. Da lahko podamo oceno strinjanja za posamezen model, nato izračunamo povprečno oceno strinjanja po posameznih stavkih istega modela. V obeh tabelah lahko vidimo tudi, koliko ocen na lestvici od 1 do 5 dobi posamezen model za vseh 90 vprašanj.



Slika 7.16: Grafikon strinjanja glede berljivosti

Na sliki 7.16 je prikazan grafikon strinjanja glede berljivosti po posameznih modelih. Modri stolpec predstavlja raven strinjanja prevajalcev, ki jo pomnožimo z vrednostjo 10, da sta lestvici vrednosti strinjanja in berljivosti lažje primerljivi. Iz grafikona lahko razberemo, da se ocenjevalci najbolj strinjajo, kateri modeli so najmanj berljivi. Strinjanje za modela E in F je po Fleisovi kappi 0,29.

Strinjanje glede ustreznosti je največje glede modelov, ki dobijo najvišje vrednosti pomenske ustreznosti. Kot kaže palični diagram na sliki 7.17, se ocenjevalci najlažje odločijo glede stavkov, ki so pomensko najustreznejši. Ocenjevalci so najbolj enotni, da je prevod pomensko najustrezneje prevedel Google Prevajalnik, nato pa domenski model.



Slika 7.17: Grafikon strinjanja glede ustreznosti

Ocene strinjanja so z najvišjo oceno 0,29 na lestvici od 0 do 1 relativno nizke, kar lahko pripisujemo delno uporabljenemu algoritmu, delno pa dejstvu, da je prevode ocenjevalo 6 ocenjevalcev. Tako je bila denimo kappa strinjanja za primer, ko je 5 ocenjevalcev dalo oceno 5, eden pa oceno 4, enaka 0,65, čeprav bi opisno lahko dejali, da so ocenjevalci skoraj enotni.

Medsebojno strinjanje po parih ocenjevalcev

V prejšnjem podpoglavju je predstavljeno medsebojno strinjanje vseh ocenjevalcev glede ocen berljivosti in ustreznosti. V tem podpoglavju predstavimo medsebojno strinjanje po parih ocenjevalcev. To oceno izračunamo s pomočjo Cohenove kappe [57].

Če želimo izračunati Cohenovo kappo po vseh parih ocenjevalcev, moramo ustvariti tabelo strinjanja glede posameznih odgovorov za vse posamezne pare ocenjevalcev. V tabeli 7.8 je prikazano medsebojno strinjanje med ocenjevalcema A in B. Iz tabele razberemo, da je od vseh 180 ocen ocena 1 dana petkrat. Ocenjevalec A in B dasta v dveh primerih oceno 1 za isto vprašanje, medtem ko v preostalih treh primerih dasta različne ocene. V zeleno obarvani diagonali vidimo število enakih ocen še za druge ocene. Ocenjevalec A in B sta enotna glede ocene v 85 primerih. Pri izračunu Cohenove kappe se upošteva še verjetnost, da ocenjevalca A in B dasta isto oceno povsem slučajno. Izračunana Cohenova kappa za par AB je 31,88.

Tabela 7.8: Medsebojno strinjanje med ocenjevalcema A in B

		A					
		1	2	3	4	5	Skupaj
B	1	**2**	8	3	0	0	13
	2	2	**11**	11	7	0	31
	3	0	12	**17**	4	0	33
	4	1	8	14	**16**	4	43
	5	0	3	7	11	**39**	60
	Skupaj	**5**	**42**	**52**	**38**	**43**	180

Strinjanje	2	11	17	16	39	85
Naključno	0,36	7,23	9,53	9,08	14,33	40,53

Kappa	**31,88**

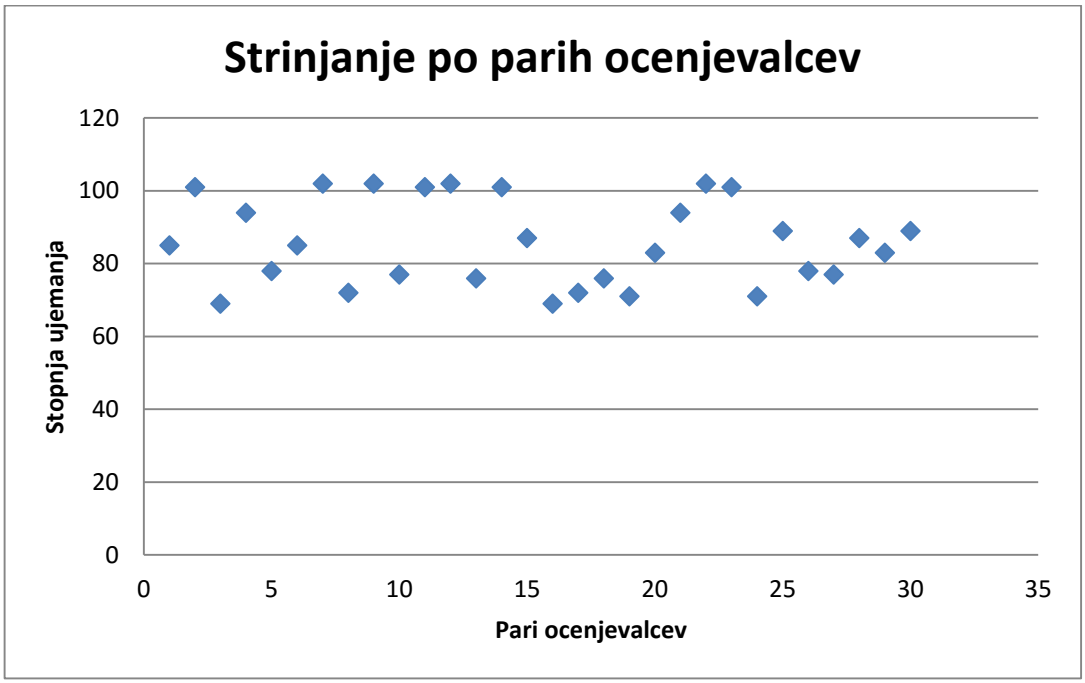
Tabelo, kot je prikazana na sliki 7.8, nato izdelamo še za vse ostale pare ocenjevalcev. Če strinjanje posameznih parov prenesemo v tabelo, dobimo tabelo s povzetkom, prikazano na sliki 7.9.

Tabela 7.9: Povzetek strinjanja med posameznimi pari ocenjevalcev

Št.	Par	Strinjanje	Slučajno	Kappa
1	AB	85	40,53	31,88
2	AC	101	41,38	43,01
3	AD	69	36,08	22,87
4	AE	94	41,06	38,10
5	AF	78	42,33	25,91
6	BA	85	40,53	31,88
7	BC	102	39,55	44,46
8	BD	72	35,61	25,20
9	BE	102	39,63	44,43
10	BF	77	41,54	25,61
11	CA	101	41,38	43,01
12	CB	102	39,55	44,46
13	CD	76	35,98	27,79
14	CE	101	39,82	43,64
15	CF	87	41,02	33,08
16	DA	69	36,08	22,87
17	DB	72	35,61	25,20

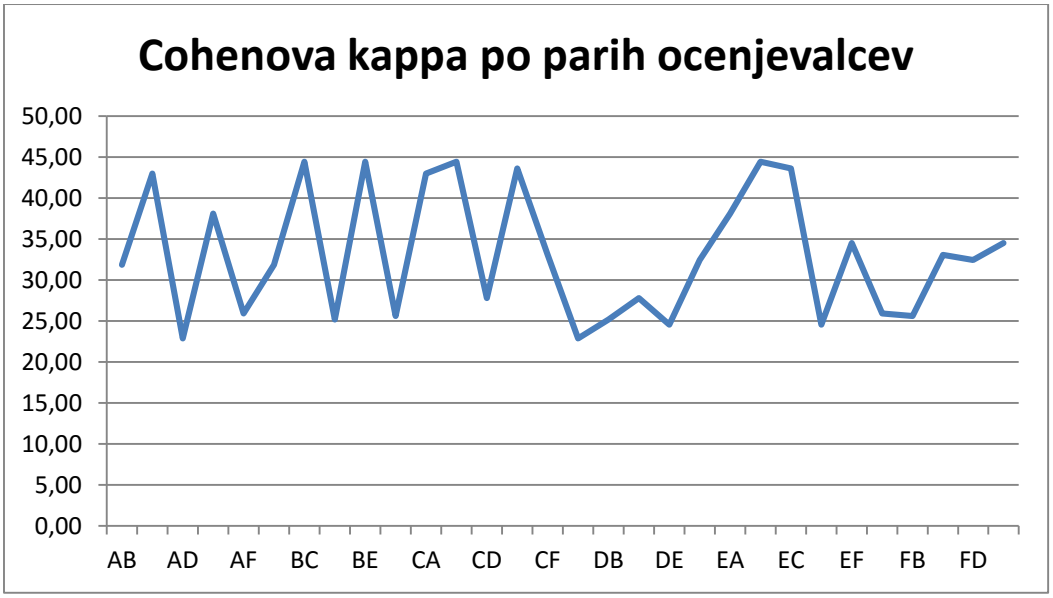
18	DC	76	35,98	27,79
19	DE	71	35,53	24,55
20	DF	83	36,38	32,46
21	EA	94	41,06	38,10
22	EB	102	39,63	44,43
23	EC	101	39,82	43,64
24	ED	71	35,53	24,55
25	EF	89	40,98	34,54
26	FA	78	42,33	25,91
27	FB	77	41,54	25,61
28	FC	87	41,02	33,08
29	FD	83	36,38	32,46
30	FE	89	40,98	34,54
	Povprečje	85,80	39,16	33,17

Iz tabele razberemo, da se med seboj najbolj strinjata ocenjevalca v parih AC, BC in BE. Ostali pari se med seboj strinjajo v od 70 do 102 ocenah od skupno 180 ocen. Iz tabele 7.8 in ostalih tabel strinjanj med posameznimi pari vidimo, da se ocenjevalci sicer ne strinjajo popolnoma glede ocene, je pa večina danih ocen za oceno višja ali nižja od izbrane. Iz tega razberemo, da se ocenjevalci potrudijo in testne stavke dobri preberejo in poskusijo objektivno oceniti. Strinjanje med posameznimi pari grafično prikažemo v grafikonu na sliki 7.18.



Slika 7.18: Strinjanje med ocenjevalci po parih ocenjevalcev

Na sliki 7.18 je prikazano, v koliko ocenah od skupno 180 se posamezen par ocenjevalcev strinja glede ocene. To je sicer dober pokazatelj strinjanja, vendar ne upošteva verjetnosti, da bi se lahko dva ocenjevalca med seboj strinjala povsem slučajno. To verjetnost upošteva Cohenova kappa, ki je prikazana tudi v tabeli 7.9, grafično pa so vrednosti prikazane tudi na sliki 7.19.



Slika 7.19: Cohenova kappa po parih prevajalcev

Za vsakega ocenjevalca lahko izračunamo tudi povprečno vrednost strinjanja z drugimi ocenjevalci. Na ta način dobimo oceno, kako blizu povprečne ocene je ocenjeval posamezen ocenjevalec. Rezultati tega izračuna so prikazani v tabeli 7.10.

Tabela 7.10: Povprečne vrednosti strinjanja posameznih ocenjevalcev

	Povprečja		
Ocenjevalci	**Strinjanje**	**Slučajno**	**Kappa**
A	85,40	40,28	32,35
B	87,60	39,37	34,32
C	93,40	39,55	38,40
D	74,20	35,92	26,57
E	91,40	39,40	37,05
F	82,80	40,45	30,32
Povprečje:	**85,8**	**39,16**	**33,17**

Iz tabele 7.10 razberemo, da je najvišjo vrednost Cohenove kappe dosegel ocenjevalec C, ki se je v povprečju tudi največkrat strinjal z ostalimi ocenjevalci.

7.2.6.3. Vpliv dolžine stavka na oceno

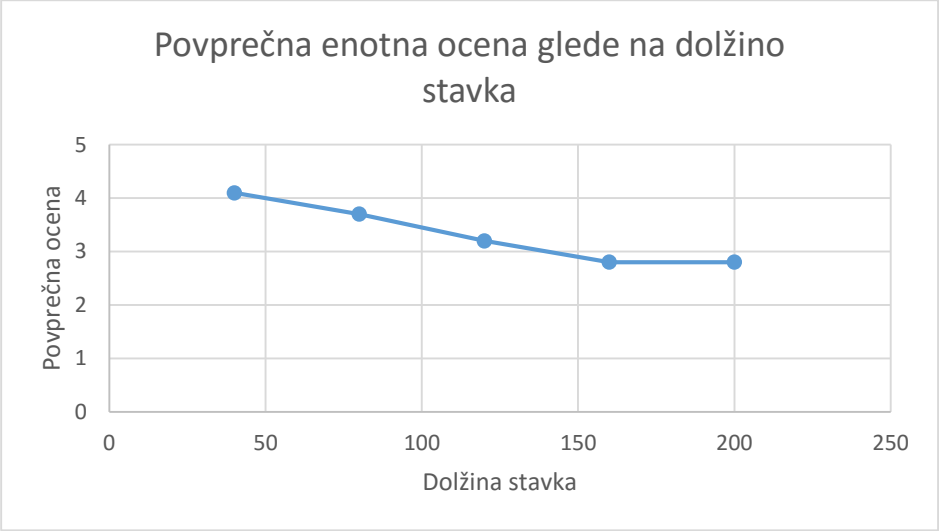
Ko v tabelo razvrstimo vseh 180 odgovorov s povprečno enotno oceno in dolžino stavka po povprečni oceni, opazimo, da je dolžina stavkov, ki dosežejo povprečno oceno 4,8 ali več, krajša od 62 znakov. Za stavke z najslabšo oceno berljivosti ali ustreznosti ne moremo jasno zatrditi, da so najslabše ocenjeni najdaljši stavki, čeprav taka tendenca obstaja. Ugotovimo, da je v delu tabele z najslabšimi ocenami večina stavkov dolžine 80 ali več, nekaj pa je izjem z krajšo dolžino stavka.

Ker nas zanima trend vpliva dolžine stavka na oceno prevoda, stavke razvrstimo po dolžinah stavkov in jih razvrstimo v pet kategorij. Zanje izračunamo povprečno enotno oceno kakovosti prevoda in dobimo vrednosti, ki so prikazane v tabeli 7.11.

Tabela 7.11: Pregled povprečnih ocen stavkov po kategorijah dolžin

Dolžina stavka	**Povprečna enotna ocena**
do 40	4,1
do 80	3,7
do 120	3,2
do 160	2,8
do 200	2,8
Std. odklon	0,51

Dobljene vrednosti smo prikazali tudi v grafikonu na sliki 7.20. Iz grafikona razberemo, da so bili krajši stavki v povprečju ocenjeni bolje od daljših.



Slika 7.20: Grafikon povprečne enotne ocene glede na dolžino stavka

7.2.7. Vpliv velikosti učne zbirke na kakovost prevoda

V večini literature, ki govori o statističnem strojnem prevajanju, poudarjajo, da je velikost učne zbirke ključnega pomena. Večja kot je, boljše rezultate lahko pričakujemo. Zanima nas, ali velja enako tudi za našo domensko zbirko? Ali obstaja morda neka zgornja meja za velikost zbirke, ko se cena BLEU ne povečuje več?

Ker imamo s področja naše domene že zgrajeno zbirko podatkov, zbirke ne moremo povečevati, zato se odločimo, da zbirko postopno zmanjšujemo. Ker zbirke ne želimo zmanjšati zgolj z brisanjem datoteke od določene vrstice naprej, temveč ohraniti raznolikost vsebine, učne datoteke pripravimo tako, da s programom awk v datotekah obdržimo vsako drugo, vsako četrto, vsako peto itn. vrstico.

Pri tem eksperimentu uporabimo jezikovni model Gigafida in različne velikosti domenskega prevajalnega modela. V tabeli 7.12 so prikazani rezultati našega eksperimenta.

Tabela 7.12: Primerjava velikosti učne zbirke in ocen BLEU

Velikost zbirke	**Ocena BLEU**
1.034.800 vrstic	28,94
700.000 vrstic	26,71
520.000 vrstic	26,24
200.000 vrstic	24,24 26,40 (210) | 30,37 (260)
Standardni odklon:	**1,93**

Ugotovimo, da ocena BLEU narašča sorazmerno s povečevanjem učne zbirke podatkov.

Ker v našem eksperimentu učno zbirko zmanjšujemo z brisanjem vrstic po določenem ključu, nam v dveh primerih brisanja z različnima ključema do velikosti 200.000 vrstic uspe doseči oceni BLEU, ki sta višji od ocene, ki jo dobimo po prevajanju z učno zbirko velikosti 500.000 vrstic. Možen vzrok za to vidimo v povečevanju variance pri manjših količinah podatkov.

8. Zaključek

V magistrskem delu uspešno prilagodimo statistični strojni prevajalnik za področje farmacije. Potrdimo naša pričakovanja, da uporaba domenske učne zbirke prinese boljše rezultate od splošne učne zbirke podatkov. Ocena BLEU testnega prevoda, prevedenega z domenskim modelom na osnovi besednih zvez, je 40,74 in je za 23,63 višja od ocene BLEU za testni prevod, preveden s splošnim modelom. V nasprotju z našimi pričakovanji pa kombinirani model, ki je učen z zbirko, sestavljeno iz splošne in domenske zbirke, ne da boljšega rezultata od domenske zbirke. Možnost za izboljšanje kakovosti prevajalnika vidimo v uporabi dveh ločenih prevajalnih modelov v času dekodiranja (prevajanja); domenskega in splošnega, pri čemer damo prednost domenskemu prevajalnemu modelu.

Prilagajanje za specifično domeno izvedemo tudi s faktorskim pristopom statističnega strojnega prevajanja. Tudi tu najboljšo oceno BLEU dobi testni prevod, preveden z domensko zbirko, vendar za 0,75 ocene BLEU nižjo od pristopa na osnovi besednih zvez. S primerjavo obeh domenskih prevodov opazimo, da so nekateri stavki prevoda s faktorskim pristopom bolj berljivi od prevoda s pristopom na osnovi besednih zvez. To pripisujemo dodatnim informacijam, ki jih prinašajo faktorji. Pomembno prednost faktorskega pristopa vidimo v možnosti dodatnega označevanja besedila, s čimer omogočimo strojno prevajanje določenih vrst besedil, ki jih s pristopom na osnovi besednih zvez ni mogoče prevesti. Tu imamo v mislih predvsem besedila z oblikovalskimi oznakami in povezavami.

Za testne prevode opravimo ročno oceno kakovosti. Šest strokovnjakov oceni berljivost in ustreznost izbranega dela testnega prevoda. Ocenjevalci najboljšo oceno berljivosti in ustreznosti dajo prevodu, prevedenim z Google Prevajalnikom (3,48 za berljivost in 4,41 za ustreznost), le malo slabšo pa domenskemu modelu (3,17 za berljivost in 4,0 za ustreznost). Prevod, preveden z Google Prevajalnikom doseže oceno BLEU zgolj 22,17 v primerjavi z domenskim modelom, ki doseže oceno 40,74. Tako velika razlika nas preseneti in nas vodi v razmišljanje, ali je ocena BLEU primerna za orientacijo glede kakovosti posameznega modela med več modeli v času načrtovanja prevajalnega modela. Za realno oceno modela je treba izvesti ročno ocenjevanje.

Po končani ročni oceni izvedemo izračun strinjanja prevajalcev. Ugotovimo, da so ocenjevalci pri ročnem ocenjevanju ustreznosti najbolj enotni glede pomensko najustreznejših prevodov, medtem ko so pri ocenjevanju berljivosti najbolj enotni glede najslabše berljivih prevodov. Iz

dobljenih ocen lahko razberemo, da so ocenjevalci najlažje prepoznali na eni strani stavke z največ slovničnimi napakami in na drugi strani stavke, ki so jim bili pomensko najrazumljivejši.

Če prevajalne modele pripravimo z domensko učno zbirko podatkov, ki daje dobre rezultate glede ustreznosti, menimo, da se za najboljšega odločimo na podlagi ocene za berljivost, ker je z vidika naravnosti prevoda ta vidik pomembnejši. Berljivost prevoda je v največji meri odvisna od kakovostnega jezikovnega modela. V našem eksperimentu ugotovimo, da jezikovni model, zgrajen iz korpusa slovenskega jezika ccGigafida, doseže najboljšo oceno berljivosti v primerjavi z drugima dvema jezikovnima modeloma, ki ju še opazujemo. Menimo, da ga je smiselno uporabiti tudi v drugih modelih, bodisi kot glavnega bodisi kot pomožnega v primeru uporabe dveh jezikovnih modelov. Jezikovni model ccGigafida damo na voljo tudi širši raziskovalni javnosti.

Kakovost strojnega prevajalnika je odvisna tudi od ustreznega uglaševanja. V našem primeru uporabimo algoritem MERT, ki je optimiziran za maksimalno oceno BLEU. Menimo, da daje dobre rezultate, če želimo imeti v prevodih izrazoslovje čim bolj enako kot v učni zbirki. V našem primeru je to zelo zaželeno, ker prilagajamo strojni prevajalnik za strokovno specifično področje, kjer se zahteva enotna terminologija in oblika stavkov. Čeprav algoritem MERT daje dobre rezultate, je vredno poskusiti tudi z drugimi optimizacijskimi algoritmi.

Za približevanje ciljni domeni uporabimo tudi domenski slovar. Testni prevod, preveden z dodatkom slovarja, je bil za 1,11 ocene BLEU boljši od prevoda brez dodanega slovarja. Dobljeni rezultat kaže, da uporaba slovarja pripomore k izboljšanju ocene BLEU. Menimo, da je uporaba slovarja še posebej koristna, kadar želimo za projekt z več dokumenti zagotoviti uporabo skladne terminologije za specifične izraze iz slovarja.

Na podlagi ocen ročnega ocenjevanja berljivosti in ustreznosti ugotovimo, da stavki, krajši od 62 znakov, dobijo skupno oceno (povprečje med ustreznostjo in berljivostjo) višjo od 4,8. Za stavke z najslabšo oceno berljivosti ali ustreznosti ne moremo jasno zatrditi, da so najslabše ocenjeni najdaljši stavki, čeprav taka tendenca obstaja.

Testni dokument sestavimo iz treh predhodno prevedenih navodil za uporabo zdravil, od katerih je eden vsebovan tudi v učni zbirki, druga dva ne. Opazimo slabost statističnega strojnega prevajalnika, da stavki, ki so vsebovani v učni zbirki, v testni prevod niso preneseni v celoti. Tako delovanje je sicer normalen način delovanja statističnega strojnega prevajalnika, vendar nas vodi v razmišljanje, da je za prevajanje dokumentov s predpisano obliko in izrazoslovjem

dobro vzpostaviti sistem, ki omogoča najprej analizo stavkov dokumenta s 100 % ujemanjem iz pomnilnika prevodov, nato pa obdelavo preostanka besedila s strojnim prevajalnikom. V tem pristopu vidimo možnost za nadaljnjo raziskavo in nadgradnjo.

Pri analiziranju kakovosti strojnega prevoda moramo upoštevati tudi, da je končna uporabnost rezultatov prevajanja odvisna tudi od namena uporabe. Če nameravamo strojne prevode uporabiti kot osnovo za popravljanje strojnih prevodov s strani prevajalcev, moramo poleg v tem delu izvedenih ocen izvesti še oceno truda, potrebnega za popravljanje strojnega prevoda (*angl. post-editing effort*), in strojni prevajalnik prilagoditi temu.

Raziskovano področje predstavlja številne možnosti za nadaljnje raziskave. V okviru modela na osnovi besednih zvez bi bilo zanimivo povečati učno zbirko s prevajanjem domenske zbirke, kot je denimo PubMed ali podobna farmacevtska zbirka, z uporabo enega od javno dostopnih strojnih prevajalnikov z dobrimi rezultati, kot denimo Google Prevajalnik. Poleg tega bi bilo za primerjavo zanimivo vzpostaviti tudi slovensko-angleški strojni prevajalnik, ki bi morda zaradi manjše oblikoslovne pestrosti angleškega jezika prinesel boljše rezultate berljivosti.

Na področju izboljševanja berljivosti največjo priložnost vidimo v vzpostavitvi prevajalnika na osnovi globokih mrež (*angl. Neural Machine Translation*), ki je z Googlovim odprtjem razčlenjevalnika SyntaxNET za javnost postal na voljo kot odprtokodni sistem, v kombinaciji z ustreznim razčlenjevalnikom slovenskega jezika.

Trenutni trendi na področju strojnega prevajanja kažejo, da se, podobno kot pred leti za prevajanje na osnovi besednih zvez, zdaj raziskovalno področje seli na področje strojnega prevajanja na osnovi globokih mrež. Menimo, da se bo kakovost strojnih prevajalnikov z uporabo kombinacije več pristopov izboljšala tudi za slovenski jezik, še posebej, ker se bo s časom povečala količina javno dostopnih zbirk v slovenskem jeziku.

9. Literatura

[1] Automatic Language Processing Advisory Committee (ALPAC), *"Language and machines: Computers in translation and linguistics,"* Wahington, 1966.

[2] Apsic. Apsic Xbench. [Spletni vir]. http://www.xbench.net/ [Dostop: 5. 5. 2016]

[3] Mihael Arčan, Marco Turchi, Sara Tonelli in Paul Buitelaar, "Enhancing statistical machine translation with bilingual terminology in a CAT environment," v zborniku *Association for Machine Translation in the Americas (AMTA) Conference 2014*, Vancouver, 2014, str. 54–68.

[4] Tadeja Artiček, *Uvajanje strojnega prevajanja v delo prevajalske agencije: analiza popravljanja strojnih prevodov*. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta, 2014.

[5] Amittai Axelrod, *Factored language models for statistical machine translation*. Edinburgh: Institute for Communicating and Collaborative Systems, University of Edinburgh, 2006.

[6] Dzmitry Bahdanu in KyungHyun Cho, "Neural machine translation by jointly learning to align and translate," v zborniku *International Conference on Learning Representations* , San Diego, 2015, str. 1–15.

[7] Nicola Bertoldi, Mauro Cettolo in Marcello Federico, "Cache-based online adaptation for machine translation enhanced computer assisted translation," v zborniku *XIV Machine Translation Summit*, Nica, 2013, str. 35–42.

[8] Jeff A. Bilmes in Katrin Kirchhoff, "Factored language models and generalized parallel backof," v zborniku *Conference of the North American Chapter of the Association for Computational Linguistics on human language technology*, Edmonton, 2003.

[9] Ivan Bulyko, Spyros Matsoukas, Richard Schwartz, Long Nguyen in John Makhoul, "Language model adaptation in machine translation from speech," v zborniku *BBN Technologies*, 2007, str. 117–120.

[10] Eugene Charniak, Kevin Knight in Kenji Yamada, "Syntax-based language models for statistical machine translation," v zborniku *MT Summit iX International Association for Machine Translation*, 2003.

[11] Jani Dugonik, *Uglaševanje parametrov pri statističnem strojnem prevajanju*. Maribor: Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru, 2013.

[12] Europarl. European Parliament Proceedings Parallel Corpus 1996–2011. [Spletni vir]. http://www.statmt.org/europarl/ [Dostop: 10. 5. 2016]

[13] Marcello Federico, Nicola Bertoldi in Mauro Cettolo, "IRSTLM: an open source toolkit for handling large scale language models," v zborniku *Interspeech*, Brisbane, 2008, str. 1618–1621.

[14] Qin Gao. Mgiza. [Spletni vir]. http://48-191.hostmonster.com/software/doku.php/mgiza:overview [Dostop: 23. 4. 2016]

[15] Google Prevajalnik. (2016) [Spletni vir]. https://translate.google.com/ [Dostop: 12. 4. 2016]

[16] Miha Grčar , Matjaž Juršič, Jan Rupnik, Simon Krek in Kaja Dobrovoljc. Obeliks. [Spletni vir]. http://oznacevalnik.slovenscina.eu/Vsebine/Sl/ProgramskaOprema/Navodila.aspx [Dostop: 15. 6. 2016]

[17] Eva Hasler, Barry Haddow in Philipp Koehn, *Margin Infused Relaxed Algorithm for Moses*. University of Edinburgh, Institute for Language, Cognition and Computation, 2011.

[18] Kenneth Heafield. KenLM Language Model Toolkit. [Spletni vir]. http://kheafield.com/code/kenlm/ [Dostop: 5. 5. 2016]

[19] Kenneth Heafield, "KenLM: Faster and Smaller Language Model Queries," v zborniku *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 2011, str. 187–197.

[20] Julia Hirschberg in Christopher D. Manning, "Science," *Advances in natural language processing*, str. 261–266, 349 2015.

[21] The-European-Commission's- Joint-research-centre. European Commission's Directorate-General for Translation. [Spletni vir]. https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory [Dostop: 13. 4. 2016]

[22] Daniel Jurafsky in James H. Martin, *Speech and Language Processing*, 2nd ed. New Yersey: Person Education Inc., 2009.

[23] Adam Kilgarriff in sod. SketchEngine. [Spletni vir]. https://downloads.sketchengine.co.uk/325445fda_models.tar.xz [Dostop: 8. 7. 2016]

[24] Philipp Koehn. Statistical Machine Translation - Syntax-Based Models. [Spletni vir]. http://homepages.inf.ed.ac.uk/pkoehn/publications/esslli-slides-day5.pdf [Dostop: 12. 4. 2016]

[25] Philipp Koehn in Hieu Hoang, *Moses - Statistical machine translation system, User manual and code guide*. Edinburgh: University of Edinburgh, 2016.

[26] Philipp Koehn in sod. Moses: Open source toolkit for statistical machine translation system. 2007 [Spletni vir]. http://homepages.inf.ed.ac.uk/pkoehn/publications/acl2007-moses.pdf [Dostop: 12. 2. 2016]

[27] Philipp Koehn, Franz Josef Och in Daniel Marcu, *Statistical phrase based translation*.: North American Chapter of the Association for Computuinal Lingustics: Human Language Technologies (HLT-NAACL-03), 2003.

[28] Philipp Koehn in Josh Schroeder, "Experiments in Domain Adaptation for Statistical Machine Translation," v zborniku *Second Workshop on Statistical Mahine Translation*, Praga, 2007, str. 224–227.

[29] Nataša Logar Berginc, Tomaž Erjavec, Simon Krek, Miha Grčar in Peter Halozan. Javnodostopna zbirka ccGigafida. [Spletni vir]. https://www.clarin.si/repository/xmlui/handle/11356/1035 [Dostop: 15. 2. 2016]

[30] Adam Lopez, "ACM Computing Surveys," *Statistical machine translation*, p. članek 8, 40 2008.

[31] Fakulteta za naravoslovje in matematiko Univerze v Mariboru. (2004) Trojezični terminološki slovar kemijskih pojmov.

[32] Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage in Yoav Goldberg, "Universal dependency annotation for multilingual parsing," v zborniku *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, 2013, str. 92–97.

[33] Moravia. Pharmaceuticals translation services. [Spletni vir]. http://www.moravia.com/en/services/life-sciences/pharmaceuticals/ [Dostop: 23. 4. 2016]

[34] Dragos Stefan Munteanu in Daniel Marcu, "Improving machine translation performance by exploting non-parallel corpora," *Computational linguistics*, vol. 41(3), str. 477–504, 2005.

[35] Jan Nieheus in Alex Waibel, *Domain adaptation in statistical machine translation using factored translation models*. Karlsruhe: Institute for Anthropomatics, Karlsruhe Institute for Technology. [Spletni vir]. http://isl.anthropomatik.kit.edu/downloads/921_eamt10.pdf [Dostop: 16. 4. 2016]

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

[37] Franz Josef Och in Hermann Ney, "Computational Linguistics," *A Systematic Comparison of Various Statistical Alignment Models*, str. 19–51, Jan. 2003.

[38] OpenTM2. The Open Translation Manager. [Spletni vir]. http://www.opentm2.org/ [Dostop: 5. 5. 2016]

[39] Kishore Papineni, Salim Roukos, Todd Ward in Wei-Jing Zhu, "BLEU: a method for Automatic Evaluation of Machine Translation," v zborniku *Proceedings of the 40th Annual Meeting for Computational Linguistic (ACL)*, Philadelphia, 2002, str. 311–318.

[40] Slav Petrov. Google Research Blog. [Spletni vir]. https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html?m=1 [Dostop: 10. 7. 2016]

[41] Helmut Schmid. TreeTagger. [Spletni vir]. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ [Dostop: 26. 4. 2016]

[42] School of medicine and health sciences. Back-translation guidelines. [Spletni vir]. http://www.med.und.edu/genacis/back-translation-guidelines.cfm [Dostop: 23. 4. 2016]

[43] Rico Sennrich, Barry Haddow in Alexandra Birch. (2016) Improving neural machine translation models with monolingual Data. [Spletni vir]. http://arxiv.org/pdf/1511.06709v4.pdf [Dostop: 10. 7. 2016]

[44] Inguna Skadina in sod., "ACCURAT - Collecting and Using Comparable Corpora for Statistical Machine Translation" , 2013.

[45] Andreas Stolcke, "SRILM – An extensible language modeling toolkit," v zborniku *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, 2002, str. 901–904.

[46] TAUS - Translation Automation User Society. TAUS Adequacy/Fluency Guidelines. [Spletni vir]. https://www.taus.net/academy/best-practices/evaluate-best-practices/adequacy-fluency-guidelines [Dostop: 7. 6. 2016]

[47] Jörg Tiedemann in Lars Nygaard. Opus - the open parallel corpus. [Spletni vir]. http://opus.lingfil.uu.se/ [Dostop: 17. 5. 2016]

[48] Nicola Ueffing, Gholamreza Haffari in Anoop Sarkar. (2008) Semi-supervised model adaptation for statistical machine translation.

[49] Aljoša Vrščaj, *Evalvalcija strojnih prevajalnikov*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za prevajalstvo, 2011.

[50] John S White, Theresa O'Connell in Francis O'Mara. The ARPA MT Evaluations Methodologies: Evolution, Lessons in Future Approaches. [Spletni vir]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.137.1288&rep=rep1&type=pdf [Dostop: 15. 2. 2016]

[51] Wikipedia. Fleiss Kappa. [Spletni vir]. https://en.wikipedia.org/wiki/Fleiss%27_kappa [Dostop: 14. 6. 2016]

[52] Wikipedia. pdftotext. [Spletni vir]. https://en.wikipedia.org/wiki/Pdftotext [Dostop: 5. 5. 2016]

[53] Angleška Wikipedia. Statistical machine translation. [Spletni vir]. https://en.wikipedia.org/wiki/Statistical_machine_translation [Dostop: 20. 4. 2016]

[54] Wikipedia. (2016, Feb.) Strojno prevajanje. [Spletni vir]. https://sl.wikipedia.org/wiki/Strojno_prevajanje [Dostop: 15. 2. 2016]

[55] Hua Wu, Haifend Wang in Chengqing Zong, "Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora," v zborniku *Coling 2008*, Manchester, 2008, str. 993–1000

[56] Deyi Xiong, Min Zhang in Haizhou Li, "Enhancing Language Models in Statistical Machine Translation with Backward N-grams and Mutual Information Triggers," v zborniku *49th Annual Meeting of the Association for Computational Linguistics*, Portland, 2011, str. 1288–1297.

[57] Charles Zaiontz. RealStatistics. [Spletni vir]. http://www.real-statistics.com/reliability/cohens-kappa/ [Dostop: 8. 7. 2016]

[58] Marija Žakelj - Mavrič in Marko Dolinar, *Angleško-slovenski slovar izbranih izrazov iz biokemije in*. Ljubljana: Slovensko biokemijsko društvo, 2012.

10. Priloge

Priloga 1: Testni dokument za ocenjevanje strojnega prevajalnika

Priloga 2: Kratek opis postopka za namestitev Mosesa in gradnjo prevajalnega modela

Priloga 3: Optimizacija MERT in zgled inicializacijskih datotek

Priloga 4: Skripta za pretvarjanje oblike ttpos v obliko za Moses (*ttpos-to-moses.sh*)

Priloga 5: Skripta za pretvarjanje Obeliks oblike .xml v Moses (*xml-to-moses.pl*)

Priloga 1: Testni dokument za ocenjevanje strojnega prevajalnika

V tej prilogi je vsebina celotnega testnega dokumenta, ki smo ga uporabljali za samodejno ocenjevanje prevajalnih modelov. Z modro barvo so označeni posamezni deli, ki smo jih uporabili za ročno ocenjevanje.

Dasselta

desloratadine

This is a summary of the European public assessment report (EPAR) for Dasselta. It explains how the Committee for Medicinal Products for Human Use (CHMP) assessed the medicine to reach its opinion in favour of granting a marketing authorisation and its recommendations on the conditions of use for Dasselta.

What is Dasselta?

Dasselta is a medicine containing the active substance desloratadine. It is available as tablets (5 mg).

Dasselta is a 'generic medicine'. This means that Dasselta is similar to a 'reference medicine' already authorised in the European Union (EU) called Aerius. For more information on generic medicines, see the question-and-answer document here.

What is Dasselta used for?

Dasselta is used to relieve the symptoms of allergic rhinitis (inflammation of the nasal passages caused by an allergy, for example, hay fever or allergy to dust mites) or urticaria (a skin condition caused by an allergy, with symptoms including itching and hives).

The medicine can only be obtained with a prescription.

How is Dasselta used?

The recommended dose for adults and adolescents (12 years of age and over) is one tablet once a day.

How does Dasselta work?

The active substance in Dasselta, desloratadine, is an antihistamine. It works by blocking the receptors on which histamine, a substance in the body that causes allergic symptoms, normally fixes itself. When the receptors are blocked, histamine cannot have its effect, and this leads to a decrease in the symptoms of allergy.

How has Dasselta been studied?

Because Dasselta is a generic medicine, studies in patients have been limited to tests to determine that it is bioequivalent to the reference medicine, Aerius. Two medicines are bioequivalent when they produce the same levels of the active substance in the body.

What are the benefits and risks of Dasselta?

Because Dasselta is a generic medicine and is bioequivalent to the reference medicine, its benefits and risks are taken as being the same as the reference medicine's.

Why has Dasselta been approved?

The CHMP concluded that, in accordance with EU requirements, Dasselta has been shown to have comparable quality and to be bioequivalent to Aerius. Therefore, the CHMP's view was that, as for Aerius, the benefit outweighs the identified risk. The Committee recommended that Dasselta be given marketing authorisation.

Other information about Dasselta

The European Commission granted a marketing authorisation valid throughout the European Union for Dasselta on 28 November 2011.

The full EPAR for Dasselta can be found on the Agency's website: ema.europa.eu/Find medicine/Human medicines/European Public Assessment Reports. For more information about treatment with Dasselta, read the package leaflet (also part of the EPAR) or contact your doctor or pharmacist.

This summary was last updated in 10-2011.

Deltyba

delamanid

This is a summary of the European public assessment report (EPAR) for Deltyba. It explains how the Agency assessed the medicine to recommend its authorisation in the EU and its conditions of use. It is not intended to provide practical advice on how to use Deltyba.

For practical information about using Deltyba, patients should read the package leaflet or contact their doctor or pharmacist.

What is Deltyba and what is it used for?

Deltyba is a tuberculosis medicine that contains the active substance delamanid. Tuberculosis is an infection caused by the bacterium Mycobacterium tuberculosis (M. tuberculosis). Deltyba is used in adults with tuberculosis that is affecting the lung and that is multi-drug resistant (resistant to at least isoniazid and rifampicin, two standard anti-tuberculosis medicines). It is used together with other standard medicines and when other combinations without this medicine cannot be used either because the disease is resistant to them or because of their side effects.

Because the number of patients with tuberculosis is low in the EU, the disease is considered 'rare', and Deltyba was designated an 'orphan medicine' (a medicine used in rare diseases) on 1 February 2008.

How is Deltyba used?

Deltyba can only be obtained with a prescription and treatment should be started and monitored by a doctor who is experienced in the treatment of multi-drug resistant tuberculosis.

The medicine is available as tablets (50 mg) and the recommended dose is two tablets twice a day taken together with food. Deltyba is given for 6 months together with other standard medicines. Treatment with these standard medicines should continue as recommended by official guidelines after completion of Deltyba treatment. For further information, see the package leaflet.

How does Deltyba work?

The active substance in Deltyba, delamanid, is an antibiotic active against M. tuberculosis. Although the precise mode of action is unclear, delamanid is known to block the production of methoxy-mycolic and keto-mycolic acids, two essential components of the cell walls of M. tuberculosis, which will cause the bacteria to die.

What benefits of Deltyba have been shown in studies?

The effects of Deltyba have been looked at in one main study involving 481 adults with tuberculosis resistant to standard treatments. Patients in the study were given Deltyba or placebo (a dummy treatment) for 2 months in addition to their other treatments. The main measure of effectiveness was the proportion of patients who no longer had the bacteria in their sputum (phlegm). After 2 months of treatment more than 40% of the patients who were taking Deltyba no longer had the bacteria in their sputum compared with 30% of the patients who were taking placebo.

After the main study had finished, patients had the option to receive treatment with Deltyba for 6 months in an extension study. In addition, a majority of patients who entered the main study were followed up for up to 24 months afterwards. Looking at the results of these follow-up studies together, 2 years after starting treatment 75% of patients who received Deltyba for 6 months or more had no bacteria in their sputum compared with 55% of patients who received Deltyba for 2 months or less.

What are the risks associated with Deltyba?

The most common side effects with Deltyba (which may affect around a third of patients) are nausea, vomiting and dizziness. The most serious side effect is QT prolongation (an alteration of the electrical activity of the heart which can cause a life-threatening abnormality of heart rhythm). Other important side effects are anxiety, paraesthesia (unusual sensations like pins and needles) and tremor (shaking). For the full list of all side effects reported with Deltyba, see the package leaflet.

Deltyba must not be used in patients who have low levels of albumin (a blood protein). It must also not be used in patients who are taking certain other medicines that affect the way Deltyba is broken down in the body. For the full list of restrictions, see the package leaflet.

Why is Deltyba approved?

The Agency's Committee for Medicinal Products for Human Use (CHMP) decided that Deltyba's benefits are greater than its risks and recommended that it be approved for use in the EU. The Committee considered that the benefits of Deltyba had been shown for patients with multi-drug resistant tuberculosis affecting the lung. Although the main study was of short duration and the follow-up studies had shortcomings, the CHMP took the view that the effects shown after the initial 2 months of treatment are likely to be sustained for the full treatment

duration. The CHMP noted that an on-going clinical study will provide confirmation on the long-term effectiveness. In addition, the CHMP required that an additional study should be carried out to confirm that the current recommended dose is the most appropriate dose.

Regarding the safety of Deltyba, the safety profile was considered manageable and several measures were introduced to minimise the risks, including a study to confirm the long-term safety. Furthermore the Committee highlighted the medical need for new agents to treat multi-drug resistant tuberculosis.

Deltyba has been given 'conditional approval'. This means that there is more evidence to come about the medicine, which the company is required to provide. Every year, the European Medicines Agency will review any new information that becomes available and this summary will be updated as necessary.

What information is still awaited for Deltyba?

Since Deltyba has been granted a conditional approval, the company that markets Deltyba will carry out further studies to confirm the long-term effectiveness and safety of Deltyba. A further study will also be carried out to confirm the most appropriate dose.

What measures are being taken to ensure the safe and effective use of Deltyba?

A risk management plan has been developed to ensure that Deltyba is used as safely as possible. Based on this plan, safety information has been included in the summary of product characteristics and the package leaflet for Deltyba, including the appropriate precautions to be followed by healthcare professionals and patients.

In addition, the company that markets Deltyba will provide educational material for healthcare professionals, explaining how to use the medicine safely in order to avoid problems such as the development of resistance and side effects on the heart, as well as the risks in pregnancy or women who are breast-feeding.

Other information about Deltyba

The European Commission granted a marketing authorisation valid throughout the European Union for Deltyba on 28 April 2014.

The full EPAR for Deltyba can be found on the Agency's website: ema.europa.eu/Find medicine/Human medicines/European public assessment reports. For more information about treatment with Deltyba, read the package leaflet (also part of the EPAR) or contact your doctor or pharmacist.

The summary of the opinion of the Committee for Orphan Medicinal Products for Deltyba can be found on the Agency's website: ema.europa.eu/Find medicine/Human medicines/Rare disease designation.

This summary was last updated in 04-2014.

Mekinist

trametinib

This is a summary of the European public assessment report (EPAR) for Mekinist. It explains how the Agency assessed the medicine to recommend its authorisation in the EU and its conditions of use. It is not intended to provide practical advice on how to use Mekinist.

For practical information about using Mekinist, patients should read the package leaflet or contact their doctor or pharmacist.

What is Mekinist and what is it used for?

Mekinist is a cancer medicine used to treat adults with melanoma (a type of skin cancer) that has spread or cannot be surgically removed. It is used on its own or combination with another cancer medicine, dabrafenib.

Mekinist is only for patients whose melanoma cells have been tested and shown to have a specific genetic mutation (change) in their genes called 'BRAF V600'.

Mekinist contains the active substance trametinib.

How is Mekinist used?

Treatment with Mekinist must be started and supervised by a doctor experienced in the use of cancer medicines. The medicine can only be obtained with a prescription.

Mekinist is available as tablets (0.5 mg, 1 mg and 2 mg). The dose of Mekinist either used alone or in combination with dabrafenib is 2 mg once a day, at a similar time every day.

Mekinist should be taken without food, at least 1 hour before or 2 hours after a meal. Treatment may need to be interrupted or stopped, or the dose reduced, if the patient experiences certain side effects. For further information, see the summary of product characteristics (also part of the EPAR).

How does Mekinist work?

In melanoma with the BRAF V600 mutation, an abnormal form of the protein BRAF is present, which switches on another protein called MEK involved in stimulating cell division. This encourages cancer to develop by allowing uncontrolled division of cells. The active substance in Mekinist, trametinib, works by blocking MEK directly and by preventing its activation by BRAF thereby slowing down the growth and spread of the cancer. Mekinist is only given to patients whose melanoma is caused by the BRAF V600 mutation.

What benefits of Mekinist have been shown in studies?

Mekinist has been studied in one main study involving 322 patients with melanoma that had spread to other parts of the body or could not be surgically removed, and whose melanoma had the BRAF V600 mutation. Mekinist alone was compared with the cancer medicines dacarbazine or paclitaxel, and the main measure of effectiveness was how long patients lived until their disease got worse (progression-free survival). In this study Mekinist was more effective than dacarbazine or paclitaxel in controlling the disease: patients taking Mekinist lived on average for 4.8 months without their disease getting worse, compared with 1.5 months for patients given dacarbazine or paclitaxel.

In an additional study Mekinist did not show any benefit when given to patients who did not respond to previous treatment with another medicine called a BRAF inhibitor.

In one of two studies of Mekinist in combination with dabrafenib, 423 patients were given either the combination or dabrafenib alone. The result was that patients given the combination lived for 11 months without their disease worsening, while those given dabrafenib alone lived for 8.8 months without their disease worsening. In a second study involving 704 patients, Mekinist with dabrafenib was compared with another medicine for melanoma, vemurafenib. Patients given the combination lived longer on average, 25.6 months versus 18 months with vemurafenib.

What are the risks associated with Mekinist?

The most common side effects with Mekinist (which may affect more than 1 in 5 people) are rash, diarrhoea, fatigue, peripheral oedema (swelling, especially of ankles and feet), nausea and dermatitis acneiform (acne-like inflammation of the skin).

When Mekinist is taken in combination with dabrafenib the most common side effects (seen more than 1 in 5 people) are fever, tiredness, nausea, headache, chills, diarrhea, rash, joint pain, high blood pressure, vomiting and cough.

For the full list of all side effects and restrictions with Mekinist, see the package leaflet.

Why is Mekinist approved?

The Agency's Committee for Medicinal Products for Human Use (CHMP) decided that Mekinist's benefits are greater than its risks and recommended that it be approved for use in the EU. The Committee considered that Mekinist when used alone or in combination with dabrafenib had shown a clinically relevant benefit in patients whose melanoma had a BRAF V600 mutation. In terms of safety, the side effects were considered acceptable and manageable with appropriate measures.

What measures are being taken to ensure the safe and effective use of Mekinist?

A risk management plan has been developed to ensure that Mekinist is used as safely as possible. Based on this plan, safety information has been included in the summary of product characteristics and the package leaflet for Mekinist, including the appropriate precautions to be followed by healthcare professionals and patients.

Further information can be found in the summary of the risk management plan.

Other information about Mekinist

The European Commission granted a marketing authorisation valid throughout the European Union for Mekinist on 30 June 2014.

The full EPAR and risk management plan summary for Mekinist can be found on the Agency's website: ema.europa.eu/Find medicine/Human medicines/European public assessment reports. For more information about treatment with Mekinist, read the package leaflet (also part of the EPAR) or contact your doctor or pharmacist.

Priloga 2: Kratek opis postopka za namestitev Mosesa in gradnjo prevajalnega modela

V tem dodatku so napisana kratka navodila za namestitev sistema Moses in gradnjo prevajalnega modela. V navodila smo dopisali nekaj naših opažanj in zapisali rešitve. Za podrobnejša navodila predlagamo, da bralec obišče Mosesovo spletno stran na naslovu http://www.statmt.org/moses/?n=Development.GetStarted.

Kratka navodila za postavitev prevajalnega sistema Moses

1. Najprej moramo namestiti predpogojne pakete Ubuntu za gradnjo Mosesa in pripadajočih programov:
   ```
   sudo apt-get install build-essential git-core pkg-config automake
   libtool wget zlib1g-dev python-dev libbz2-dev
   ```

2. Za regresijske teste potrebujemo tudi

   ```
   sudo apt-get install libsoap-lite-perl
   ```

3. Zdaj lahko prenesemo in namestimo Moses iz repozitorija. Nato se postavimo v imenik za gradnjo Mosesa
   ```
   git clone https://github.com/moses-smt/mosesdecoder.git
   cd mosesdecoder
   ```

4. Z naslednjim ukazom namestimo najnovejšo različico Boosta, ker je privzeta različica v našem sistemu morda prestara. Namestimo tudi nekaj ostalih z Mosesom povezanih programov:

   ```
   make -f contrib/Makefiles/install-dependencies.gmake
   ```

5. Za prevajanje Mosesa zaženemo ukaz:

   ```
   ./compile.sh [dodatni parametri]
   ```

- `--prefix=/destination/path --install-scripts`
  ... če želimo Moses namestiti na drugo mesto v našem sistemu
- `--with-mm`
  ...če želimo omogočiti prevajalne tabele kot podatkovna polja

Navodila za postavitev osnovnega prevajalnega modela za prevajanje z modelom na osnovi besednih zvez

Pregled

Ko imamo nameščen sistem Moses, lahko z vzporednimi podatki zgradimo realen prevajalski sistem na osnovi besednih zvez. Proces lahko izvajamo na prenosnem računalniku, vendar lahko to traja kakšen dan, potrebujemo pa vsaj 2 GB pomnilnika RAM ter 10 GB prostora na disku. (To je samo ocena, dejanska količina je odvisna od velikosti zbirke.)

Minimalne zahteve glede programske opreme so:

- Moses,
- GIZA++ za poravnavo besed v vzporedni zbirki,
- Eden od programov za gradnjo jezikovnega modela (IRSTLM, SRILM ali KenLM).

KenLM je privzeto vključen v Moses.

Za vsa orodja in podatke smo izbrali domači imenik (tj. `~/`) (*angl. home*). Moses je nameščen v imeniku `~/mosesdecoder` in ga bomo v tem imeniku tudi izvajali.

Nameščanje orodja GIZA++

Orodje GIZA++ potrebujemo za učenje prevajalnega sistema. Čeprav je namestitev na voljo na spletni strani Google Code, je najboljše, da najnovejšo različico prenesemo z orodjem svn:

```
git clone https://github.com/moses-smt/giza-pp.git
cd giza-pp
make
```

S tem ukazom se ustvarijo binarne datoteke `~/giza-pp/GIZA++-v2/GIZA++`, `~/giza-pp/GIZA++-v2/snt2cooc.out` in `~/giza-pp/mkcls-v2/mkcls`. Te moramo prekopirati v lokacijo, kjer jih Moses lahko najde, na primer:

```
cd ~/mosesdecoder
mkdir tools
cp ~/giza-pp/GIZA++-v2/GIZA++ ~/giza-pp/GIZA++-v2/snt2cooc.out \
~/giza-pp/mkcls-v2/mkcls tools
```

Ko zaženemo učenje, moramo skriptu za učenje z argumentom `-external-bin-dir` povedati, kje je nameščeno orodje GIZA++.

```
train-model.perl -external-bin-dir $HOME/mosesdecoder/tools
```

Pozor! Če boste namesto orodja GIZA++ uporabljali orodje MGIZA, morate Mosesu v parametrski datoteki podati pot do programa. V nasprotnem primeru se bo izvajanje Mosesa morda ustavilo z napako:

```
ERROR: Cannot find mkcls, GIZA++/mgiza, &
snt2cooc.out/snt2cooc in /home/kjoze/mosesdecoder/tools.
You MUST specify the parameter -external-bin-dir at
/home/kjoze/mosesdecoder/scripts/training/train-model.perl
line 488.
```

V našem primeru Moses ni želel delovati tudi, ko smo podali ustrezne parametre. Potem ko smo poleg MGIZA-e namestili program GIZA++, je Moses začel delovati normalno.

Postopek za pripravo zbirk

Za učenje prevajalskega sistema potrebujemo vzporedne podatke (angleško in slovensko oz. dva različna jezika), ki so poravnani na ravni stavka. Za začetek priprave najprej izdelamo imenik, v katerega damo obe datoteki zbirke podatkov. Angleški dodamo pripono *.en, slovenski pa *.sl.

```
mkdir corpus
cd corpus
```

Če želimo pripraviti podatke za učenje prevajalskega sistema, moramo storiti naslednje:

- **Tokenizacija** (*angl. tokenisation*): program vstavi presledke med (npr.) besede in ločila.
- **Določanje pravilnega zapisa velikih in malih črk** (*angl. truecasing*): zapis začetnih besed v vsakem stavku se pretvori v najverjetnejši zapis z malo ali veliko črko. S tem se zmanjša pomanjkanje podatkov.
- **Rezanje** (*angl. cleaning*): dolgi stavki in prazni stavki se odstranijo, saj lahko povzročajo težave z učnim cevovodom. Odstranijo se tudi stavki, za katere je očitno, da niso ustrezno poravnani.

Postopek tokenizacije zaženemo na naslednji način:

```
~/mosesdecoder/scripts/tokenizer/tokenizer.perl -l sl \
< ~/corpus/training/ime-slovenske-datoteke.sl    \
> ~/corpus/ime-slovenske-datoteke.tok.sl
~/mosesdecoder/scripts/tokenizer/tokenizer.perl -l en \
< ~/corpus/training/ime-angleške-datoteke.en    \
> ~/corpus/ime-angleške-datoteke.tok.en
```

Orodje za določanje pravilnega zapisa velikih in malih črk najprej zahteva učenje, da lahko izvleče določene statistične podatke o besedilu:

```
~/mosesdecoder/scripts/recaser/train-truecaser.perl \
--model ~/corpus/truecase-model.en --corpus     \
~/corpus/ime-slovenske-datoteke.tok.sl
~/mosesdecoder/scripts/recaser/train-truecaser.perl \
--model ~/corpus/truecase-model.en --corpus     \
~/corpus/ime-angleške-datoteke.tok.en
```

Pri določanju pravilnega zapisa velikih in malih črk se uporablja drug skript v distribuciji Moses:

```
~/mosesdecoder/scripts/recaser/truecase.perl \
--model ~/corpus/truecase-model.sl         \
< ~/corpus/ime-slovenske-datoteke.tok.sl \
> ~/corpus/ime-slovenske-datoteke.true.en
~/mosesdecoder/scripts/recaser/truecase.perl \
--model ~/corpus/truecase-model.en         \
< ~/corpus/ime-angleške-datoteke.tok.en \
> ~/corpus/ime-angleške-datoteke.true.en
```

Na koncu izvedemo rezanje in omejimo dolžino stavka na 80:

```
~/mosesdecoder/scripts/training/clean-corpus-n.perl \
~/corpus/skupno-ime-zbirke.true en sl \
~/corpus/skupno-ime-zbirke.clean 1 80
```

Zadnji ukaz obdela datoteki obeh jezikov. <`skupno-ime-zbirke`> je del imena datoteke pred pripono, ki je običajno enak za obe datoteki. V zgornjih ukazih smo pisali ime-slovenske-datoteke ter ime-angleške-datoteke zaradi večje preglednosti.

Učenje jezikovnega modela

Z jezikovnim modelom zagotovimo tekoč prevod. Zgradimo ga z zbirko v ciljnem jeziku (slovenskem). Spodaj je naveden osnovni ukaz, s katerim zgradimo 3-gramski jezikovni model. Več možnosti je opisanih na spletni strani programa KenLM [**19**]. Jezikovni model ima po koncu obdelave pripono .arpa

```
mkdir ~/lm
cd ~/lm
~/mosesdecoder/bin/lmplz  -o  3  <~/corpus/ime-slovenske-datoteke.true.sl  >
ime-slovenske-datoteke.arpa.sl
```

Nato datoteko `*.arpa.en` z orodjem KenLM zaradi hitrejšega nalaganja pretvorimo v dvojiški zapis:

```
~/mosesdecoder/bin/build_binary \
ime-slovenske-datoteke.arpa.en \
Ime-slovenske-datoteke.blm.en
```

Uporabili bi lahko tudi program IRSTLM, ki ima prav tako dvojiški zapis, ki ga podpira Moses.

Učenje prevajalnega sistema

Glavni del priprave prevajalnega modela je učenje prevajalnega modela. To storimo tako, da z orodjem GIZA++ zaženemo poravnavo besed, izvlečemo in ocenimo besedne zveze, ustvarimo tabele leksikaliziranega preurejanja in ustvarimo Mosesovo konfiguracijsko datoteko. Vse to v Mosesu naredimo z enim samim ukazom. V našem primeru smo glede na navodila na Mosesovi spletni strani ustvarili imenik `/working` in zagnali ukaz za učenje, ki med postopkom učenja dogodke beleži v dnevnike:

```
mkdir ~/working
cd ~/working
nohup nice ~/mosesdecoder/scripts/training/train-model.perl -root-dir train
\
-corpus ~/corpus/skupno-ime-zbirke.clean                                \
-f en -e sl -alignment grow-diag-final-and -reordering msd-bidirectional-fe
\
-lm 0:3:$HOME/lm/ime-jezikovnega-modela.blm.en:8                          \
 -external-bin-dir ~/mosesdecoder/tools >& training.out &
```

Ker smo imeli na voljo zmogljiv 15 procesorski računalnik, smo uporabili še argument `-cores 15`, da smo proces razdelili na več vzporednih niti.

Pozor! Ta proces je časovno kar zahteven in lahko traja več ur, če nimate ustreznega računalnika. Bodite pripravljeni in potrpežljivi, da ne boste procesa prekinili, ker bi mislili, da se je proces obesil.

Ko je obdelava končana, boste v imeniku `~/working/train/model` našli datoteko `moses.ini`. Z modelom, ki ga podaja ta datoteka ini, lahko že začnemo prevajati, vendar je z njim nekaj težav. Prva težava je v tem, da se model nalaga zelo počasi, kar pa lahko izboljšamo tako, da tabelo besednih zvez in tabelo preurejanja pretvorimo v dvojiško obliko, ki jo sistem lahko hitro naloži. Druga težava je v tem, da uteži, s katerimi Moses določi težo različnih primerjanih modelov, niso optimizirane. Če pogledamo datoteko `moses.ini`, vidimo, da so nastavljene na različne privzete vrednosti, na primer 0.2, 0.3 itd. Če želite najti boljše uteži, moramo prevajalni sistem uglasiti.

Uglaševanje

Pozor! To je najpočasnejši del procesa, zato predlagamo, da si v primeru, da se s strežnikom povezujete prek ukaznega poziva, kot je denimo putty, namestite program screen. Ta vam omogoča, da proces zaženete, nato pa povezavo s sejo prekinete, seja pa bo še vedno tekla v ozadju. Ko vas bo zanimalo, kaj se s procesom dogaja, se s sejo lahko znova povežete. Navodila za program screen so na voljo na spletni strani

https://www.gnu.org/software/screen/manual/screen.html.

Za uglaševanje potrebujemo majhno količino vzporednih podatkov, ki so drugačni od učnih podatkov. V našem primeru smo v vseh primerih uporabili manjši del dvojezične zbirke ebooks. Zbirko za uglaševanje moramo zopet najprej tokenizirati in določiti pravilen zapis velikih in malih črk po postopku, ki je opisan zgoraj v razdelku o pripravi zbirke.

Uglaševanje zaženemo tako, da se pomaknemo v imenik, ki smo ga uporabili za učenje, in zaženemo naslednji ukaz:

```
cd ~/working
nohup nice ~/mosesdecoder/scripts/training/mert-moses.pl \
~/corpus/zbirka-za-uglaševanje.true.en                    ~/corpus/zbirka-za-
uglaševanje.true.sl \
~/mosesdecoder/bin/moses train/model/moses.ini --mertdir ~/mosesdecoder/bin/
\
&> mert.out &
```

Ker smo imeli na voljo več jeder, smo izvajanje procesa pohitrili tako, da smo Moses zagnali v večjedrnem načinu. Na konec zgornjega ukaza smo dodali parameter `--decoder-flags="-threads 15"`, da se je dekoder zagnal s 15 nitmi. Tudi s to nastavitvijo lahko proces traja več ur, celo dni.

Končni rezultat naravnavanja je datoteka ini z naučenimi utežmi, ki se nahaja v imeniku `~/working/mert- work/moses.ini`.

Preizkušanje prevajalnega sistema

Po končanem uglaševanju lahko Moses moses zaženemo s spodnjim ukazom ter želeno besedilo vnesemo kar v ukazni poziv:

```
~/mosesdecoder/bin/moses -f ~/working/mert-work/moses.ini
```

Dekoder za zagon potrebuje vsaj nekaj minut. Če želimo, da se zažene hitreje, lahko tabelo

besednih zvez in tabelo leksikaliziranega preurejanja pretvorimo v dvojiški zapis. To storimo tako, da ustvarimo primeren imenik in modele na naslednji način pretvorimo v dvojiški zapis:

```
mkdir ~/working/binarised-model
cd ~/working
~/mosesdecoder/bin/processPhraseTableMin \
-in train/model/phrase-table.gz -nscores 4 \
-out binarised-model/phrase-table
~/mosesdecoder/bin/processLexicalTableMin \
-in train/model/reordering-table.wbe-msd-bidirectional-fe.gz \
-out binarised-model/reordering-table
```

Nato ustvarimo kopijo datoteke `~/working/mert-work/moses.ini` v imeniku binarised-model in spremenimo tabelo besednih zvez in tabelo preurejanja, tako da kaže na različico v dvojiškem zapisu:

1. `PhraseDictionaryMemory` spremenimo v `PhraseDictionaryCompact`
2. Nastavimo pot funkcije `PhraseDictionary`, tako da kaže na

   ```
   $HOME/working/binarised-model/phrase-table.minphr
   ```

3. Nastavimo pot funkcije `LexicalReordering`, tako da kaže na

   ```
   $HOME/working/binarised-model/reordering-table
   ```

Prevajanje datotek in ocenjevanje prevoda s samodejnim algoritmom BLEU

Ko je prevajalni sistem zgrajen do te mere, nas zanima, kako dober je prevajalni sistem. To izmerimo tako, da uporabimo drugi dve datoteki (testna dokumenta – izvirnik in njegov prevod), ki se razlikujeta od tistih, ki smo jih uporabili do zdaj.

Angleški in slovenski dokument (oba morata biti med seboj poravnana na ravni stavka) moramo najprej tokenizirati in določiti pravilen zapis velikih in malih črk tako kot zgoraj.

```
~/mosesdecoder/scripts/tokenizer/tokenizer.perl -l sl \
< ~/corpus/training/ime-slovenske-datoteke.sl    \
> ~/corpus/ime-slovenske-datoteke.tok.sl
~/mosesdecoder/scripts/tokenizer/tokenizer.perl -l en \
< ~/corpus/training/ime-angleške-datoteke.en    \
>~/corpus/ime-angleške-datoteke.tok.en
~/mosesdecoder/scripts/recaser/truecase.perl \
--model ~/corpus/truecase-model.sl         \
< ~/corpus/ime-slovenske-datoteke.tok.sl \
> ~/corpus/ime-slovenske-datoteke.true.en
~/mosesdecoder/scripts/recaser/truecase.perl \
--model ~/corpus/truecase-model.en         \
```

```
< ~/corpus/ime-angleške-datoteke.tok.en \
> ~/corpus/ime-angleške-datoteke.true.en
```

Pred začetkom preizkusnega prevajanja lahko naučeni model za ta preizkusni nabor prefiltriramo, kar pomeni, da bomo obdržali samo vnose, potrebne za prevod preizkusnega nabora. S tem bo prevajanje veliko hitrejše.

```
cd ~/working
~/mosesdecoder/scripts/training/filter-model-given-input.pl             \
Filtrirani-testni-model mert-work/moses.ini ~/corpus/ime-angleškega-
dokumenta.true.en \
-Binarizer ~/mosesdecoder/bin/processPhraseTableMin
```

Dekoder lahko preizkusimo tako, da najprej prevedemo preizkusni dokument, nato pa za prevod zaženemo skript BLEU:

```
nohup nice ~/mosesdecoder/bin/moses            \
-f ~/working/Filtrirani-testni-model/moses.ini   \
< ~/corpus/ime-angleškega-dokumenta.true.en               \
> ~/working/ime-prevedenega-dokumenta.translated.sl         \
> ~/working/newstest2011.out
~/mosesdecoder/scripts/generic/multi-bleu.perl \
-lc ~/corpus/ime-slovenskega-dokumenta.true.sl              \
< ~/working/ime-prevedenega-dokumenta.translated.sl
```

Ocena BLEU je odvisna od števila ujemajočih se n-gramov med ročno predprevedenim testnim prevodom in prevodom, prevedenim s sistemom Moses.

Priloga 3: Optimizacija MERT in zgled inicializacijskih datotek

Na sliki 10.1 je prikazan parametrski del datoteke *moses.ini*, iz katere lahko razberemo privzete vrednosti parametrov.

```
# dense weights for feature functions
[weight]
# The default weights are NOT optimized for translation quality. You MUST tune
the weights.
#         Documentation         for         tuning         is         here:
http://www.statmt.org/moses/?n=FactoredTraining.Tuning
UnknownWordPenalty0= 1
WordPenalty0= -1
PhrasePenalty0= 0.2
TranslationModel0= 0.2 0.2 0.2 0.2
LexicalReordering0= 0.3 0.3 0.3 0.3 0.3 0.3
Distortion0= 0.3
LM0= 0.5
```

Slika 10.1: Parametrski del privzete datoteke moses.ini

Algoritem MERT nato optimizira te parametre tako, da večkrat prevede manjšo učno zbirko in nato v datoteko *moses.ini* (slika 10.2) zapiše vrednosti parametrov, pri katerih je bila dosežena najboljša vrednost BLEU. To je časovno najzahtevnejši postopek in lahko za večjo zbirko na slabšem računalniku traja tudi več dni, celo tednov. V našem primeru je bil proces razdeljen na 14 niti, ki je na 15 procesorskem računalniku tekel, odvisno od velikosti zbirke, od 10 do 30 ur.

```
[weight]

LexicalReordering0= 0.0195308 0.0582737 0.332593 0.0121416 0.0357293 0.211438
Distortion0= 0.0213123
LM0= 0.0414591
WordPenalty0= -0.0963664
PhrasePenalty0= 0.0227121
TranslationModel0= 0.0572243 0.00790737 0.0593157 0.0239968
UnknownWordPenalty0= 1
```

Slika 10.2: Parametrski del optmizirane datoteke moses.ini

Priloga 4: Skripta *ttpos-to-moses.sh*

```
#!/bin/bash
# Skripto izdelal Jože Kadivec
# Uporaba: ./ttpos-to-moses.sh <ime označene zbirke>

if [ -z "$1" ]
  then
    echo "No argument supplied"
    exit
fi

echo "Kopiram datoteko v začasni imenik"
mkdir "temp"
cp $1 temp/temp.txt
echo "Menjam EOL znake s presledki"
perl -p -e 's/\n/ /g' -i temp/temp.txt
echo "Zamenjujem vse <novavrst> z znaki za novo vrstico"
sed -i 's/<novavrst> /\n/g' temp/temp.txt
echo "Zamenjujem tabulatorje z znaki |"
perl -p -e 's/\t/|/g' -i temp/temp.txt
cp temp/temp.txt $1.moses.out
echo "Brišem začasni imenik"
rm -rf temp
```

Priloga 5: Skripta za pretvarjanje Obeliks oblike .xml v Moses

```
binmode STDIN, ":utf8";
binmode STDOUT, ":utf8";

# Skripto izdelali Simon Krek, Sašo Kuntarič in Jože Kadivec
# Uporaba: perl xml-to-moses.pl <XML-datoteka> <txt-Moses-datoteka>

$in1 = @ARGV [0];
$out1 = @ARGV [1];

open (IN, "$in1") || die "Cannot open file [$in1]:$!";

open (OUT, ">$out1") || die "Cannot open file [$out1]:$!";

    while ($line = <IN>) {
               if ($line =~ m/msd="(.)(.*?)" lemma="(.*?)">(.*?)</) {
               $msd1 = $1;
        $msd2 = $2;
        $lemma = $3;
        $form = $4;
        print OUT "$form|$lemma|$msd1|$msd1$msd2";
    }
               elsif ($line =~ m/<S\/>/) {
        print OUT " ";
    }
               elsif ($line =~ m/<c>(.?)<\/c>/) {
               $punct = $1;
        print OUT " $punct|locilo|locilo";
    }
               elsif ($line =~ m/<p>/) {
        print OUT "\n";
    }
        }
close IN;
close OUT;
```