Izgradnja sustava za neuralno strojno prevođenje

Saratlija, Jelena

Master's thesis / Diplomski rad

2020

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, University of Zagreb, Faculty of Humanities and Social Sciences / Sveučilište u Zagrebu, Filozofski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:131:440947

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-28**

Filozofski fakultet
Sveučilišta u Zagrebu

Repository / Repozitorij:

ODRAZ - open repository of the University of Zagreb
Faculty of Humanities and Social Sciences

zir.nsk.hr

dabar
DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

SVEUČILIŠTE U ZAGREBU

FILOZOFSKI FAKULTET

ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI

SMJER Informatika (istraživački smjer)

Ak. god. 2019./ 2020.

Jelena Saratlija

Izgradnja sustava za neuralno strojno prevođenje

Diplomski rad

Mentori: dr. sc. Ivan Dunđer, prof. dr. sc. Sanja Seljan,

Zagreb, rujan 2020.

Izjava o akademskoj čestitosti

Izjavljujem i svojim potpisom potvrđujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

_____

(potpis)

Sadržaj

1. Uvod 1
2. Razvoj strojnog prevođenja 5
   2.1. Povijest strojnog prevođenja 5
   2.2. Pristupi strojnom prevođenju 10
   2.3. Neuralne mreže 11
      2.3.1. Procesi učenja neuralnih mreža 13
   2.4. Neuralno strojno prevođenje 16
      2.4.1. Treniranje dubokih neuralnih mreža 16
      2.4.2. Neuralni jezični modeli 17
      2.4.3. Dinamičke povratne (rekurentne) neuralne mreže 19
         2.4.3.1. Duga kratkoročna memorija (LSTM) 21
         2.4.3.2. Mreža propusno povratnih ćelija (GRU) 23
      2.4.4. Enkoder-dekoder arhitektura 24
      2.4.5. Generiranje prijevoda 25
         2.4.5.1. Greedy dekodiranje 25
         2.4.5.2. Beam dekodiranje 26
      2.4.6. Mehanizam pažnje 27
   2.5. Usporedba statističkog i neuralnog strojnog prevođenja 30
3. Istraživanje 35
   3.1. Cilj istraživanja 35
   3.2. Metodologija i tijek istraživanja 35
   3.3. OpenNMT 36
   3.4. Podatkovni skupovi za treniranje i testiranje sustava 38

3.5. Izgradnja sustava za neuralno strojno prevođenje 40
3.6. Automatska evaluacija kvalitete strojnog prijevoda 47
3.6.1. BLEU metoda evaluacije 49
3.6.2. Analiza BLEU rezultata 51
4. Zaključak 61
5. Literatura 62
Popis slika 69
Popis tablica 71
Sažetak 72
Summary 73

1. Uvod

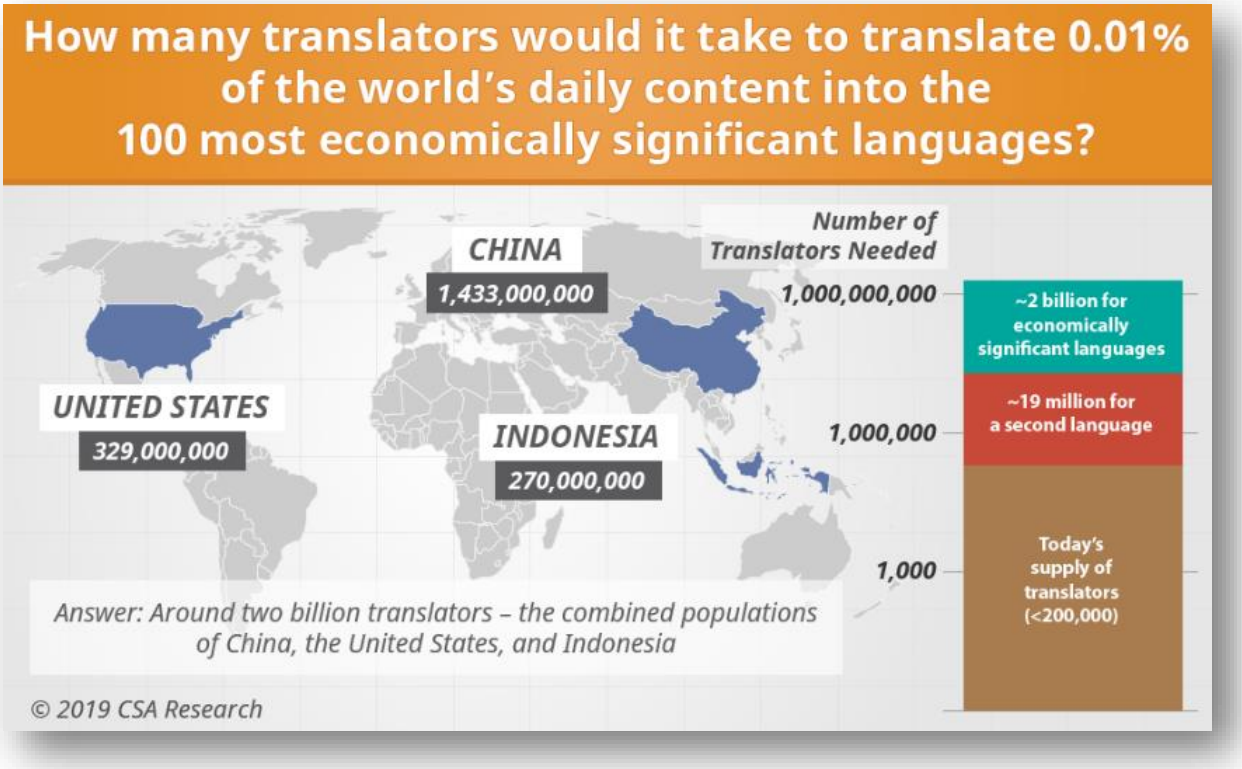
Jezici čine temelj komunikacije i uistinu bi bilo izvanredno imati jedinstven jezik na kojem bi svi mogli komunicirati i međusobno se razumjeti. To bi srušilo mnoge barijere i olakšalo protok informacija kojih je svakog dana sve više. Prema Popadić (2017[1]) automatizacija strojnog prevođenja nedvojbeno je važna tema, kako s društvenog, političkog i komercijalnog, tako i sa znanstvenog i filozofskog gledišta.

Društvena i politička važnost očituje se ponajviše u zemljama ili udrugama u kojima se govori više od jednog jezika. Svaki narod ponosan je na svoj jezik i ne želi dopustiti dominaciju drugog jezika nad vlastitim. Osim toga, uz jezik se vežu i kultura, običaji i tradicije. Kako bi svaki narod očuvao integritet svoga jezika, a uspijevao komunicirati sa svim ostalim narodima svijeta, važno je osigurati izgradnju jezičnih digitalnih resursa kojima se potiče razvoj i očuvanje jezika. Globalizacija i razvoj tehnologije utjecali su na razvoj interneta, alata i resursa pomoću kojih je moguće ostvariti kontakt s osobom iz bilo kojeg dijela svijeta. Da bi se razumjeli, ljudi moraju ili pronaći zajednički jezik ili iskoristiti prednosti strojnog prevođenja. Iako takvi sustavi nisu nepogrešivi i savršeni, napredak ostvaren u zadnjih 30 godina od velikog je značaja (Seljan, 2002[2]). U posljednih 5 godina, u središtu pozornosti je neuralni pristup nad obradom teksta općenito, kao i nad strojnim prevođenjem, što predstavlja okosnicu ovoga rada.

Prevođenje je od znatne važnosti jer dostupnost uputa i informacija o proizvodu na vlastitom jeziku uvelike utječu na korisnika informacije (Seljan, 2011[3], Dunđer, 2020[4], Seljan i sur., 2015[5]) i na poslovanje u poduzećima (Seljan, 2018a[6], Seljan 2018b[7]). Za primjer se može uzeti bilo kakav uređaj, poput televizora, proizvedenog u Kini. Ako kupac ne poznaje taj jezik, a nije osiguran prijevod uputa pri radu s uređajem, gotovo je sigurno da kupac neće kupiti ni koristiti taj proizvod. Kada se samo pomisli na količinu proizvoda koja se svaki dan proizvede,

[1] Popadić, D. (2007). Usporedna analiza alata za strojno prevođenje. Diplomski rad: Sveučilište Josipa Jurja Strossmayera u Osijeku, Fakultet elektrotehnike, računarstva i informacijskih tehnologija.
[2] Dovedan, Z.; Seljan, S.; Vučković, K. (2002). Strojno prevođenje kao pomoć u procesu komunikacije. Informatologija 35 (4), 283-291.
[3] Seljan, S. (2011). Translation technology in education and business. Informatologia 44 (4), 279-286.
[4] Dunđer, I. (2020). Machine Translation System for the Industry Domain and Croatian Language. Journal of information and organizational sciences. 44. 33-50.
[5] Seljan, S.; Klasnić, K., Stojanac, M.; Pešorda, B., Mikelić Preradović, N. (2015). Information Transfer through Online Summarizating and Translation Technology.
[6] Seljan, S. (2018a). Total Quality Management Practice in Croatian Language Service Provider Companies. EntreNova 18, 4 (1), 461-469. INFuture2015: e-Institutions–Openness, Accessibility, and Preservation.
[7] Seljan, S. (2018b). Quality Assurance (QA) of Terminology in a Translation Quality Management System (QMS) in the Business Environment. European Parliament: Translation Services in the Digital World, 92-145.

na tržišta i zemlje u koje će biti izvezeni, na potencijalne kupce i jezike, očigledno je da potreba za prevođenjem nadilazi ljudske resurse koji bi to mogli brzo i kvalitetno odraditi. CSA (Common Sense Advisory) Research, tvrtka koja se bavi istraživanjima fokusiranima na globalna tržišta i tržišta jezičnih usluga 2019. godine je provela istraživanje kako bi otkrila koliko je prevoditelja potrebno da bi se prevelo 0.01% globalnog dnevnog sadržaja u 100 ekonomski najvažnijih svjetskih jezika (CSA Research, 2019[8]).



Slika 1. Važnost prevođenja (CSA, Research, 2019)

Prema rezultatima istraživanja (CSA Research, 2019[9]) prikazanima na slici 1 potrebno je otprilike 2 milijarde prevoditelja, tj. sveukupna populacija SAD-a, Indonezije i Kine, kako bi se prevelo 0,01% dnevnog svjetskog sadržaja u 100 ekonomski najznačajnijih jezika. Štoviše, čak 19 milijuna prevoditelja bilo bi potrebno da bi se sadržaj preveo samo u drugi ekonomski najvažniji svjetski jezik, a ono što najviše zapanjuje u usporedbi s ovim rezultatima je činjenica da u svijetu postoji manje od 200.000 profesionalnih prevoditelja. Idealno bi bilo kada bi postojali sustavi koji bi mogli prevoditi podjednako dobro kao i profesionalni

[8] CSA Research. (2019). Why buy CAT tools when NMT rules?
[9] Ibid.

prevoditelj. U ovom trenutku to još nije postignuto. Pomisli li se samo na bogatstvo jezika, jezičnih pravila, načina izražavanja, raznih riječi, njihovih sinonima, antonima, lažnih parova među jezicima, dijalekte, pa i činjenicu da riječi konstantno ulaze i izlaze iz jezika, upitno je može li sustav za strojno prevođenje pratiti sve trendove i osigurati izražajno prihvatljive prijevode kakve bi ostvario profesionalni prevoditelj. Ono što je sigurno jest da su današnji sustavi itekako korisni. Iako proizvode pogreške, ljudima je često dovoljno i da dobiju samo općenitu ideju o tekstu koji ih zanima.

Iz znanstvene perspektive, strojno prevođenje je vrlo zanimljivo područje primjene umjetne inteligencije, a neki od najvažnijih razvitaka na tom području započeli su upravo strojnim prevođenjem. Umjetna opća inteligencija predstavlja mogućnost stroja da razumije ili nauči bilo koji zadatak za čije je obavljanje potrebna ljudska inteligencija. Temelji se i na dubokom učenju čije tehnike unaprjeđuju mogućnost računala da klasificira, prepoznaje, detektira, opisuje i razumije podatke.

S filozofskog gledišta strojno prevođenje dovodi u pitanje mogućnost potpune automatizacije prevođenja, odnosno traži odgovor na pitanje koliko je moguće automatizirati razmišljanje. Aktivnost prevođenja zahtijeva široki spektar ljudskog znanja, od informatičkog, informacijskog, matematičkog, lingvističkog, logičkog i kulturnog. Jezik je živ, stalno se mijenja, riječi ulaze i izlaze iz uporabe, a izrazi s vremenom mogu izgubiti ili promijeniti svoje značenje. S obzirom na to, upitno je hoće li potpuna automatizacija prevođenja ikada biti moguća.

Rad je podijeljen na dvije osnovne cjeline: teorijski i praktični dio. U teorijskom dijelu je opisana i objašnjena svrha prevođenja, povijest i različiti pristupi. Posebno su analizirane neuralne mreže, različiti modeli i arhitektura neuralnog strojnog prevođenja. Analizirana je i usporedba između statističkog i neuralnog strojnog prevođenja. U prvom poglavlju opisana je svrha i važnost prevođenja. Drugo poglavlje podijeljeno je na 5 potpoglavlja. U prvom potpoglavlju opisan je nastanak kao i povijesni razvoj strojnoga prevođenja od samoga početka do danas. U drugom potpoglavlju opisani su postojeći pristupi strojnom prevođenju, dok je u trećem potpoglavlju opisan koncept neuralnih mreža koje čine temelj za ono što će biti opisano u četvrtom potpoglavlju, neuralno strojno prevođenje. U tom poglavlju opisan je način na koji se duboke neuralne mreže treniraju te kako procesuiraju podatke kako bi generirale prijevode. U petom potpoglavlju iznesena je usporedba neuralnog i statističkog pristupa strojnom prevođenju.

U praktičnom dijelu rada prikazano je istraživanje. Opisan je postupak pripreme sustava za neuralno strojno prevođenje primjenom OpenNMT okruženja. Sustav je treniran nad pripremljenim skupom podataka, a zatim je testiran prevođenjem podatkovnog skupa od 1001 rečenice. Rezultati su evaluirani opisanom BLEU metodom evaluacije. Na kraju je iznesen zaključak, iza kojeg slijedi pregled korištene literature, popis slika, tablica te sažeci s ključnim riječima na engleskom i hrvatskom jeziku.

2. Razvoj strojnog prevođenja

2.1. Povijest strojnog prevođenja

Automatizacija prevođenja jedan je od ciljeva koji bi uvelike olakšao razmjenu informacija diljem svijeta. Prema Hutchins i Somers (1992[10]) korištenje mehaničkih rječnika u svrhu prevladavanja jezičnih barijera prvi put je predloženo još u 17. stoljeću kada Descartes i Leibniz raspravljaju o kreiranju rječnika temeljenih na univerzalnim numeričkim kodovima. Inspiraciju su pronašli u tada aktualnom pokretu „univerzalnog jezika" koji označava ideju o kreiranju nedvosmislenog jezika temeljenog na logičkim principima i simbolima s kojima bi cijelo čovječanstvo moglo komunicirati bez straha od nesporazuma. Najpoznatiji takav međujezik (eng. *interlingua*) objavljen je u radu *Essay towards a Real Charachter and a Philosophical Language* autora Johna Wilkinsa iz 1668. (Hutchins i Somers, 1992[11]). U dvadesetom stoljeću dolazi do značajnijeg napretka u istraživanjima strojnog prevođenja. Automatizacija prevođenja postaje stvarnost u obliku kompjutorskih programa koji mogu prevoditi tekstve iz jednog prirodnog jezika u drugi. Ti prijevodi nisu bili savršeni, a postići savršeni strojni prijevod bez ljudskog faktora je ono čemu se teži. Do kraja 20. stoljeća razvijeni su programi koji mogu proizvesti „sirove" prijevode u relativno dobro definiranim domenama (Seljan, 2000[12]). Ti prijevodi mogu se revidirati kako bi se ostvario prijevod dobre kvalitete po ekonomski održivoj stopi ili su to prijevodi koje u svom nerevidiranom stanju mogu pročitati i razumjeti stručnjaci za to područje u svrhu općenitog informiranja.

Od 17. stoljeća do polovice 20. stoljeća bilo je pokušaja predlaganja raznih internacionalnih međujezika među kojima je najpoznatiji Esperanto (Hutchins i Somers, 1992[13]). Pokušaji mehanizacije prevođenja započinju polovicom 20. stoljeća (Dovedan i sur., 2002[14]). 1930-ih godina pojavljuju se prvi patenti u Francuskoj i Rusiji. 1933. godine Artsrouni prijavljuje patent za „stroj koji prevodi", odnosno za automatski dvojezični rječnik zapisan na bušenim trakama dok ruski istraživač Troyanskii iste godine prijavljuje opsežniji patent za dvojezični rječnik koji se oslanja i na metodu prepoznavanja gramatičkih uloga u raznim

[10] Hutchins, J.; Somers, H. L. (1992). An introduction to machine translation. London: Academic Press.
[11] Ibid.
[12] Seljan, S. (2000). Sublanguage in Machine Translation. Mipro 2000.
[13] Hutchins, J.; Somers, H. L. (1992). op. cit.
[14] Dovedan, Z.; Seljan, S.; Vučković, K. op.cit.

jezicima (Dunđer, 2015[15], Hutchins, 2004[16].). Prema Dunđer (2015[17]) jedan od događaja koji je omogućio nagli razvoj i istraživanja na području strojnog prevođenja jest pojava prvog programabilnog računala ENIAC predstavljenog 1946. godine. Nekoliko godina poslije W. Weaver i A. D. Booth sastaju se i raspravljaju o mogućnosti korištenja računala za prevođenje. No, tek je memorandumom W. Weavera iz 1949. godine ideja o strojnom prevođenju predstavljena javnosti (Hutchins, 1995[18]). U svom radu Weaver predlaže razne metode: korištenje kriptografskih tehnika, statističkih metoda, Shannonove teorije informacija, istraživanje logike i univerzalnih značajki jezika koje predstavljaju zajedničke temelje ljudske komunikacije (Hutchins, 1995[19]).

Yehoshua Bar-Hillel, znanstvenik s MIT-a koji se bavio istraživanjem strojnog prevođenja potpomognutog rječnicima, 1952. godine organizirao je prvu međunarodnu konferenciju o strojnom prevođenju (Dunđer, 2015[20]). Na konferenciji je izneseno kako potpuno automatsko prevođenje ne može biti ostvareno bez dugoročnih istraživanja i bez ljudskog faktora, tj. osobe koja će tekstove pripremiti ili revidirati nakon prijevoda (eng. *pre- and post-editing*). Sudionici konferencije iznijeli su kako je prvi zahtjev za nastavak ulaganja u istraživanje ovoga područja bio dokazati izvedivost sustava za strojno prevođenje. Zatim je u siječnju 1954. godine održana prva demonstracija sustava za strojno prevođenje na Sveučilištu Georgetown. Pažljivo je odabran uzorak od 49 rečenica na ruskom jeziku koje su prevedene na engleski jezik korištenjem ograničenog rječnika od 250 riječi i 6 pravila (Hutchins, 1995[21]). Iako ovaj eksperiment nije bio od velike znanstvene važnosti, bio je dovoljan kako bi potaknuo brojna ulaganja u istraživanja strojnog prevođenja diljem svijeta.

1950-ih i 1960-ih godina istraživanjima dominiraju sustavi za strojno prevođenje temeljeni na dvojezičnim rječnicima i pravilima za održavanje ispravnog poretka riječi u rečenici. Tadašnji sustavi rabili su 3 različite metode i pristupe prevođenju (Dunđer, 2015[22],

[15] Dunđer, I. (2015). Sustav za statističko strojno prevođenje i računalna adaptacija domene. Doktorski rad. Zagreb: Filozofski fakultet.
[16] Hutchins, J. (2004). Two precursors of machine translation: Artsrouni and Trojanskij. International Journal of Translation, vol. 16, no. 1, pp. 11-31.
[17] Dunđer, I. op. cit.
[18] Hutchins, J. (1995). Concise history of the language sciences: from the Sumerians to the cognitivists. U E.F.K. Koerner i R.E. Asher, (431-445), Pergamon Press, Oxford.
[19] Ibid.
[20] Dunđer, I. op. cit.
[21] Hutchins, J. (1995). op. cit.
[22] Dunđer, I. op. cit.

Chéragui, 2012[23], Koehn, 2010[24]). Prva od njih je direktna metoda koja pomoću pravila uparuje riječi, zatim metoda transfera koja koristi morfološku i sintaktičku analizu te metoda međujezika koja koristi apstraktnu reprezentaciju značenja.

1964. godine vladini sponzori strojnog prevođenja Sjedinjenih Američkih Država formiraju ALPAC (eng. *Automatic Language Processing Advisory Committee*) kako bi se ispitalo stanje istraživanja i njihovi rezultati. Prema Hutchins (1995[25]) u izvješću iz 1966. godine izneseno je da je strojno prevođenje sporije, manje točno i dvostruko skuplje nego ljudski prijevod i da se ne može predvidjeti kako će se nastavak financiranja istraživanja strojnog prevođenja isplatiti. Umjesto toga preporučen je nastavak razvoja alata za pomoć prevoditeljima kao što su automatski rječnici i kontinuirani nastavak generičkog istraživanja računarske lingvistike (Hutchins, 1995[26]). Ovo izvješće za rezultat je imalo prestanak razvoja strojnog prevođenja diljem svijeta, međutim, istraživanja su nastavljena u Kanadi, Francuskoj i Njemačkoj. Unatoč napuštanju istraživanja strojnog prevođenja, prema Dunđer (2015[27]) u SAD-u s radom nastavljaju Peter Toma, osnivač tvrtke „Systran" (1968.) čiji je sustav koristilo američko ministarstvo obrane, i Bernard Scott, osnivač tvrtke „Logos" (1970.). Zatim se pojavljuje kanadski sustav „TAUM Météo" koji prevodi vremenske prognoze s engleskog na francuski (Dunđer, 2015[28], Hutchins, 2001[29]). Krajem istog desetljeća, pojavio se METAL (eng. *Mechanical Translation and Analysis of Languages*) koji se bazirao na prevođenju između engleskoga i njemačkoga jezika (Popadić, 2017[30]).

80-ih godina 20. stoljeća metoda međujezika dolazi u središte istraživanja. Tom metodom pokušava se opisati i prikazati značenje neovisno o odabranom jeziku. Među najpoznatijim sustavima temeljenima na ovoj metodi bili su „Catalyst" i „Pangloss" (Dunđer 2015[31], Koehn 2010[32]). Prema Čanić (2018[33]) u ovom razdoblju pojavljuju se nove metode koje naglasak stavljaju na samostalno učenje, odnosno treniranje sustava koji bi učio pomoću

[23] Chéragui, M. A. (2012). Theoretical Overview of Machine translation. Proceedings of the 4th International Conference on Web and Information Technologies (ICWIT 2012), pp. 160-169.
[24] Koehn, P. (2010), Statistical Machine Translation. Cambridge: University Press.
[25] Hutchins, J. (1995). op. cit.
[26] Ibid.
[27] Dunđer, I. op. cit.
[28] Ibid.
[29] Hutchins, J. (2001). Machine translation over fifty years. Histoire, Epistémologie, Langage: Le traitement automatique des langues, vol. 23, no. 1, pp. 7-31.
[30] Popadić, D. (2007). op. cit.
[31] Dunđer, I. op. cit.
[32] Koehn, P. op. cit.
[33] Čanić, J. (2018). Komparativna analiza strojnog prijevoda sa švedskog na hrvatski jezik. Diplomski rad. Zagreb: Filozofski fakultet.

prevedenih tekstova što rezultira pojavom metoda strojnog prevođenja temeljenog na primjerima i metoda temeljenih na podacima (Brkić i sur., 2009[34]). Na temelju ovih koncepata nastale su prijevodne memorije koje sadrže već prevedeni tekst s jednog jezika na drugi te ga je ugrađivanjem u alate za strojno potpomognuto prevođenje (eng. *CAT tools*) moguće iskoristiti pri prevođenju za isti jezični par, koje se mogu izgrađivati tijekom postupka prevođenja ili preuzeti već gotove, kao npr. sa specijaliziranih sustava (Jaworski i sur., 2017[35]). Dakle, pri prevođenju novog teksta sustav prijevodne memorije pretražit će postojeće prijevode te ponuditi adekvatne prijevode, ukoliko se novi tekst do određene razine podudara s tekstom koji se već nalazi u prijevodnoj memoriji (Brkić i sur., 2008[36], Seljan i sur., 2020[37]) i predstavljaju temelj za daljnju izgradnju sustava (Seljan i Pavuna, 2006[38]).

1990-ih godina dolazi do snažnog tehnološkog napretka, cijena računala opada, kućanstva ih kupuju te se industrija okreće većoj proizvodnji osobnih računala. S obzirom na to da u isto vrijeme raste broj informacija dostupnih na internetu, odnosno broj tekstualnih korpusa na raznim jezicima, raste i potreba za korištenjem strojnog prevođenja. Prema Dunđer (2015[39]), u ovom razdoblju pojavljuje se besplatni prevodilački alat „BabelFish“ izrađen prema Systranovoj tehnologiji strojnog prevođenja temeljenog na pravilima. Ovo razdoblje završava važnim događajem za najnoviji, neuralni pristup strojnom prevođenju. Prema Medium (2017[40]), godine 1997. Ramon Neco i Mikel Forcada predlažu korištenje „enkoder-dekoder“ arhitekture za strojno prevođenje, da bi desetljeće kasnije – 2003. godine – grupa istraživača sa Sveučilišta u Montrealu predvođena istraživačem Yoshua Bengio razvila jezični model temeljen na neuralnim mrežama. Ovaj rad postavio je temelje ubrzanom razvoju neuralnih mreža u svrhu strojnog prevođenja.

[34] Brkić, M.; Seljan, S.; Vičić, T. (2009). Evaluation of the statistical machine translation service for Croatian-English. INFuture 2009: Digital resources and knowledge sharing, 319-322.
[35] Jaworski, R.; Seljan, S.; Dunđer, I. (2017). Towards educating and motivating the crowd – a crowdsourcing platform for harvesting the fruits of NLP students' labour. Human Language Technologies as a Challenge for Computer Science and Linguistics 1, 332-336.
[36] Brkić, M.; Seljan, S.; Bašić Mikulić, B. (2009). Using translation memory to speed up translation process. INFuture 2009 : Digital resources and knowledge sharing, 353-363.
[37] Seljan, S.; Škof Erdelja, N.; Kučiš, V.; Dunđer, I.; Pejić Bach, M. (2020). Quality Assurance in Computer-Assisted Translation in Business Environments. Natural Language Processing for Global and Local Business. IGI-Global, 247-270.
[38] Seljan, S.; Pavuna, D. (2006). Translation Memory Database in the Translation Proces.. Proceedings of Information and Intelligent Systems IIS, 327-332.
[39] Dunđer, I. op. cit.
[40] Medium. (2017). History and Frontier of the Neural Machine Translation.

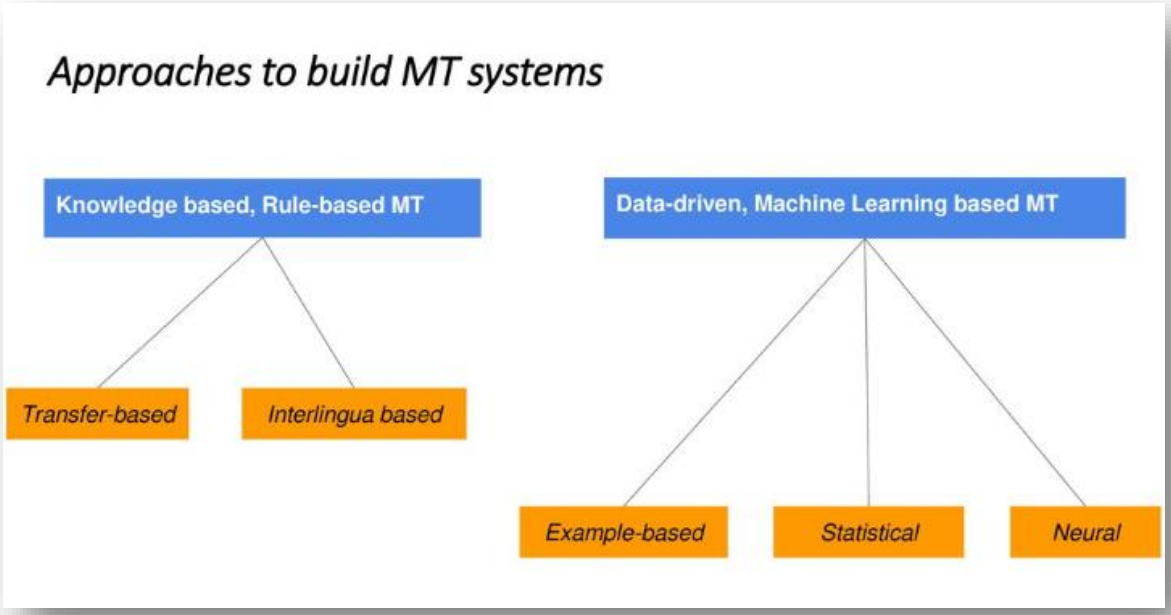
U 2000-im godinama s pojavom sustava za statističko strojno prevođenje naglo raste interes za istraživanje strojnog prevođenja. Prema Koehn i Haddow (2012[41]), ti sustavi se izgrađuju za specifičnu domenu jer na taj način sustavi postižu najbolje prijevode. Dillinger i Marciano (2012[42]) kao prednosti ovog pristupa navode relativno jeftinu izgradnju sustava za statističko strojno prevođenje, jednostavno dodavanje novih jezika i automatsko ugađanje sustava. Iako su ovakvi prijevodi uglavnom fluentni, ovaj pristup ima poteškoća s gramatičkim aspektom jezika poput glagolskih vremena, brojeva, padeža, slaganja i slično (Seljan i sur., 2015[43], Kučiš i Seljan, 2014[44]).

U 2010-im godinama dolazi do ubrzanog razvoja neuralnih mreža koje se počinju primjenjivati za rješavanje kompleksnih problema poput strojnog prevođenja, prepoznavanja govora i vizualnog prepoznavanja objekata (Sutskever i sur., 2014[45]). Prema Medium (2017[46]) 2013. godine N. Kalchbrenner i P. Blunsom predlažu novu „*end-to-end"* enkoder-dekoder arhitekturu za strojno prevođenje. Ovaj model kodira rečenicu iz izvornog jezika u kontinuirani vektor korištenjem konvolucijske neuralne mreže (eng. *convolutional neural network*), a zatim koristi dinamičku povratnu (rekurentnu) neuralnu mrežu (eng. *recurrent neural network*) koja dekodira i transformira vektor u rečenicu u ciljnom jeziku. Kalchbrenner i Blunsom ovaj su koncept predstavili u djelu *Recurrent Continuous Translation Models* čime je određen nastanak neuralnog strojnog prevođenja.

[41] Koehn, P.; Haddow, B. (2012). Interpolated backoff for factored translation models. Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA).
[42] Dillinger, M.; Marciano, J. (2012). Introduction to MT. The Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2012).
[43] Seljan, S.; Tucaković, M.; Dunđer, I. (2015). Human evaluation of online machine translation services for english/russian-croatian. New Contributions in Information Systems and Technologies, 1089-1098.
[44] Kučiš, V.; Seljan, S. (2014). The role of online translation tools in language education. Babel 60 (3).
[45] Sutskever, I.; Vinyals, O.; Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104–3112).
[46] Medium. op. cit.

2.2. Pristupi strojnom prevođenju

Sumirajući kratak pregled povijesti strojnog prevođenja, prema Bhattacharyya (2015[47]) tri su paradigme dominirale strojnim prevođenjem. Vremenski poredane to su strojno prevođenje temeljeno na pravilima (eng. *rule-based machine translation* ili RBMT), strojno prevođenje temeljeno na primjerima (eng. *example-based machine translation* ili EBMT) i statističko strojno prevođenje (eng. *statistical machine translation* ili SMT).



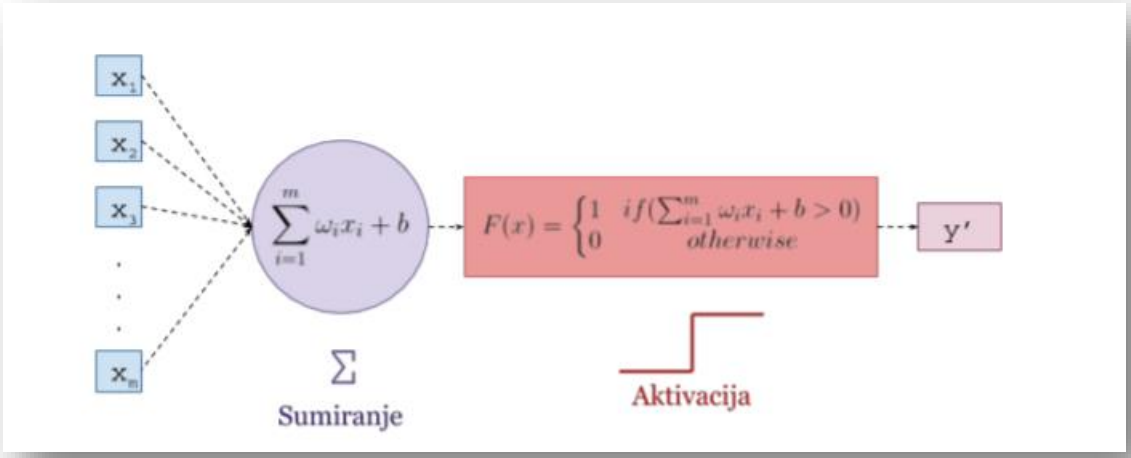
Slika 2. Pristupi izgradnji sustava za strojno prevođenje (Kunchukuttan, 2018)

Prema slici 2 u osnovi se pristupi strojnom prevođenju razlikuju prema načinu na koji odrađuju tri temeljna procesa strojnog prevođenja: analiza, transfer i generiranje (*analysis, transfer and generation*, ATG). RBMT koristi pravila i intermedijarni prikaz, interlinguu ili transfer koji se razlikuju prema dubini provedene analize (Kunchukuttan, 2018[48]). SMT koristi podatke, a EBMT kombinira oba pristupa što znači da podaci osiguravaju dijelove prijevoda koji se zatim ponovno kombiniraju prema pravilima. Trenutno je najaktualniji neuralni pristup strojnom prevođenju (eng. *neural machine translation* ili NMT) koji predstavlja i središnju temu ovog rada.

[47] Bhattacharyya, P. (2015). Machine Translation. CRC PRESS Taylor & Francis Group, xix str (Preface)
[48] Kunchukuttan, A. (2018). An introduction to Machine Translation. Center for Indian Language Technology, Indian Institute of Technology Bombay. Ninth IIT-H Advanced Summer School on NLP, 27th June 2018.

2.3. Neuralne mreže

Duboko učenje je vrsta strojnog učenja koja koristi višeslojne neuralne mreže. Prema Vasić (2018[49]) umjetna neuralna mreža (eng. *Artificial neural network* ili ANN) je nelinearni model neuralne mreže baziran na strukturi ljudskog mozga. Prema Neubig (2017[50]) neuralne mreže mogu biti shvaćene kao lanac funkcija koji uzima ulazne podatke i izračunava izlazne podatke. Moć neuralnih mreža krije se u činjenici da povezivanje više jednostavnih funkcija omogućava reprezentaciju kompleksnijih funkcija. Osnovna jedinica neuralne mreže je neuron, a najjednostavnija vrsta neurona je perceptron (Rosenblatt, 1958[51]). Prema Vasić (2018[52]) perceptron, razvijen 1950-ih godina, u najjednostavnijem obliku iz više ulaza (eng. *input*) stvara jedan jedinstveni izlaz (eng. *output*).



Slika 3. Prikaz jednoslojnog perceptrona (Vasić, 2018)

Prema Vasić (2018[53]) na slici 3 je prikazan jednoslojni perceptron koji ima ulaze $X = x_1, x_2, x_3, \ldots, x_m$ i izlaz *y'* koji ovisi o parametrima neuralne mreže. Parametri ($w_1, w_2, w_3, \ldots, w_m$) reprezentiraju važnost ulaza u odnosu na izlaz, a vrijednost *b* određuje pristranost (eng. *bias*) određenoj klasi.

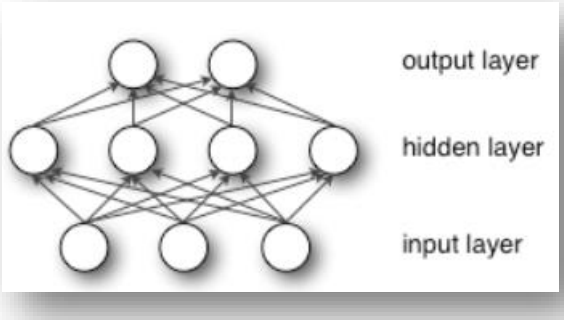
[49] Vasić, D. (2018). Analiza i primjena metoda automatskog semantičkog označavanja teksta. Split: Sveučilište u Splitu. Fakultet elektrotehnike, strojarstva i brodogradnje.
[50] Neubig, G. (2017). Neural Machine Translation and Sequence-to-sequence Models: A Tutorial. Language Technologies Institute, Carengie Mellon University. arXiv:1703.01619v1 [cs.CL].
[51] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in te brain. Psychol. Rev., vol. 65, no. 6, pp. 386–408.
[52] Vasić, D. op. cit.
[53] Ibid.

$$y' = \begin{cases} 0 & if \sum_{i=0}^{m} \omega_i x_i \leq granica \\ 1 & if \sum_{i=0}^{m} \omega_i x_i > granica \end{cases}$$

Slika 4. Određivanje izlaza neurona (Vasić, 2018)

Prema slici 4 izlaz neurona *y'* određuje se ispitivanjem je li suma jednadžbe iznad određene granične vrijednosti. Granična vrijednost u ovoj jednadžbi označava koliko je sustav pristran pojedinoj klasi (Vasić, 2018[54]), a još se naziva i vrijednost pristranosti. Težinske i granične vrijednosti podesivi su parametri pa je stoga moguće prilagoditi ih kako bi se neuralna mreža ponašala na željeni način. Prema Pathmind (2019[55]) neuralnu mrežu čini samo jedan perceptron kao što je prikazano u ovom primjeru, dok duboka neuralna mreža sadrži arbitraran broj perceptrona. Dakle, mreže dubokog učenja razlikuju se od uobičajenih jednoslojnih neuralnih mreža prema svojoj dubini, odnosno prema broju slojeva kroz koje podaci moraju proći u procesu prepoznavanja uzoraka i formiranju izlaznog podatka (Pathmind, 2019[56]).

Prema Božić-Štulić (2017[57]) neuralna se mreža s obzirom na svoju strukturu može podijeliti na tri sloja: ulazni sloj (eng. *input layer*), skriveni sloj (eng. *hidden layer*) i izlazni sloj (eng. *output layer*). Ulazni sloj prima podatke iz okruženja, zatim skriveni sloj, sastavljen od neurona, ekstrahira uzorke povezane s procesom koji se analizira, a izlazni sloj prikazuje izlaz mreže.



Slika 5. Struktura duboke neuralne mreže (Pathmind, 2019)

[54] Vasić, D. op. cit.
[55] Pathmind. (2019). A Beginner's Guide to Neural Networks and Deep Learning.
[56] Ibid.
[57] Božić-Štulić, D. (2017). Semantička segmentacija slika metodama dubokog učenja. Kvalifikacijski ispit. Split: Sveučilište u Splitu, Fakultet elektrotehnike, strojarstva i brodogradnje.

Na slici 5 prikazana je osnovna struktura duboke neuralne mreže. S obzirom na pozicije neurona i broj slojeva, postoji cijeli niz različitih neuralnih mreža. U svrhu objašnjenja neuralnog strojnog prevođenja u ovom radu će biti prikazane samo određene duboke neuralne mreže kao što su:

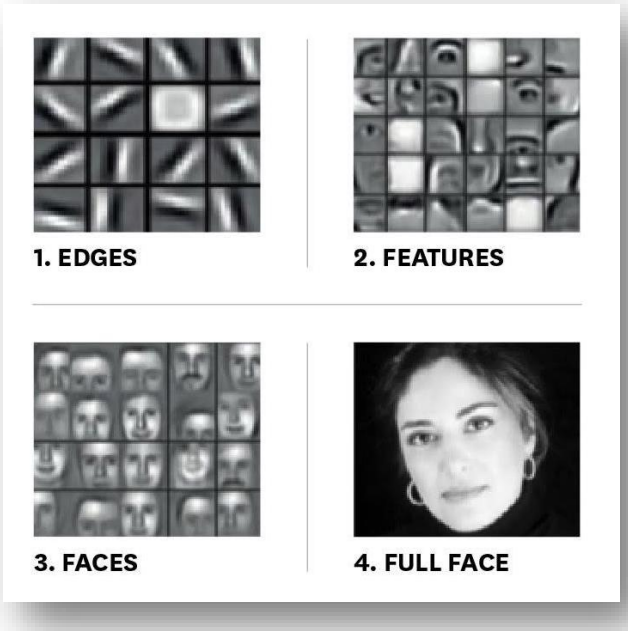
- dinamička povratna (rekurentna) neuralna mreža (eng. *recurrent neural network,* RNN),
- mreža duge kratkoročne memorije (eng. *long short-term memory,* LSTM) i
- mreža propusno povratnih ćelija (eng. *gated recurrent unit,* GRU).

2.3.1. Procesi učenja neuralnih mreža

Važno je naglasiti da je duboko učenje zapravo strojno učenje, a razlika je sadržana u tome kakve podatke model uzima. Prema Towards Data Science (2020[58]) algoritmi strojnog prevođenja trebaju strukturirane podatke dok mreže dubokog učenja mogu obrađivati i nestrukturirane podatke koristeći slojeve umjetnih neuralnih mreža (eng. *artificial neural networks*) kako bi odredile je li pretpostavka točna ili ne. Upravo ta karakteristika duboko učenje čini iznimno korisnim jer omogućuje obradu nestrukturiranih podataka kao što su slike, zvučni zapisi i tekstualni podaci.

Neuralne mreže dizajnirane su za prepoznavanje uzoraka. Opisane su i kao "univerzalni aproksimator" jer mogu naučiti kako približno odrediti nepoznatu funkciju $f(x)=y$ između bilo kakvog ulaznog podatka x i bilo kakvog izlaznog podatka y, pretpostavljajući da su podaci u korelaciji ili da postoji uzročna veza (Pathmind, 2019[59]). U mrežama dubokog učenja, svaki sloj trenira se na različitim skupinama značajki temeljenim na ishodu prethodnog sloja. Što se više napreduje kroz neuralnu mrežu, karakteristike koje slojevi mogu prepoznati su sve kompleksnije jer sakupljaju i kombiniraju karakteristike iz prethodnih slojeva s trenutnima (Pathmind, 2019[60]). Na slici 6 prikazano je kako se povećava razina kompleksnosti koje učenjem sustav može prepozati.

[58] Towards Data Science. (2020). A Beginner's Guide to Deep Learning.
[59] Pathmind. op. cit.
[60] Ibid.



Slika 6. Razine kompleksnosti značajki (Pathmind, 2019)

Neuralne mreže mogu klasificirati podatke kada postoji označeni set/skup podataka nad kojima se treniraju, grupirati neoznačene podatke prema sličnostima među ulaznim podacima ili analitički predvidjeti događaje (Pathmind, 2019[61]).

Klasifikacijski zadaci ovise o označenim setovima podataka, odnosno, ljudi trebaju ručno na temelju svojih znanja, označiti setove podataka kako bi neuralna mreža mogla učiti korelacije. Ovaj princip poznat je kao nadzirano učenje (eng. *supervised learning*) (Pathmind, 2019[62]). U ovu skupinu pripadaju aktivnosti prepoznavanja lica, identificiranja ljudi na slikama, prepoznavanje izraza lica (sreća, ljutnja i sl.), identificiranje objekata na slikama, klasificiranje tekstova (e-mailova) kao spam pošte ili zlonamjernih tekstova i slično.

Grupiranje označava prepoznavanje sličnosti. Duboko učenje ne zahtijeva postojanje oznaka kako bi prepoznalo sličnosti. Ovakva vrsta učenja naziva se nenadzirano učenje (eng. *unsupervised learning*). Neoznačeni podaci predstavljaju većinu podataka u svijetu. Jedan od zakona strojnog učenja tvrdi da je ponašanje algoritma točnije što je skup podataka na kojem se trenira veći (Pathmind, 2019[63]). Stoga, nenadzirano učenje ima potencijal da proizvede

[61] Pathmind. op. cit.
[62] Ibid.
[63] Ibid.

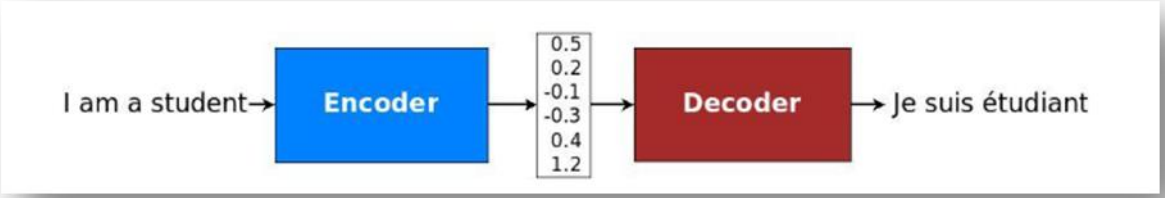
iznimno točne modele. Kao primjeri stvarne primjene mogu se navesti uspoređivanje dokumenata, slika ili zvukova sa sličnim objektima. Osim toga, nenadziranim učenjem moguće je i detektirati anomalije, odnosno neobična ponašanja unutar promatranog skupa podataka.

Prediktivna analiza odnosi se na pronalaženje korelacija između sadašnjih i budućih događaja. Izloženo dovoljnoj količini validnih podataka, duboko učenje može pronaći korelacije između sadašnjih i budućih događaja. Budući događaj je kao oznaka, a pružanjem podataka u seriji s nekoliko vremena, duboko učenje može pročitati niz brojeva i predvidjeti broj koji je najvjerojatnije da će se sljedeći pojaviti (Pathmind, 2019[64]). Prediktivnu analizu moguće je primijeniti, na primjer, u poslovanju, u učenju, u slučaju predviđanja gubitka korisnika, reakcija korisnika itd.

[64] Pathmind. op. cit.

2.4. Neuralno strojno prevođenje

Neuralno strojno prevođenje je pristup strojnom prevođenju koji koristi duboke neuralne mreže kako bi predvidio vjerojatnost ciljnog prijevoda. Za razliku od prethodnih pristupa, neuralni strojni pristup, baš kao i ljudi, uzima cijelu izvornu rečenicu, razumije njezino značenje i zatim ju prevodi (Bahdanau i sur. [65]). Kao što je prikazano na slici 7, sustav za neuralno strojno prevođenje koristi enkoder-dekoder arhitekturu. Na početku ovoga poglavlja je iznesen kratki pregled funkcioniranja i treniranja dubokih neuralnih mreža.



Slika 7. Opći prikaz enkoder-dekoder arhitekture (Luong i sur., 2017)

2.4.1. Treniranje dubokih neuralnih mreža

Proces izračuna vjerojatnosti kod neuralnih mreža podijeljen je u dvije faze. Sustav čita ulaznu rečenicu pomoću enkodera (eng. *encoder*) i pretvara ju u vektor (Luong i sur., 2017[66]). Ovaj vektor još se naziva i vektorskom reprezentacijom riječi (eng. *word embedding*) i označava vektore realnih brojeva koji odgovaraju određenim riječima u rječniku (Neubig, 2017[67]) i predstavljaju njihovo značenje. Preciznije rečeno, prema Neubig (2017[68]) u prvom koraku izračunava se skriveni sloj *h* uzimajući u obzir ulazne podatke *x* i proizvodi se vektor skrivenih varijabli *h*. Zatim se izračunava izlazni sloj na način da dekoder (eng. *decoder*) uzima varijablu te izračunava finalni rezultat *y*. Prema Neubig (2017[69]) u prvoj fazi korak funkcija (eng. *step function*) primjenjuje se na zbroj težine (eng. *weights*) pomnožene s ulaznim podacima i vrijednosti pristranosti (eng. *bias*), dok se u drugoj fazi težine množe s izračunom

[65] Bahdanau, D.; Cho, K.; Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Proceedings of the International Conference on Learning Representations (ICLR).
[66] Luong, T.; Brevdo E.; Zhao R. (2017). Neural Machine Translation (seq2seq Tutorial).
[67] Neubig, G. (2017). Neural Machine Translation and Sequence-to-sequence Models: A Tutorial.
[68] Ibid.
[69] Ibid.

skrivenog sloja te se tom iznosu dodaje vrijednost pristranosti. Korak funkcija uzima varijablu *x* te iznosi 1 ako je x > 0, inače iznosi -1 (Neubig, 2017[70]).

$$h = step\ (W_{xh}x + b_h)$$

$$y = w_{hy}h + b_y$$

Uzimajući u obzir cijelu rečenicu, procesuirajući ju i zatim prevodeći, modeli neuralnog strojnog prevođenja mogu povezati udaljene elemente u rečenici što utječe na pravilno slaganje oblika riječi s obzirom na rod, broj i lice te kvalitetniju sintaktičku strukturu i proizvodi fluentniji prijevod, čak i za manje govorene jezike, kao što je hrvatski (Seljan i sur., 2020[71]; Dunđer i sur., 2020[72]).

Kako bi se prethodno navedeni parametri W_{mh}, b_h, w_{hy} i b_y trenirali, potrebno je definirati funkciju gubitka. Treniranje je proces optimizacije težinskih vrijednosti koji se svodi na minimizaciju pogreške predviđanja neuralne mreže i istinitih vrijednosti (Vasić, 2018[73]). Prema Neubig (2017[74]) funkcija koja računa koliko predviđene vrijednosti *y* odstupaju od ispravne vrijednosti *y** naziva se *funkcija gubitka*. Ona se računa kao kvadratna razlika dviju vrijednosti.

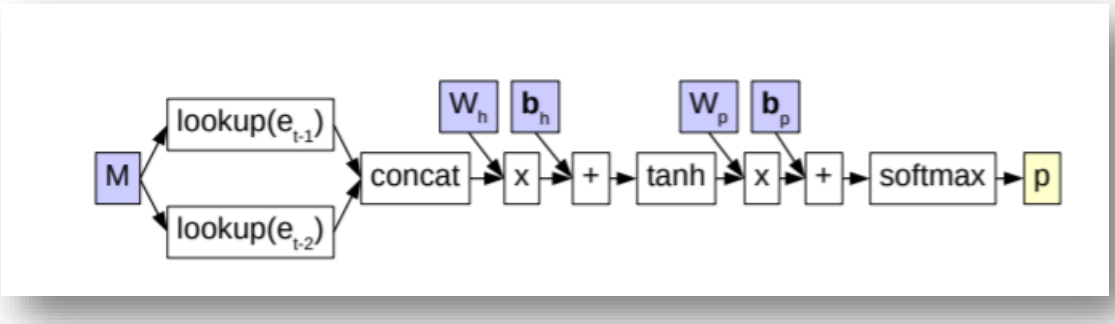
$$l(y^*, y) = (y^* - y)^2$$

Prema Vasić (2018[75]), kako bi se odredila pogreška između svakog neurona koristi se metoda nazvana propagacija unatrag/unazad (eng. *backpropagation*) preko koje se računa spust funkcije gubitka.

2.4.2. Neuralni jezični modeli

Uzimajući u obzir prethodno pojašnjene osnove funkcioniranja i treniranja neuralnih mreža, potrebno je primijeniti ih na jezično modeliranje. Za primjer je uzet 3-gramski neuralni model s jednim slojem te je koncept objašnjen na temelju znanstvenog rada *Neural Machine Translation and Sequence-to-sequence Models: A Tutorial* autora Neubig iz 2017. godine.

[70] Neubig, G. op. cit.
[71] Seljan, S.; Dunđer, I.; Pavrlovski, M. (2020). Human Quality Evaluation of Machine-Translated Poetry. MIPRO 2020.
[72] Dunđer, I.; Seljan, S.; Pavlovski, M. (2020). Automatic Machine Translation of Poetry and a Low-Resource Language Pair. MIPRO 2020.
[73] Vasić, D. op. cit.
[74] Neubig, G. op. cit.
[75] Vasić, D. op. cit.



Slika 8. Graf izračuna vjerojatnosti za 3-gramski jednosmjerni neuralni jezični model (Neubig, 2017)

Prema Neubig (2017[76]) u modelu prikazanom na slici 8, pri izračunu vjerojatnosti za neuralni jezični model prvo je potrebno dobiti rezultat vektora *m* koji predstavlja kontekst e_{i-n+1}^{i-1}. *M* ovdje predstavlja matricu sa |*V*| stupcima i L_m retcima, gdje svaki stupac odgovara vektoru za L_m podatak te predstavlja jednu riječ. Ta vrijednost označava vektorsku reprezentaciju riječi (eng. *word embedding*). Prema formuli:

$$m = concat(M._{et-2}, M._{et-1})$$

Vrijednost vektora *m* rezultat je konkatenacije vektora svake riječi danog konteksta, odnosno:

$$|m| = L_m*(n-1)$$

Vrijednost vektora *m* zatim se primjenjuje kako bi se dobio izračun skrivenog sloja:

$$h = tanh(W_{mh}m + b_h)$$

Ovim postupkom model može naučiti kombinacije značajki koje odražavaju više riječi u kontekstu što omogućava bolje rezultate pri obrađivanju kompleksnijih rečenica (Neubig, 2017[77]).

U sljedećem koraku potrebno je izračunati vektor rezultata (eng. *score vector*) za svaku riječ $s \in \mathbb{R}^{|V|}$. U teoriji procjene najveće vjerojatnosti, vektor rezultata predstavlja stupanj funkcije vjerojatnosti s obzirom na parametre koji se procjenjuju (StatLect, 2020[78]). U ovom, konkretnom slučaju on se izračunava prema sljedećoj formuli:

[76] Neubig, G. op. cit.
[77] Ibid.
[78] StatLect. (2020). Score Vector.

$$s = W_{hs}h + b_s$$

Vektor skrivenog sloja h množi se s matricom težinskih vrijednosti $W_{hs} \in \mathbb{R}^{|V|x|h|}$ te se tom umnošku dodaje vektor pristranosti $b_s \in \mathbb{R}^{|V|}$.

Vjerojatnost p je zatim izračunata primjenjivanjem *softmax* funkcije na prethodno dobiveni izračun vektora rezultata s. Prema DeepAI (2020[79]) softmax funkcija pretvara vektore čija vrijednost može biti pozitivna, negativna ili nula u p vrijednost $0 \leq p \leq 1$ kako bi mogle biti interpretirane kao vjerojatnosti.

$$p = softmax(s)$$

U slučaju treniranja moguće je izračunati funkciju gubitka, ako je poznata vrijednost e_t:

$$l = - log(pe_t)$$

2.4.3. Dinamičke povratne (rekurentne) neuralne mreže

NMT modeli variraju s obzirom na njihovu arhitekturu. Većina NMT modela za rad sa sekvencijalnim podacima koristi dinamičke povratne (rekurentne) neuralne mreže (RNNs) (Neubig, 2017[80]). Obično su i koder i dekoder rekurentne neuralne mreže, no različitih karakteristika. RNN modeli razlikuju se prema smjeru (jednosmjerni ili dvosmjerni), broju slojeva (jednoslojni ili višeslojni) i vrsti (*vanilla* rekurentna neuralna mreža, duga kratkoročna memorija ili LSTM, i mreža propusno povratnih ćelija ili GRU), prema Luong i sur. (2017[81]). Neuralni jezični modeli koriste upravo rekurentne neuralne mreže jer one imaju sposobnost obuhvaćanja jezičnih međuovisnosti (Neubig, 2017[82]). Ovo je važno kako bi došlo do pravilnog slaganja elemenata u rečenici, kao npr. u:

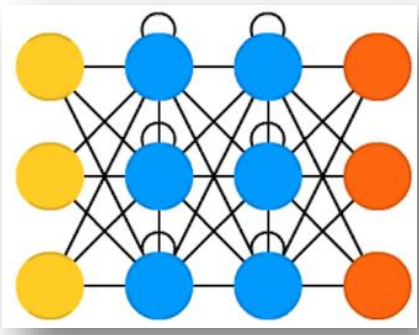
She doesn't like talking about her feelings.

Kao drugi primjer moguće je uzeti stil nekog teksta. Ukoliko se radi o formalnom jeziku službenog dokumenta, bilo bi neprirodno pronaći riječi iz svakodnevnog, govornog jezika.

Dinamička povratna neuralna mreža obuhvaća jezične međuovisnosti zahvaljujući mogućnosti da prenosi informacije među vremenskim koracima. Na primjer, ako neki od skrivenih slojeva h_{t-1} enkodira informaciju da je subjekt ulazne rečenice ženski, tu informaciju moguće je prenijeti u trenutni skriveni sloj h_t, pa zatim u h_{t+1} i tako dalje sve do kraja rečenice.

[79] DeepAI. 2020. Softmax Function.
[80] Neubig, G. op. cit.
[81] Luong, M. i sur. (2017). op. cit.
[82] Neubig, G. op. cit

Dakle, rekurentne neuralne mreže omogućavaju prijenos informacije preko arbitrarnog broja konsekutivnih vremenskih koraka (Neubig, 2017[83]) te zahvaljujući tome sustavi za neuralno strojno prevođenje proizvode fluentnije prijevode. Primjer arhitekture rekurentne neuralne mreže koji navodi Van Veen (2016[84]) prikazan je na slici 9.



Slika 9. Primjer rekurentne neuralne mreže (Van Veen, 2016)

Ideja izračuna ovog prijenosa informacija jest dodavanje veze koja referencira prethodno skriveno stanje h_{t-1} pri izračunu trenutnog skrivenog stanja *h:*

$$\boldsymbol{h}_t = \begin{cases} \tanh(W_{xh}\boldsymbol{x}_t + W_{hh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_h) & t \geq 1, \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

Slika 10. Izračun trenutnog skrivenog stanja h_t (Neubig, 2017)

Prema formuli sa slike 10, jedina razlika između standardne i rekurentne neuralne mreže jest dodavanje veze $W_{hh}h_{t-1}$ iz skrivenog sloja prethodnog koraka (*t-1*) u trenutni vremenski korak (*t*). Isto tako, važno je naglasiti da se u sljedeći skriveni sloj šalje samo skriveni sloj prethodne riječi, odnosno vremenskog koraka jer bi informacije prethodnih riječi već trebale biti uključene u korak koji se želi prenijeti u trenutni skriveni sloj h_t. Najčešće se prethodno navedeni izračun h_t kratko zapisuje s funkcijom RNN(·) te prema tome izračun vjerojatnosti izgleda kao:

[83] Neubig, G. op. cit.
[84] Van Veen, F. (2016). The Asimov Institute: The Neural Network Zoo.

$$m_t = M., e_{t-1}$$

$$h_t = RNN(m_t, h_{t-1})$$

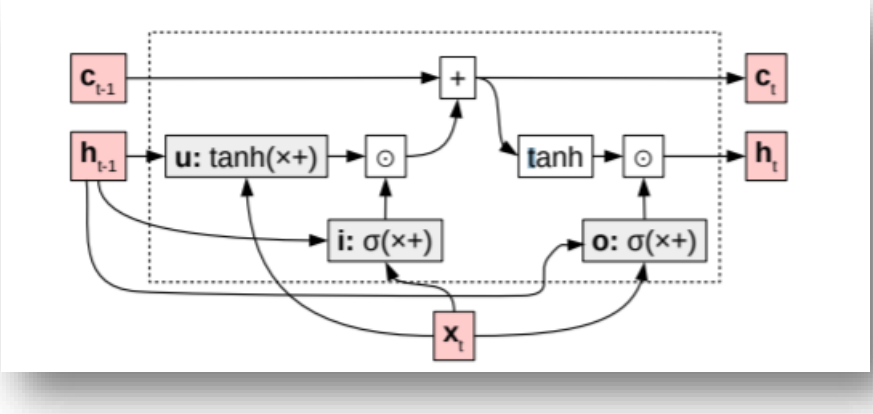
$$p_t = softmax(W_{hs}h_t + b_s).$$

Međutim, prema Vasić (2018[85]), primjenom prijenosne funkcije u intervalu od -1 do 1 i matričnim množenjem težinskih vrijednosti, slanjem podataka iz jednog koraka u drugi, informacije se vrlo brzo smanjuju i postaju 0. Ovaj problem naziva se problemom nestajućeg gradijenta (eng. *vanishing gradient* problem), a onemogućava obuhvaćanje dugoročnih ovisnosti. Kako bi se ovaj problem izbjegao koriste se druge dvije vrste dubokih neuralnih mreža, mreža duge kratkoročne memorije (eng. *long short-term memory*, LSTM) i mreža propusno povratnih ćelija (eng. *gated recurrent unit*, GRU).

2.4.3.1. Duga kratkoročna memorija (LSTM)

Duga kratkoročna memorija je vrsta duboke neuralne mreže koja osigurava da je izlazna vrijednost rekurentne funkcije ove mreže uvijek 1 (Neubig, 2017[86]). Temeljna ideja LSTM mreže je da osim skrivenog stanja *h* postoji i memorijska ćelija *c* čija je vrijednost uvijek 1 te zbog toga problem nestajućeg gradijenta nema utjecaja na informacije pohranjene u nju. Najčešća arhitektura LSTM bloka sastoji se od memorijske ćelije, ulaznih vrata (eng. *input gate*), izlaznih vrata (eng. *output gate*) i zaboravljajućih vrata (eng. *forget gate*). Prema Vasić (2018[87]) ulazna vrata kontroliraju ulazne informacije, izlazna vrata kontroliraju izlazne informacije, a zaboravljajuća vrata određuju koje će informacije ostati u memorijskoj ćeliji.

[85] Vasić, D. op. cit.
[86] Neubig, G. op. cit.
[87] Vasić, D. op. cit.



Slika 11. Prikaz arhitekture LSTM-a (Neubig, 2017)

Kao što se vidi na slici 11 arhitekture LSTM-a ulazne podatke predstavljaju ulazni vektor x_t, izlaz prethodnog LSTM bloka h_{t-1} i memorija prethodnog bloka c_{t-1}, dok izlazne podatke predstavljaju izlaz trenutnog bloka h_t i memorija trenutnog bloka c_t. Prema Neubig (2017[88]) nad ulaznim podacima, podacima prethodnog bloka i mjerom pristranosti ažurirajuća vrata u_t primjenjuju *tanh* funkciju, dok ulazna vrata i_t i izlazna vrata o_t primjenjuju sigmoidnu funkciju:

$$u_t = tanh(W_{xu}x_t + W_{hu}h_{t-1} + b_u)$$

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_t{-}1 + b_o)$$

Zatim se izračunava vrijednost memorijske ćelije množenjem vektora u_t s vektorom i_t te se zbraja s vrijednošću memorijske ćelije prethodnog koraka c_{t-1}:

$$c_t = i_t \circ u_t + c_{t-1}$$

U zadnjem koraku se izračunava skriveni sloj primjenom *tanh* funkcije na izračun vektora c_t te njegovim množenjem s vektorom o_t

$$h_t = o_t \circ tanh(c_t)$$

Postoji mnogo varijanti LSTM mreže, a najučestalija je ona varijanta koja standardnoj arhitekturi pridodaje i zaboravljajuća vrata (eng. *forget gate*). Zaboravljajuća vrata korisna su jer omogućavaju da se iz ćelije obrišu nepotrebne informacije i tako ne prenose u svaku sljedeću

[88] Neubig, G. op. cit.

memorijsku ćeliju c_t. Njihova se vrijednost izračunava primjenom sigmoidne funkcije nad ulaznim podacima, podacima prethodnog bloka i mjerom pristranosti:

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f)$$

Isto tako, u odnosu na standardni LSTM blok razlika se uočava u izračunu memorijske ćelije c_t koja se sada izračunava zbrajanjem umnoška vektora i_t s vektorom u_t s umnoškom vektora f_t i c_{t-1}.

$$c_t = i_t \circ u_t + f_t \circ c_{t-1}$$

2.4.3.2. Mreža propusno povratnih ćelija (GRU)

Drugi tip rekurentne neuralne mreže koji uspješno rješava problem nestajućeg gradijenta naziva se mreža propusno povratnih ćelija (eng. *gated recurrent unit*). Prema Vasić (2018[89]) GRU ima manje parametara od LSTM mreže jer ne sadrže izlazna vrata te nema odvojeni koncept ćelije. GRU izračunava skriveni sloj h_t množenjem prethodnog skrivenog sloja h_{t-1} s razlikom $(1\text{-}z_t)$ koja predstavlja ažurirajuća vrata (eng. *update gate*) te se zatim zbraja s umnoškom ažurirajućih vrata s kandidatom skrivenog sloja h'_t. GRU će iskoristiti vrijednost kandidata skrivenog sloja ako je vrijednost z_t blizu 0, a u suprotnom, ako je vrijednost blizu 1 iskoritit će prethodnu vrijednost h_{t-1}.

$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z)$$

$$h_t = (1 - z_t)\, h_{t-1} + z_t h'_t$$

Vrijednost kandidata skrivenog sloja h'_t izračunava se primjenom *tanh* funkcije nad ulaznim podacima, podacima prethodnog bloka i mjerom pristranosti gdje su podaci prethodnog bloka dodatno pomnoženi umnoškom resetirajućih vrata (eng. *reset gate*) r_t s vrijednošću prethodnog skrivenog sloja h_{t-1}.

$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r)$$

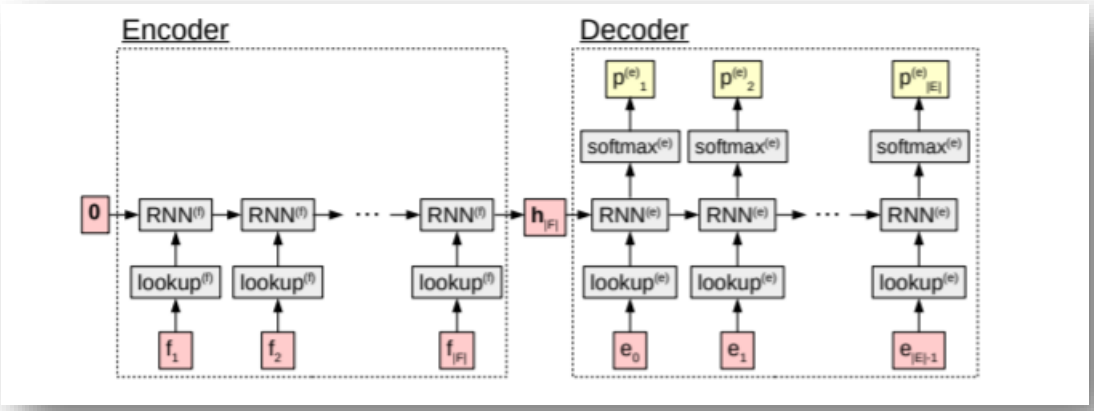
$$h'_t = tanh(W_{xh}x_t + W_{hh}(r_t \circ h_{t-1}) + b_h)$$

Ova vrsta RNN mreže, zbog jednostavnije arhitekture, uglavnom se koristi kako bi se umanjila memorija i vrijeme potrebno za izračun.

[89] Vasić, D. op. cit.

2.4.4. Enkoder-dekoder arhitektura

Kao što je već navedeno, model koji koriste sustavi za neuralno strojno prevođenje naziva se enkoder-dekoder model (eng. *encoder-decoder model*). Prema Neubig (2017[90]) prva neuralna mreža koja obrađuje ulaznu rečenicu *F* kodira informaciju kao vektor realnih brojeva (skriveno stanje *h*), a zatim je druga neuralna mreža korištena da predvidi ciljnu rečenicu *E i* dekodira ovu informaciju.



Slika 12. Prikaz enkoder-dekoder modela

Na slici 12 enkoder je prikazan kao RNN$^{(f)}$. On uzima vektorsku reprezentaciju riječi u određenom vremenskom koraku $m_t^{(f)}$ te izračunava skriveno stanje za pojedinu riječ u izvornoj rečenici $h_t^{(f)}$. Kada je sustav došao do kraja rečenice i proizveo skriveno stanje koje obuhvaća sve informacije o izvornoj rečenici, ta informacija šalje se u dekoder. Dekoder predviđa vjerojatnost pojedine riječi e_t u određenom vremenskom koraku. Dekoder uzima vektorsku reprezentaciju prethodne riječi e_{t-1} jer je vjerojatnost trenutne riječi e_t uvjetovana njome. Zatim dekoder izračunava skriveno stanje $h_t^{(e)}$ koje je uvjetovano finalnim skrivenim stanjem enkodera. Naposljetku, vjerojatnost $p_t^{(e)}$ izračunava se primjenom softmax funkcije na skriveno stanje pojedinog vremenskog koraka u ciljnoj rečenici $h_t^{(e)}$.

$$p_t^{(e)} = softmax\,(W_{hs}h_t^{(e)} + b_s)$$

[90] Neubig, G. op. cit.

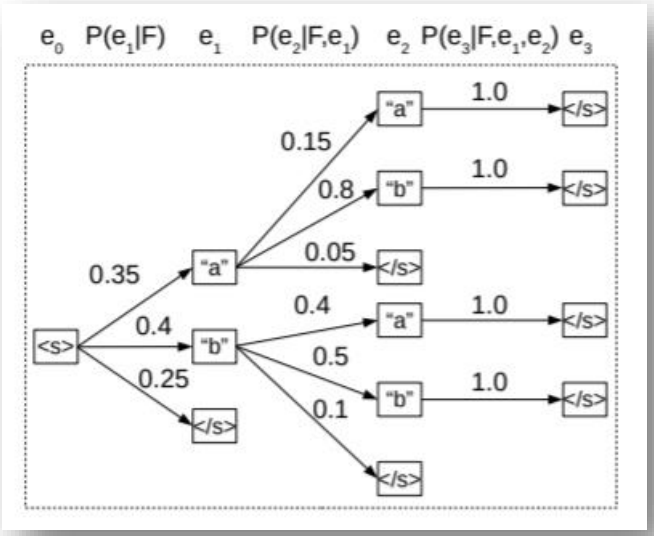
2.4.5. Generiranje prijevoda

Sustav izračunate vjerojatnosti koristi za generiranje prijevoda. Prema Luong i sur. (2017[91]) razlika između procesa treniranja i generiranja prijevoda (eng. *inference*) je u tome što kod generiranja prijevoda sustav ima pristup samo izvornoj rečenici. Postoji više načina kako ona može biti dekodirana, a prema Neubig (2017[92]) te metode uključuju nasumično uzorkovanje *(*eng. *random sampling*), *greedy* dekodiranje (eng. *greedy decoding,* 1-best search) i *beam* pretraživanje (*eng. beam search,* n-best search).

2.4.5.1. Greedy dekodiranje

Najčešće korištene metode su *greedy* dekodiranje i *beam* pretraživanje. Metoda *greedy* dekodiranja korisna je pri strojnom prevođenju i ostalim ostalim aplikacijama gdje se želi dobiti rezultat koji sustav smatra najboljim. Ova metoda izračunava vjerojatnost sljedeće riječi u vrmemenskom koraku p_t, odabire onu s najvećom vjerojatnošću i uzima ju kao sljedeću riječ u slijedu (Neubig, 2017[93]). Budući da ova metoda odabire najveću vjerojatnost u datom vremenskom koraku, sve ostale vjerojatnosti koje su izračunate na temelju varijable koja nije odabrana se ignoriraju. Međutim, moguće je da se upravo u njima kriju elementi s većim vjerojatnostima.

[91] Luong, T. i sur. (2017). op. cit.
[92] Neubig, G. op. cit.
[93] Ibid.



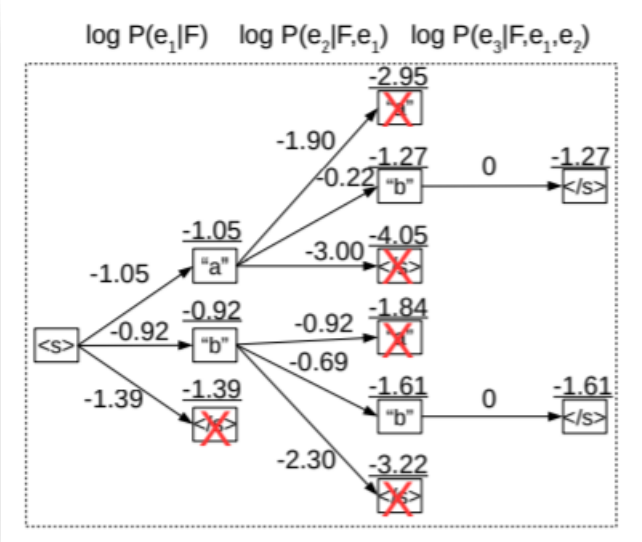
Slika 13. Greedy metoda dekodiranja (Neubig, 2017)

Na slici 13 prikazan je primjer metode *greedy* dekodiranja. U prvom koraku metodom *greedy* dekodiranja bit će odabrana varijabla „b" koja ima najveću vjerojatnost (0.4), a zatim se u odnosu na nju računa sljedeća najveća vjerojatnost koja iznosi 0.5 a pripada varijabli „b". Dakle, rezultat generiranja izlaznog podatka je „ab". Pogledaju li se vjerojatnosti nastale odabirom varijable „a" u prvom koraku, očigledno je da najveću vjerojatnost zatim ostvaruje varijabla „b" (0.8) te kao varijabla u drugom vremenskom koraku, ona ostvaruje najveću vjerojatnost, ali se ignorira.

2.4.5.2. Beam dekodiranje

Kako bi se ovaj problem izbjegao, koristi se metoda *beam* dekodiranja. Ova metoda uzima u obzir *b* pretpostavku, odnosno riječi s najvećom vjerojatnošću, gdje *b* označava širinu snopa (Neubig, 2017[94]). Na slici 14 u nastavku prikazana je metoda *beam* pretraživanja čiji snop *b* iznosi 2.

[94] Neubig, G. op. cit.



Slika 14. Beam metoda dekodiranja (Neubig, 2017)

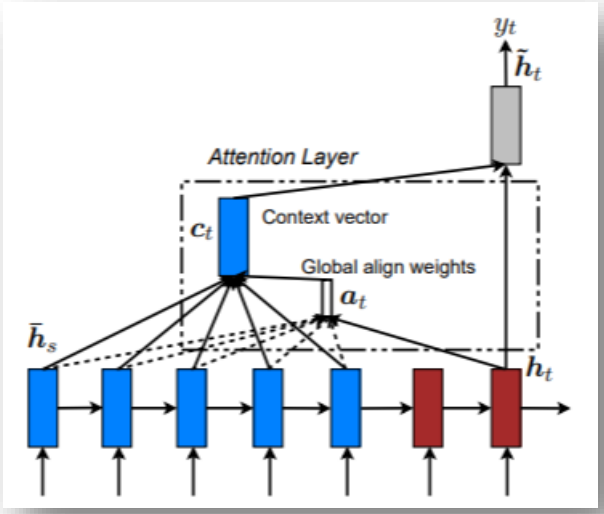
U prvom vremenskom koraku odabiru se dvije vrijednosti (*b*=2) s najvećom vjerojatnošću, u ovom slučaju to su „b” (-0.92) i „a” (-1.05), a zatim se u sljedećem vremenskom koraku za svaku od njih odabire sljedeća riječ s najvećom vjerojatnošću. Dakle nakon riječi „a” slijedila bi „b” (-0.22), a nakon „b” slijedila bi „b” (-0.69). Brojevi koji se nalaze iznad varijabli označavaju ukupnu vjerojatnost u pojedinom vremenskom koraku. Za sekvencu „ab” ona iznosi -1.27 dok za sekvencu „bb” iznosi -1.61. Stoga je moguće zaključiti kako će sustav generirati sekvencu „ab”.

2.4.6. Mehanizam pažnje

Prema Neubig (2017[95]) postoje dva problematična aspekta cjelokupne enkoder-dekoder arhitekture. Prvi problem predstavlja postojanje velike udaljenosti među riječima čija ispravna interpretacija ovisi o jednoj i drugoj, a koje trebaju biti prevedene iz izvornog jezika u ciljni. Drugi problem odnosi se na pokušaj enkoder-dekoder arhitekture da pohrani informacije o rečenicama bilo koje duljine u skriveni vektor fiksne veličine. To znači da ako je mreža premala neće biti u mogućnosti pohraniti sve informacije o dužim rečenicama, a ako je mreža prevelika, za procesuiranje kraćih rečenica će biti korištene nepotrebno velike količine memorije i vremena.

[95] Neubig, G. op. cit.

Navedeni problemi mogu se riješiti korištenjem mehanizma pažnje (eng. *attention mechanism*). Prema Luong i sur. (2017[96]) cilj primjene ovog mehanizma je uspostavljanje direktnih, kratkih veza između ciljne i izvorne rečenice obraćajući pozornost na relevantne podatke izvorne rečenice dok se prevodi. Korištenje mehanizma pažnje postalo je standardom te se koristi i u OpenNMT sustavu korištenom za istraživanje u ovom radu. Luong i sur. (2015[97]) navode globalni i lokalni model temeljen na mehanizmu pažnje. Globalni pristup uzima u obzir cijelu izvornu rečenicu, dok lokalni pristup u jednom vremenskom koraku uzima u obzir samo određene riječi izvorne rečenice. Oba pristupa u svakom vremenskom koraku *t* faze dekodiranja kao ulazni podatak uzimaju skriveni sloj ciljne rečenice $h_{t.}$ a zatim izračunavaju kontekstualni vektor c_t koji obuhvaća relevantne informacije iz izvorne rečenice kako bi pomogao u predviđanju ciljne riječi y_t. Glavna razlika između globalnog i lokalnog modela upravo je u načinu izračuna kontekstualnog vektora $c_{t.}$

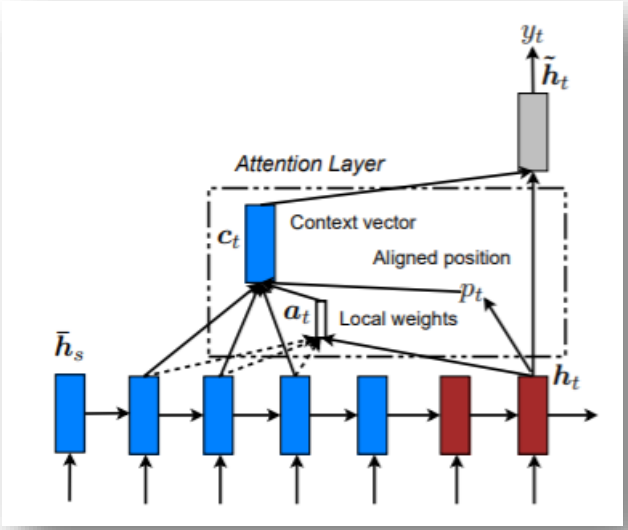


Slika 15. Globalni model mehanizma pažnje (Luong i sur., 2015)

U globalnom modelu pažnje prikazanom na slici 15, u svakom vremenskom koraku proizvodi se vektor dužine ulazne rečenice a_t na temelju vrijednosti skrivenog sloja ciljne rečenice h_t i svih skrivenih stanja izvorne rečenice $h'_{s.}$ Globalni kontekstualni vektor c_t zatim je

[96] Luong, T. i sur. (2017). op. cit.
[97] Luong, T.; Pham H.; Manning C. D. (2015). Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1412–1421.

izračunat kao ponderirani, težinski prosjek umnoška vrijednosti a_t u svim stanjima izvorne rečenice.



Slika 16. Lokalni model mehanizma pažnje (Luong i sur., 2015)

S druge strane, u lokalnom modelu pažnje prikazanom na slici 16, prvo se proizvodi jedinstveni poravnati položaj p_t (eng. *single aligned position*) za trenutnu riječ u ciljnoj rečenici. Prostor koji se nalazi oko izvorne pozicije varijable p_t koristi se za izračun kontekstualnog vektora c_t. Prema Luong i sur. (2015[98]) on se izračunava kao ponderirani, težinski prosjek skrivenog sloja izvorne rečenice koja se nalazi u tom prostoru. Težinski vektori pažnje a_t izračunati su iz trenutnog skrivenog sloja ciljne rečenice h_t i skrivenog sloja izvorne rečenice h'_s koji se nalaze u okolnom prostoru određenom varijablom p_t.

Kombinacijom vrijednosti h_t i c_t izračunava se vektor pažnje skrivenog sloja h_t (eng. *attentional hidden state*) te se u sljedećem koraku primjenom softmax funkcije na vrijednost h_t izračunava vjerojatnost (Luong i sur., 2015[99]):

$$h'_t = tanh\ (W_c[c_t\ ;h^t\ ])$$

$$p\ (y_t\ |y_{<t}, x) = softmax\ (W_s h'_t)$$

[98] Luong, T. i sur. (2015). op. cit.
[99] Ibid.

2.5. Usporedba statističkog i neuralnog strojnog prevođenja

Pored najnovijeg i najefikasnijeg pristupa strojnom prevođenju, neuralnog, dugi niz godina primjenjivao se statistički pristup strojnom prevođenju. S obzirom da je i ono još uvijek aktualno, u ovom poglavlju prikazane su prednosti i nedostaci obaju pristupa.

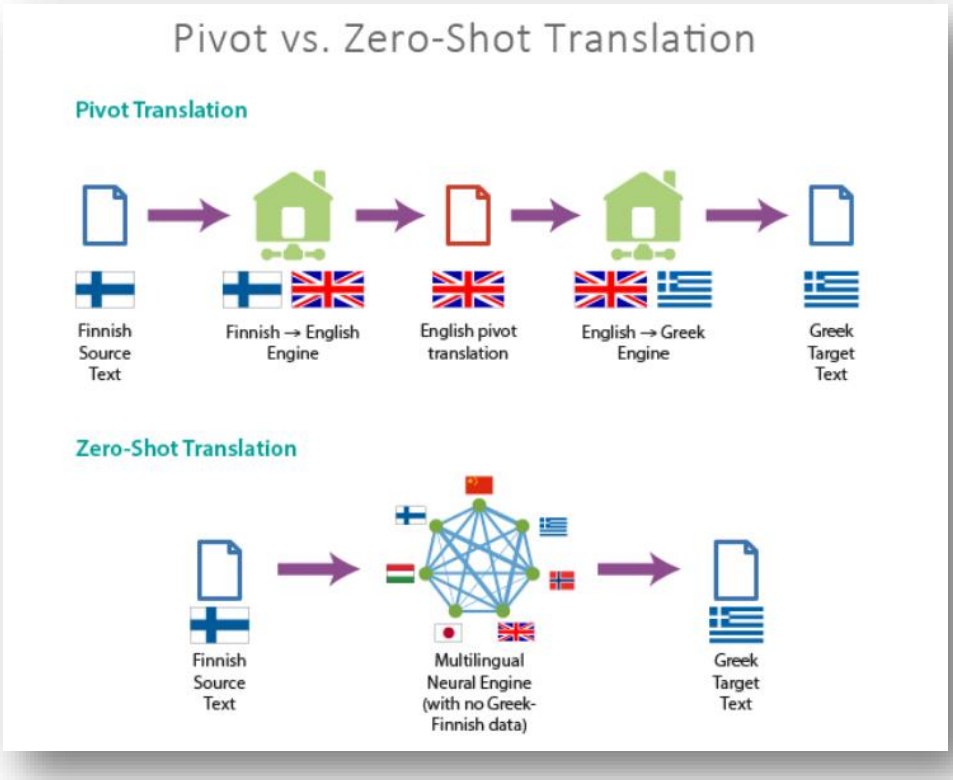
Prema Medium (2017[100]) u usporedbi sa statističkim strojnim prevođenjem, neuralnim strojnim prevođenjem je omogućeno istovremeno treniranje više značajki i nije potrebno prethodno domensko znanje, što omogućava *zero-shot* prevođenje. Prema CSA Research (2017[101]) to je sposobnost sustava da prevodi tekst iz jednog jezika u drugi, a da prethodno nije bio treniran za taj jezični par. Suprotno ovom konceptu, SMT sustavi grade dvojezične tablice za odabrani jezični par i povezuju individualne jezike po principu jedan na jedan (Dunđer, 2015[102]) te ovise o domeni u kojoj je treniran sustav (Seljan i Dunđer, 2015b[103], Brkić i sur., 2013[104], Brkić i sur., 2009[105], Seljan i sur., 2015[106]). Dakle, ovakvi sustavi ne mogu prevoditi između dva jezika za koja ne postoji trenirani, individualni sustav, osim ako ne koriste treći, dijeljeni jezik – pivot jezik (CSA Research, 2017[107]). Prema primjeru prikazanom na slici 17, ako sustav treba prevesti iz finskog jezika u grčki, ali ne postoji sustav treniran na skupu podataka posebno za taj jezični par, sustav će prijevod odraditi korištenjem finsko-engleskog sustava kako bi preveo tekst u engleski (pivot jezik), a zatim će iskoristiti englesko-grčki sustav kako bi proizveo prijevod u ciljnom jeziku. Iako je prijevod ostvaren, rezultati najčešće nisu dobri jer je mogućnost grešaka znatno veća pri prevođenju između 3 jezika. Greške iz prvog prijevoda, od izvornog do pivot jezika zatim se preslikavaju u prijevod u ciljnom jeziku.

Suprotno tome, prema Google AI Blog (2016[108]) Google neuralni sustav podacima trenira jedan zajednički sustav što dopušta izgradnju veza između više jezika umjesto

[100] Medium. op. cit.
[101] CSA Research. op. cit.
[102] Dunđer, I. (2015). Sustav za statističko strojno prevođenje i računalna adaptacija domene. Doktorska disertacija. Sveučilište u Zagrebu.
[103] Seljan, S.; Dunđer, I. (2015). Automatic quality evaluation of machine-translated output in sociological-philosophical-spiritual domain. Central European Conference on Information and Intelligent Systems (CECIIS), 318.
[104] Brkić, M.; Seljan, S.; Vičić, T. (2013). Automatic and human evaluation on english-croatian legislative test set. Intelligent Text Processing and Computational Linguistics, Springer, 311-317.
[105] Brkić, M.; Vičić, T.; Seljan, S. (2009). Evaluation of the statistical machine translation service for Croatian-English. INFuture 2009 : Digital resources and knowledge sharing, 319-332.
[106] Seljan, S.; Klasnić, K.; Stojanac, M.; Pašorda, B.; Mikelić Preradović, N. (2015). Information Transfer through Online Summarizing and Translation Technology. INFuture2015: e-Institutions–Openness, Accessibility, and Preservation.
[107] CSA Research. op. cit.
[108] Google AI Blog. (2016). Zero-Shot Translation with Google's Multilingual Neural Machine Translation System.

individualno između svakog para zasebno. Ako ne postoje podaci za pojedini jezični par, softver koristi inferencijalnu logiku kako bi izdvojio ispravni prijevod od ostalih potencijalnih prijevoda koji sadrže relevantne podatke. Ovaj pristup naziva se *zero-shot* prevođenje, a prema Johnson i sur. (2016[109]) ono označava sposobnost sustava da prevede rečenicu iz izvornog u zadani ciljni jezik bez da je prethodno bio treniran na primjerima za taj specifični, traženi jezični par. U prethodno prikazanom slučaju, ako ne postoje korisni podaci za finsko-grčki jezični par, sustav promatra korelacije između drugih jezika i proizvodi prijevod. Prijevod očekivano neće biti jednako kvalitetan kao onaj u slučaju kada je sustav treniran za taj jezični par. Važnost *zero-shot* prijevoda najviše se očituje pri prevođenju između rijetkih jezičnih parova (CSA Research, 2017[110]), a osim finskog-grčkog tu možemo svrstati i hrvatski jezik u kombinaciji s mnogim jednako tako manje rasprostranjenim jezicima. Primjer i usporedba *zero-shot* prevođenja i prevođenja pomoću pivot jezika prikazani su na slici 17. Ova prednost NMT-a također dolazi do izražaja u kontekstu Europske unije koja se na dnevnoj razini suočava s poteškoćama osiguravanja pristupa informacijama u svih 27 službenih i radnih jezika (Seljan i Gašpar, 2009[111]).

[109] Johnson, M.; Schuster, M; Q., Le V.; Krikun, M.; Wu, Y.; Chen, Z.; … & Hughes, M. (2016). Google's multilingual neural machine translation system: enabling zero-shot translation. arXiv preprint arXiv:1611.04558.
[110] CSA Research. op. cit.
[111] Seljan, S.; Gašpar, A. (2009). Primjena prevoditeljskih alata u EU i potreba za hrvatskim tehnologijama. Jezična politika i jezična stvarnost, 617-625.



Slika 17. Usporedba pivot i "zero-shot" prijevoda (CSA Research, 2017)

Nadalje, prema Medium (2017[112]) osim jasnije rečenične strukture, strojni prijevodi nastali korištenjem neuralnog pristupa imaju manje morfoloških i sintaktičkih grešaka te grešaka u rasporedu riječi u rečenici u odnosu na one nastale statističkim pristupom. Međutim, i neuralni pristup ima svoje nedostatke, poput sporog procesa treniranja i dekodiranja, nekonzistentnih prijevoda iste riječi (Medium, 2017[113]).

Prema podacima prikazanima na slici 18, prednosti neuralnog pristupa strojnom prevođenju su:

- moguće je izgraditi sustav za prevođenje s manjim setom podataka za treniranje sustava (no njegova kvaliteta neće biti visoka),
- prevodi rečenicu po rečenicu što rezultira ispravnijom i fluentnijom strukturom rečenice,

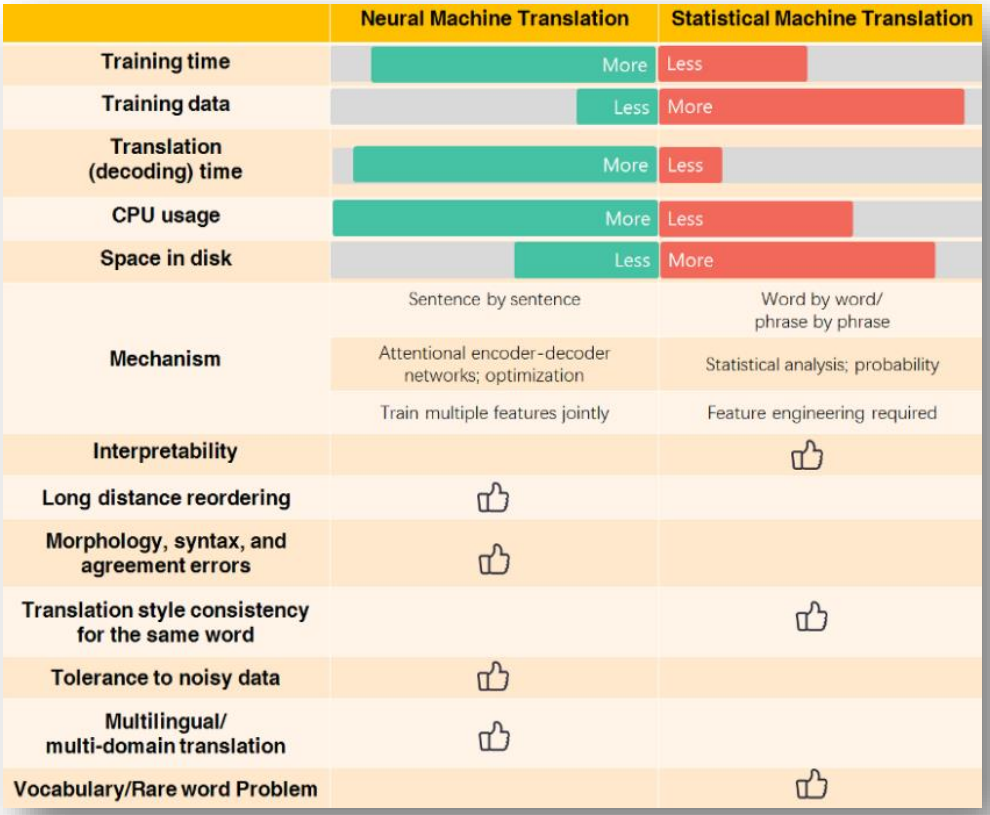
[112] Medium. op. cit.
[113] Ibid.

- omogućava kvalitetne prijevode dužih rečenica razmještanjem riječi u polaznoj rečenici,
- prijevodi sadrže manje morfoloških i sintaktičkih grešaka te grešaka u slaganju prema licu, broju i rodu,
- "bučni podaci" ne onemogućavaju potpuno proces prevođenja,
- višedomenski i višejezični prijevodi omogućeni *zero-shot* principom prevođenja.

S druge strane, prednosti statističkog strojnog prevođenja u odnosu na neuralno su:

- vrijeme treniranja sustava je kraće,
- potrebno je manje vremena za prijevod (dekodiranje),
- potrebna je manja CPU snaga stroja na kojem se pokreće sustav,
- interoperabilnost sustava,
- konzistentan stil prevođenja iste riječi kroz više prijevoda,
- točniji prijevodi "rijetkih" riječi.

Svaki od pristupa bolji je u nekom segmentu, no uzimajući u obzir svaki aspekt od prikupljanja podataka, izgradnje sustava, treniranja, prevođenja i rezultirajućih prijevoda, može se zaključiti kako neuralni pristup strojnom prevođenju prednjači nad svim dosadašnjim pristupima.



Slika 18. Usporedba karakteristika NMT-a i SMT-a (Medium, 2017)

3. Istraživanje

U sljedećim poglavljima opisan je tijek istraživanja, korištene metode i postupci. Prvo je definiran cilj istraživanja, metodologija i tijek, korišteni sustav, podatkovni skupovi korišteni za treniranje i testiranje sustava te proces izgradnje sustava za neuralno strojno prevođenje. Na kraju su prikazani rezultati istraživanja, rezultati automatske evaluacije i analiza rezultata.

3.1. Cilj istraživanja

Cilj ovog istraživanja je prikazati mogućnost pripreme i primjene sustava za neuralno strojno prevođenje za englesko-hrvatski jezični par korištenjem otvorenog koda namijenjenog istraživanju i razvijanju sustava temeljenih na dubokim neuralnim mrežama i dubokom učenju. Nadalje, cilj je i prikazati koliko je vremena potrebno za izgradnju sustava s obzirom na performanse računala te kakve rezultate takav sustav može polučiti. Važno je naglasiti kako cilj ovog istraživanja nije izgradnja visokokvalitetnog sustava za neuralno stojno prevođenje čiji bi prijevodi bili usporedivi sa svjetski poznatim online sustavima za strojno prevođenje kao što su Google Translate, Bing Microsoft Translator i ostalima, zbog tehničkih ograničenja računala na kojem je sustav izgrađen. Nakon izgradnje sustava izvršena je evaluacija prijevoda korištenjem automatske metrike te je provedena analiza prevedenih rečenica.

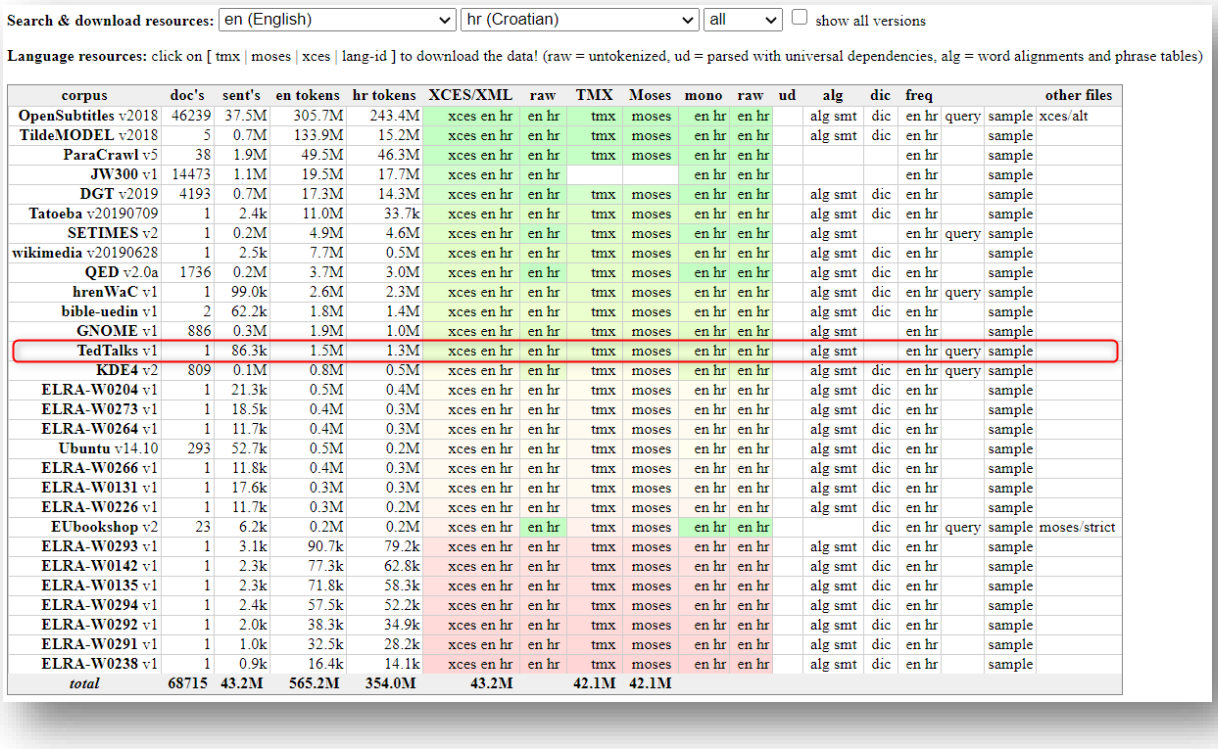
3.2. Metodologija i tijek istraživanja

Kako bi se dokazali svi prethodno navedeni koncepti neuralnog strojnog prevođenja, izgrađen je sustav za neuralno strojno prevođenje korištenjem otvoren koda OpenNMT[114]. Nakon ostvarivanja preduvjeta postavljanja radne okoline, prikupljeni su podatkovni skupovi korišteni za treniranje sustava.

Korpus je preuzet s web stranice projekta OPUS[115]. OPUS predstavlja stalno rasteću zbirku prevedenih i poravnatih tekstova s interneta (OPUS, 2012[116]). Na stranici je moguće filtrirati prikaz korpusa prema izvornom i ciljnom jeziku. Na prikazu se nalaze informacije o nazivu korpusa, broju datoteka od kojih se pojedini korpus sastoji, broju paralelnih, poravnatih

[114] https://github.com/OpenNMT/OpenNMT-py
[115] http://opus.nlpl.eu/
[116] Tiedemann, J. (2012) Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012).

rečenica i broju tokena za izvorni i ciljni jezik. Na slici 19 prikazani su razni korpusi i formati u kojima je moguće preuzeti korpus za englesko-hrvatski jezični par.



Search & download resources: en (English) hr (Croatian) all show all versions

Language resources: click on [tmx | moses | xces | lang-id] to download the data! (raw = untokenized, ud = parsed with universal dependencies, alg = word alignments and phrase tables)

corpus	doc's	sent's	en tokens	hr tokens	XCES/XML	raw	TMX	Moses	mono	raw	ud	alg	dic	freq			other files
OpenSubtitles v2018	46239	37.5M	305.7M	243.4M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr	query	sample	xces/alt
TildeMODEL v2018	5	0.7M	133.9M	15.2M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
ParaCrawl v5	38	1.9M	49.5M	46.3M	xces en hr	en hr	tmx	moses	en hr	en hr				en hr		sample	
JW300 v1	14473	1.1M	19.5M	17.7M	xces en hr	en hr			en hr	en hr				en hr		sample	
DGT v2019	4193	0.7M	17.3M	14.3M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
Tatoeba v20190709	1	2.4k	11.0M	33.7k	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
SETIMES v2	1	0.2M	4.9M	4.6M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt		en hr	query	sample	
wikimedia v20190628	1	2.5k	7.7M	0.5M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
QED v2.0a	1736	0.2M	3.7M	3.0M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
hrenWaC v1	1	99.0k	2.6M	2.3M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr	query	sample	
bible-uedin v1	2	62.2k	1.8M	1.4M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
GNOME v1	886	0.3M	1.9M	1.0M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt		en hr		sample	
TedTalks v1	1	86.3k	1.5M	1.3M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt		en hr	query	sample	
KDE4 v2	809	0.1M	0.8M	0.5M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr	query	sample	
ELRA-W0204 v1	1	21.3k	0.5M	0.4M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
ELRA-W0273 v1	1	18.5k	0.4M	0.3M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
ELRA-W0264 v1	1	11.7k	0.4M	0.3M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
Ubuntu v14.10	293	52.7k	0.5M	0.2M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
ELRA-W0266 v1	1	11.8k	0.4M	0.3M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
ELRA-W0131 v1	1	17.6k	0.3M	0.3M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
ELRA-W0226 v1	1	11.7k	0.3M	0.2M	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
EUbookshop v2	23	6.2k	0.2M	0.2M	xces en hr	en hr	tmx	moses	en hr	en hr			dic	en hr	query	sample	moses/strict
ELRA-W0293 v1	1	3.1k	90.7k	79.2k	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
ELRA-W0142 v1	1	2.3k	77.3k	62.8k	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
ELRA-W0135 v1	1	2.3k	71.8k	58.3k	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
ELRA-W0294 v1	1	2.4k	57.5k	52.2k	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
ELRA-W0292 v1	1	2.0k	38.3k	34.9k	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
ELRA-W0291 v1	1	1.0k	32.5k	28.2k	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
ELRA-W0238 v1	1	0.9k	16.4k	14.1k	xces en hr	en hr	tmx	moses	en hr	en hr		alg smt	dic	en hr		sample	
total	**68715**	**43.2M**	**565.2M**	**354.0M**	**43.2M**		**42.1M**	**42.1M**									

Slika 19. OPUS korpusi za EN-HR jezični par

Za potrebe izgradnje ovog sustava odabran je TedTalks korpus u formatu Moses koji sprema tekstualnu datoteku u UTF-8 zapisu (Tiedemann, 2012[117]). Korpus se sastoji od dviju tekstualnih datoteka od kojih svaka sadrži 86.300 rečenica, prva u izvornom jeziku (engleski), a druga u ciljnom jeziku (hrvatski). Za potrebe treniranja sustava korišten je dio korpusa od 25.000 segmenata u oba jezika te je pročišćen i tokeniziran. Nakon treniranja sustava pripremljena je datoteka na engleskom jeziku sa setom rečenica koje će se prevoditi iz istog TedTalks korpusa. Rezultati strojnog prijevoda evaluirani su BLEU metodom evaluacije.

3.3. OpenNMT

OpenNMT je sustav otvorenog koda za neuralno strojno prevođenje i učenje neuralnih nizova (eng. *neural sequence learning*) uveden u prosincu 2016. godine. Sustav ima veliku

[117] Tiedemann, J. op. cit.

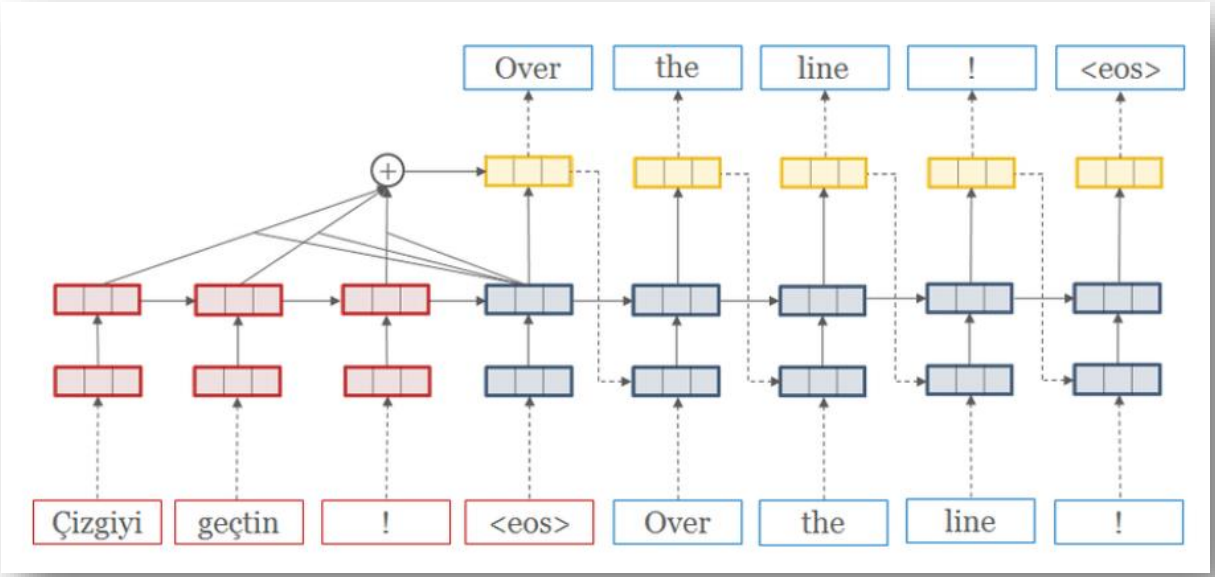
zajednicu razvojnih programera. Odlike koda su modularnost, efikasnost i proširivost (Klein i sur., 2018)[118] što znatno olakšava istraživanje drugim znanstvenicima. Projekt su započeli Harvard NLP grupa i SYSTRAN u prosincu 2016. godine, a OpenNMT je od tada korišten u nekoliko istraživanja. OpenNMT pruža implementacije u dva različita okvira za duboko učenje – Pytorch i TensorFlow (Klein i sur., 2018[119]).

Prema Klein i sur. (2018[120]) OpenNMT funkcionira na način da oblikuje vjerojatnost ciljne rečenice $w_{1:T}$ s obzirom na izvornu rečenicu $x_{1:S}$ kao

$$p(w_{1:T}|x) = \Pi_1^T p(w_t|w_{1:t-1}, x; \theta)$$

gdje je distribucija parametrizirana kao 0. Distribucija je određena korištenjem enkoder-dekoder arhitekture. Rekurentna neuralna mreža (RNN) koja kodira izvornu rečenicu mapira svaku riječ u vektor te riječi i zatim ih procesuira u slijed skrivenih vektora $h_1, \ldots, h_s$. Ciljni dekoder kombinira RNN skrivene reprezentacije prethodno generiranih riječi ($w_1, \ldots, w_{t-1}$) sa skrivenim vektorima izvornih riječi, odnosno rečenice kako bi predvidio rezultat, tj. vjerojatnost za svaku moguću sljedeću riječ. U sljedećem koraku iskorišten je *softmax* sloj, odnosno *softmax* funkcija kako bi proizvela distribuciju za sljedeću riječ $p(w_t|w_{1:t-1}, x; 0)$. Skriveni vektori izvorne rečenice utječu na distribuciju kroz sloj sažimanja s mehanizmom pažnje koji važe svaku riječ izvorne rečenice u odnosu na njezin očekivani doprinos predviđanju ciljne riječi. Cjelokupni model treniran je od početka do kraja (eng. *end-to-end*).

[118] Klein G.; Kim Y.; Deng Y.; Nguyen V.; Senellart J.; Rush A. M. (2018). OpenNMT: Neural Machine Translation Toolkit. arXiv:1805.11462 [cs.CL].
[119] Ibid.
[120] Ibid.



Slika 20. Prikaz koncepta neuralnog strojnog prevođenja (Klein i sur., 2018)

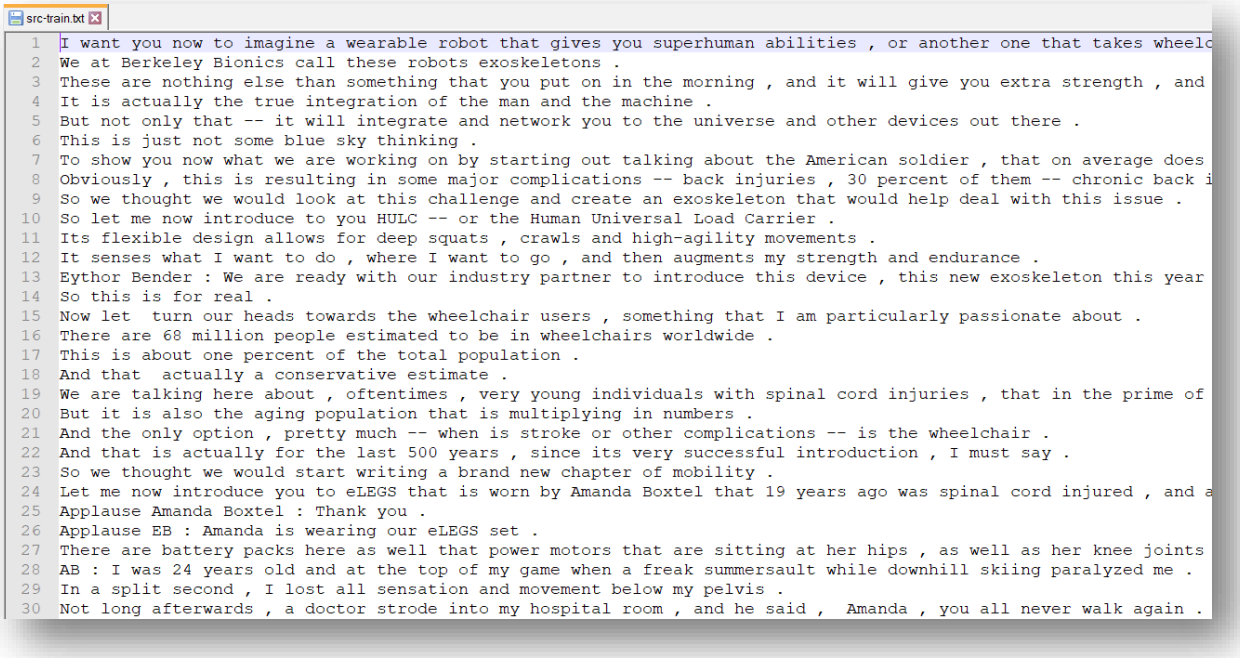
Na slici 20 prikazan je koncept neuralnog strojnog prevođenja prema Klein i sur. (2018[121]). Crveno označene riječi prvo su mapirane u vektorske reprezentacije riječi i zatim proslijeđene u rekurentnu neuralnu mrežu. Kada se dostigne <*eos*> simbol, finalni vemenski korak pokreće ciljnu (plavo označenu) rekurentnu neuralnu mrežu. U svakom vremenskom koraku ciljne rečenice, mehanizam pažnje iznova se primjenjuje preko izvorne rekurentne neuralne mreže i kombinira s trenutnim skrivenim stanjima kako bi se proizvele pretpostavke o sljedećoj riječi. Ova pretpostavka je zatim pohranjena natrag u ciljnu rekurentnu neuralnu mrežu, a postupak se ponavlja dok se ne proizvede pretpostavka o ciljnoj rečenici.

3.4. Podatkovni skupovi za treniranje i testiranje sustava

Korpus odabran za treniranje sustava sastoji se od jedne datoteke u kojoj se nalaze dvije netokenizirane, tekstualne datoteke od kojih svaka sadrži 86.300 rečenica, prva u izvornom jeziku (engleski), a druga u ciljnom jeziku (hrvatski). U tom korpusu su prikupljeni govori koje su pojedine osobe održale tijekom raznih TedTalks konferencija. S obzirom na to da se radi o govornom jeziku i mnoštvu tema koje govornici iznose, može se zaključiti da korpus pripada

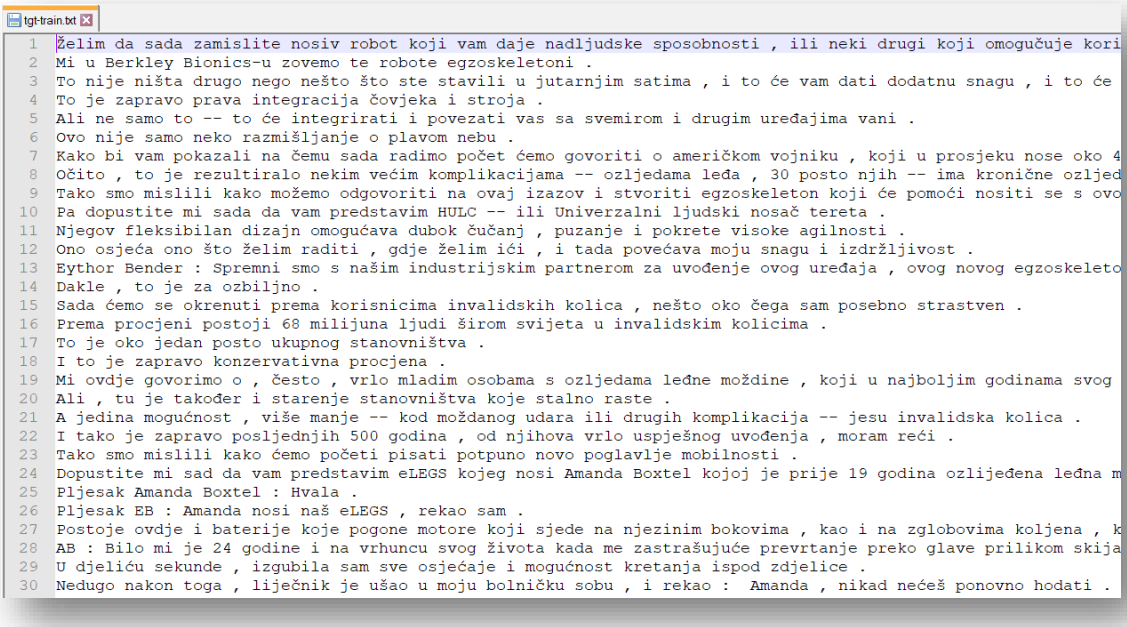
[121] Klein, G. i sur. op. cit.

općoj domeni. Za potrebe treniranja sustava uzet je dio korpusa, 25.000 rečenica za engleski jezik i prijevodi tih 25.000 rečenica na hrvatski jezik. Kako bi sustav bio kvalitetan prije pokretanja treniranja bilo je potrebno pročistiti tekst od redundatnih znakova i tokenizirati ga. U nastavku je prikazan dio korpusa pripremljenog za treniranje sustava. Slika 21 prikazuje korpus izvornog, a slika 22 korpus ciljnog jezika.


src-train.txt

```
1  I want you now to imagine a wearable robot that gives you superhuman abilities , or another one that takes wheelc
2  We at Berkeley Bionics call these robots exoskeletons .
3  These are nothing else than something that you put on in the morning , and it will give you extra strength , and
4  It is actually the true integration of the man and the machine .
5  But not only that -- it will integrate and network you to the universe and other devices out there .
6  This is just not some blue sky thinking .
7  To show you now what we are working on by starting out talking about the American soldier , that on average does
8  Obviously , this is resulting in some major complications -- back injuries , 30 percent of them -- chronic back i
9  So we thought we would look at this challenge and create an exoskeleton that would help deal with this issue .
10 So let me now introduce to you HULC -- or the Human Universal Load Carrier .
11 Its flexible design allows for deep squats , crawls and high-agility movements .
12 It senses what I want to do , where I want to go , and then augments my strength and endurance .
13 Eythor Bender : We are ready with our industry partner to introduce this device , this new exoskeleton this year
14 So this is for real .
15 Now let  turn our heads towards the wheelchair users , something that I am particularly passionate about .
16 There are 68 million people estimated to be in wheelchairs worldwide .
17 This is about one percent of the total population .
18 And that  actually a conservative estimate .
19 We are talking here about , oftentimes , very young individuals with spinal cord injuries , that in the prime of
20 But it is also the aging population that is multiplying in numbers .
21 And the only option , pretty much -- when is stroke or other complications -- is the wheelchair .
22 And that is actually for the last 500 years , since its very successful introduction , I must say .
23 So we thought we would start writing a brand new chapter of mobility .
24 Let me now introduce you to eLEGS that is worn by Amanda Boxtel that 19 years ago was spinal cord injured , and a
25 Applause Amanda Boxtel : Thank you .
26 Applause EB : Amanda is wearing our eLEGS set .
27 There are battery packs here as well that power motors that are sitting at her hips , as well as her knee joints
28 AB : I was 24 years old and at the top of my game when a freak summersault while downhill skiing paralyzed me .
29 In a split second , I lost all sensation and movement below my pelvis .
30 Not long afterwards , a doctor strode into my hospital room , and he said ,  Amanda , you all never walk again .
```

Slika 21. SRC datoteka: engleski korpus za treniranje sustava



Slika 22. TGT datoteka: hrvatski korpus za treniranje sustava

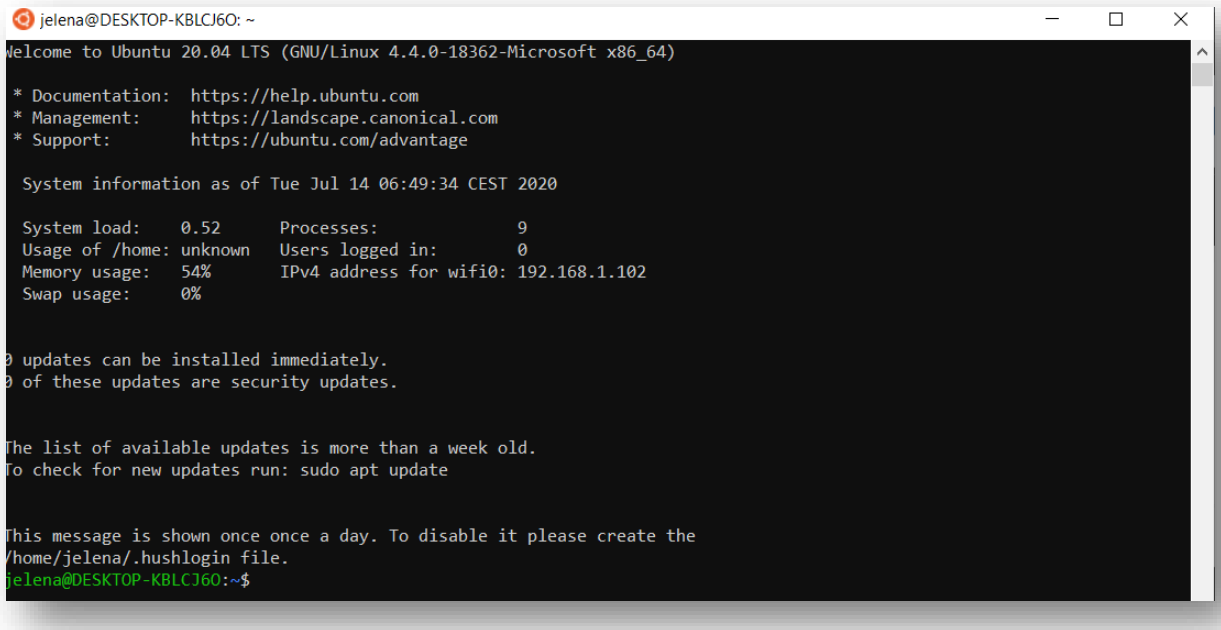
Podatkovni skup za validaciju predstavlja 5163 nasumično odabranih rečenica na izvornom jeziku i njihove odgovarajuće prijevode na ciljnom jeziku. Te rečenice ekstrahirane su iz istog korpusa, no ne pripadaju skupu rečenica odabranih za treniranje sustava. Validacijski set predstavlja skup rečenica korištenih za evaluaciju konvergencije procesa treniranja. Model koji postigne najnižu perpleksnost na ovom setu podataka, smatra se najboljim. Nakon treniranja sustava pripremljena je datoteka sa setom rečenica koje će se prevoditi. Te rečenice također čine dio TedTalks korpusa. Rezultati strojnog prijevoda evaluirani su BLEU metodom evalucije.

3.5. Izgradnja sustava za neuralno strojno prevođenje

Za izgradnju sustava korišteno je računalo sljedećih karakteristika:

- operativni sustav: OS Windows 10
- procesor: Intel(R) Core (TM) i3-8145U CPU @ 2.10 GHz, 2304 Mhz, 2 Core(s), 4 Logical Processor(s)
- Memorija: 8192 MB RAM

Za izgradnju sustava potrebno je imati instaliran Linux operativni sustav. U ovom slučaju odabrana je opcija instalacije Linux podsustava, tzv. WSL (Windows Subsystem for Linux). Kako bi treniranje sustava bilo što brže idealno bi bilo da računalo na kojem se radi posjeduje GPU (eng. *graphics processing unit*) koja omogućava znatno brže treniranje sustava s većim setovima podataka. Posjedovanje GPU nije obavezno, sami CPU (eng. *central processing unit*) može djelomično odraditi tu zadaću iako znatno sporije i vrlo ograničeno. Budući da računalo na kojem je izgrađen sustav ne posjeduje GPU, samo treniranje sustava trajalo je mnogo duže. Nakon instalacije WSL-a potebno je instalirati Ubuntu Linux operativni sustav[122]. U trenutku kada je sustav izgrađivan, 30. svibnja 2020., instalirana je verzija 20.04 što je i prikazano na slici 23 koja prikazuje sučelje Ubuntu konzole. Posljednji preduvjet je instaliranje programskog jezika.



Slika 23. Ubuntu konzola

Prije izgradnje sustava potrebno je osigurati da je sustav ažuriran.

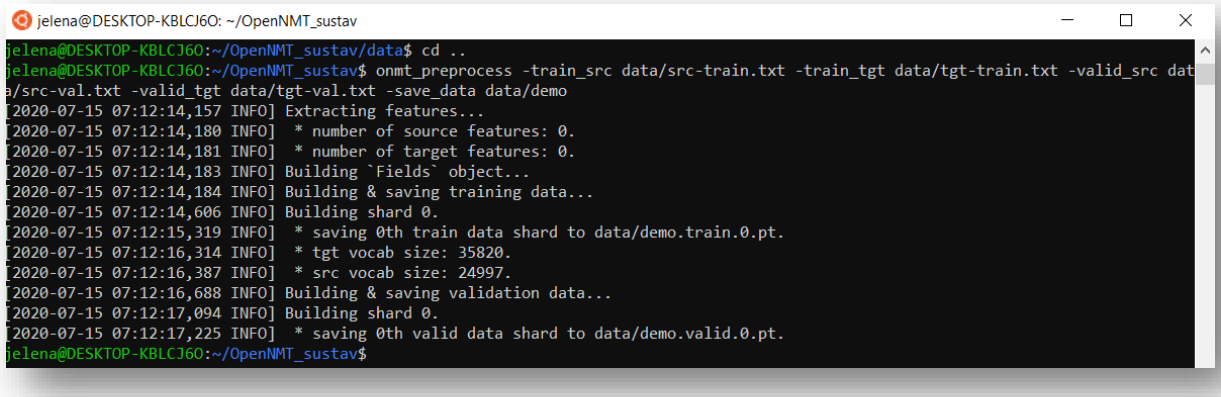
Kada su svi preduvjeti zadovoljeni, potrebno je instalirati OpenNMT sustav u PyTorch ili TensorFlow varijanti. CPU se sastoji od nekoliko jezgri optimiziranih za sekvencijalnu, serijsku obradu podataka, dok se GPU sastoji od tisuću paralelnih manjih jezgri, namijenjenih

[122] https://ubuntu.com/download/desktop

za rad s više zadataka istovremeno (PCChip, 2016[123]). Posjedovanje GPU-a u ovom području primjene znatno bi ubrzalo proces treniranja sustava i prevođenja. Nakon toga slijedi instalacija sustava.

U sljedećem koraku su pripremljeni setovi podataka opisani u poglavlju 3.4. Pripremljene datoteke prebačene su iz Windows operativnog sistema u Linux omogućavanjem pristupa i učitavanjem željene lokacije particije na lokaciju u Linux operativnom sustavu.

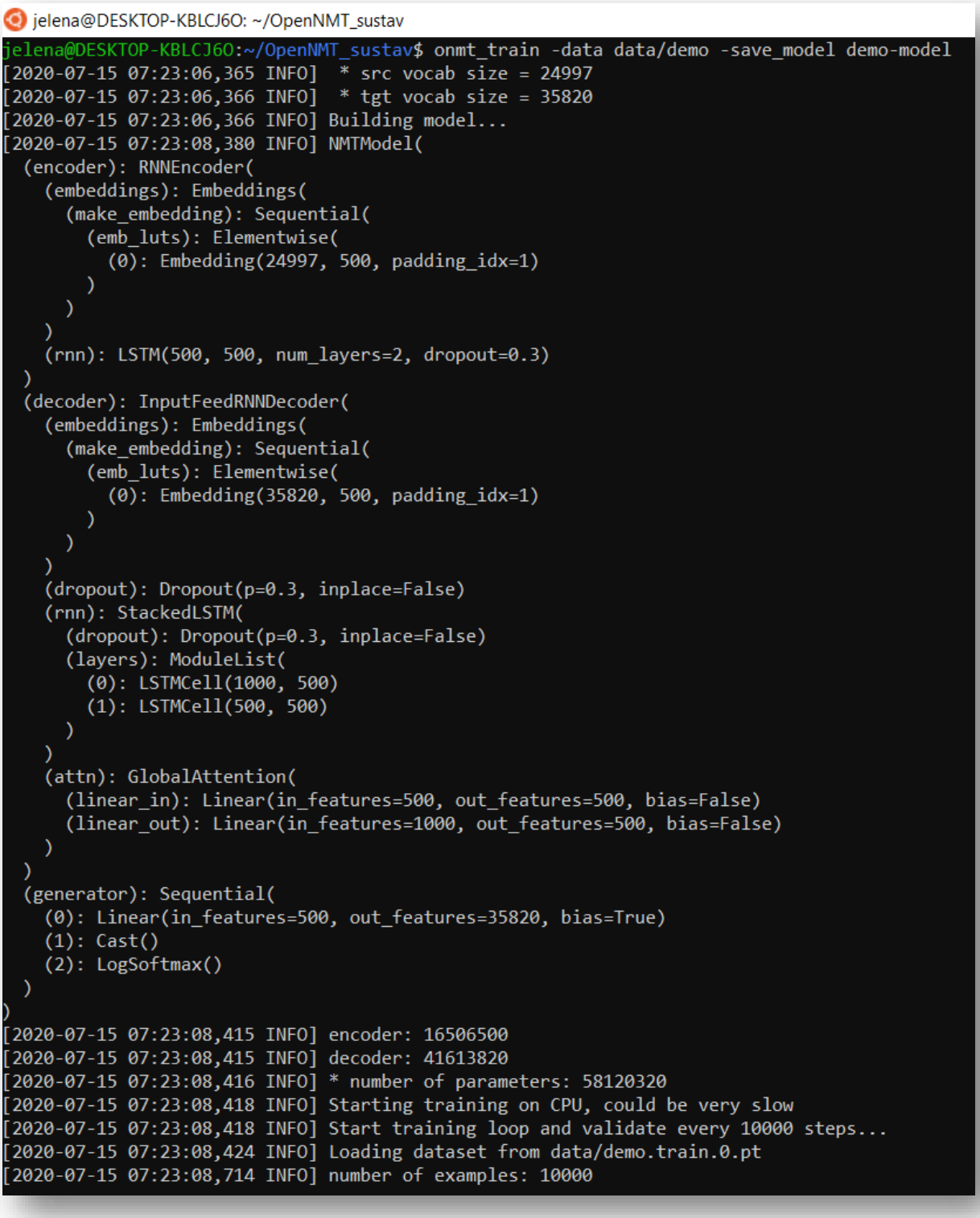
Nakon pripreme i pohrane datoteka na slici 24 je prikazan sljedeći korak, pokretanje pretprocesiranja podataka koje kreira tri različite datoteke koje sadrže podatke za strojno prevođenje (Luong i sur., 2017[124]).



Slika 24. Pretprocesiranje podataka

Nadalje, prebacivanjem i pretprocesiranjem podataka ostvareni su preduvjeti za najdugotrajniji korak u procesu izgradnje sustava, tj. treniranje sustava, kao što je prikazano na slici 25.

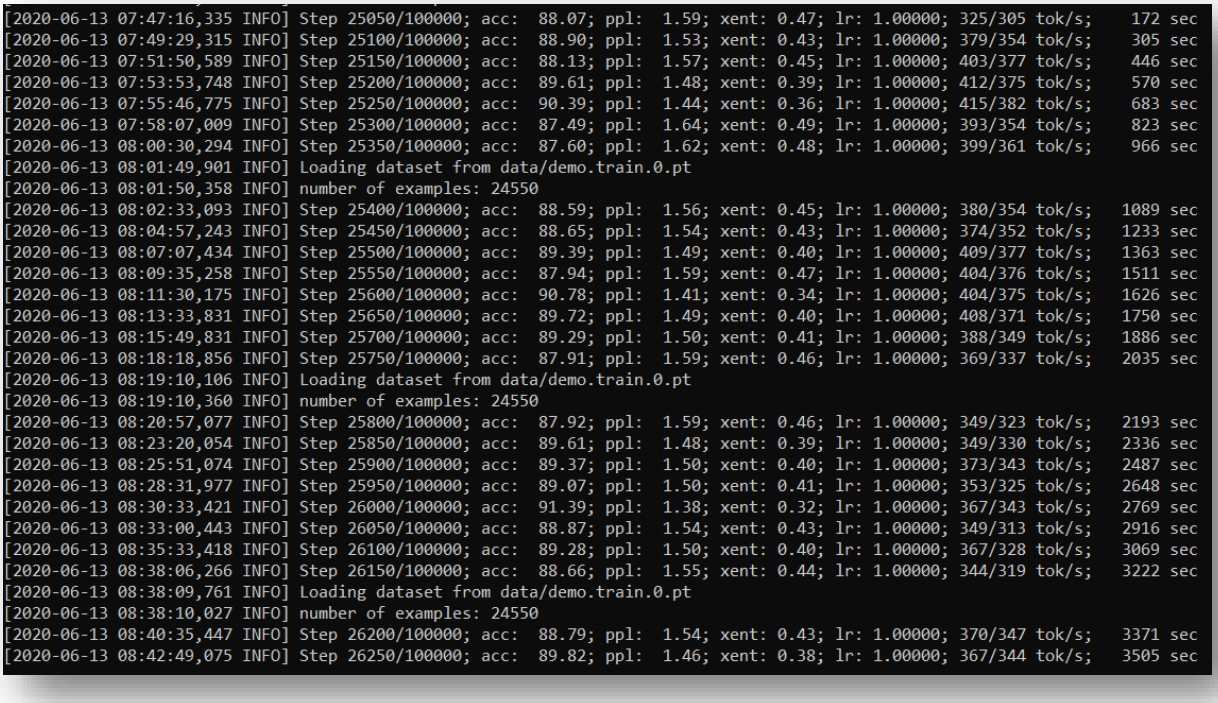
[123] PCChip. (2016). Koja je razlika između APU, CPU i GPU procesora?
[124] Luong, T. i sur. op. cit.



Slika 25. Treniranje sustava

Naredba za treniranje sustava pokreće „data" datoteku i „save" datoteku. Pokretanjem navedene naredbe pokreće se zadani model koji se sastoji od dvoslojne duge kratkoročne memorije s po 500 skrivenih jedinica na enkoder i dekoder strani (Luong i sur., 2017[125]).

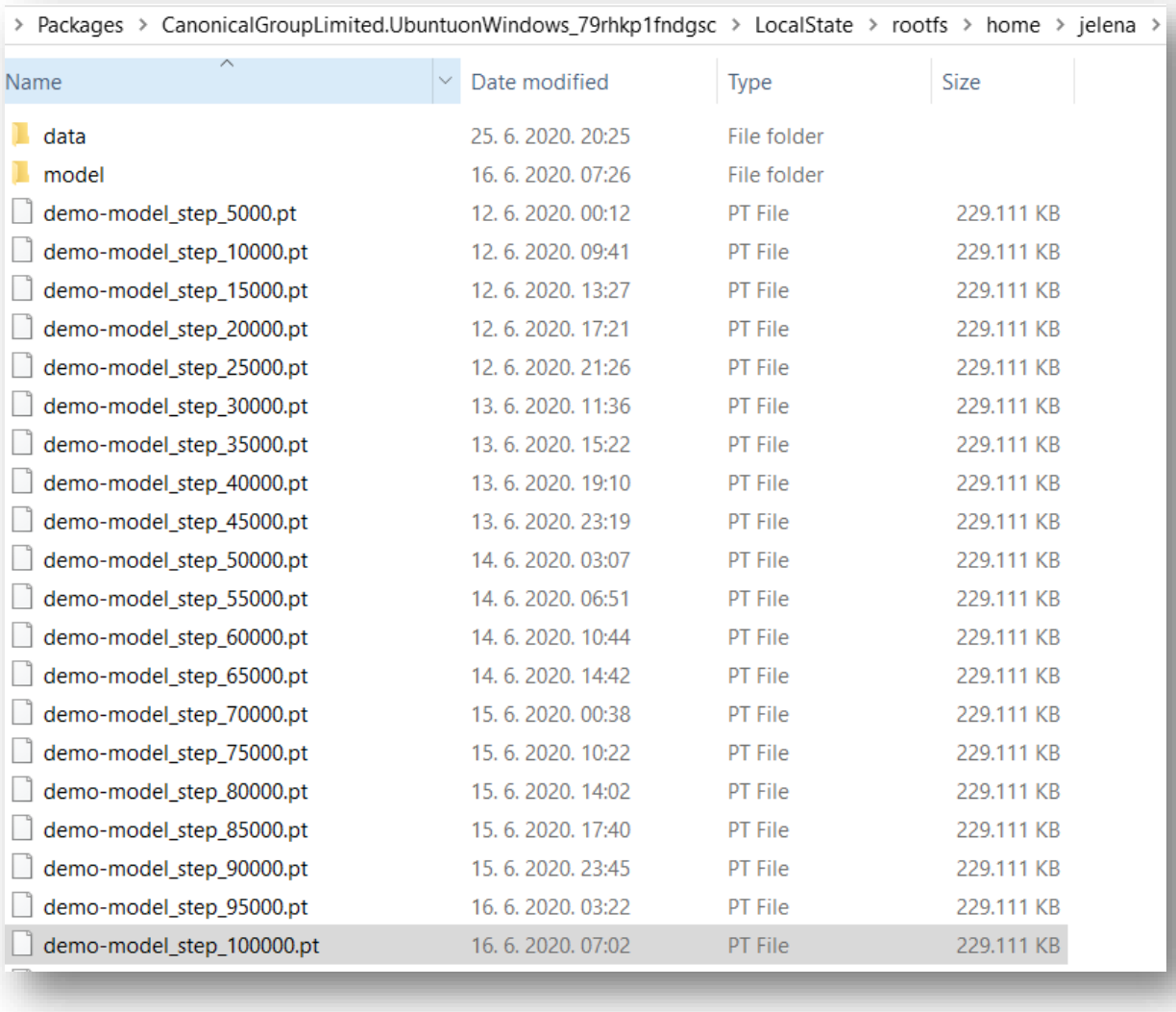
Treniranje sustava sačinjenog od 25.000 paralelnih, tokeniziranih rečenica i 5163 paralelne rečenice za validaciju sustava, trajalo je oko 108 sati, odnosno 4.5 dana. Na slici 26 moguće je pratiti tijek treniranja sustava. Analizom priloženog zaključuje se da je sami sustav proces treniranja podijelio u 100.000 koraka te svaki 50. korak ispisuje liniju s informacijama o vremenu dostizanja novog koraka, broj koraka, točnost (*acc*), složenost (*ppl*), maksimizirana vjerojatnost proučavanih podataka prema modelu (*xent*), stopa učenja (*lr*).



Slika 26. Proces treniranja sustava

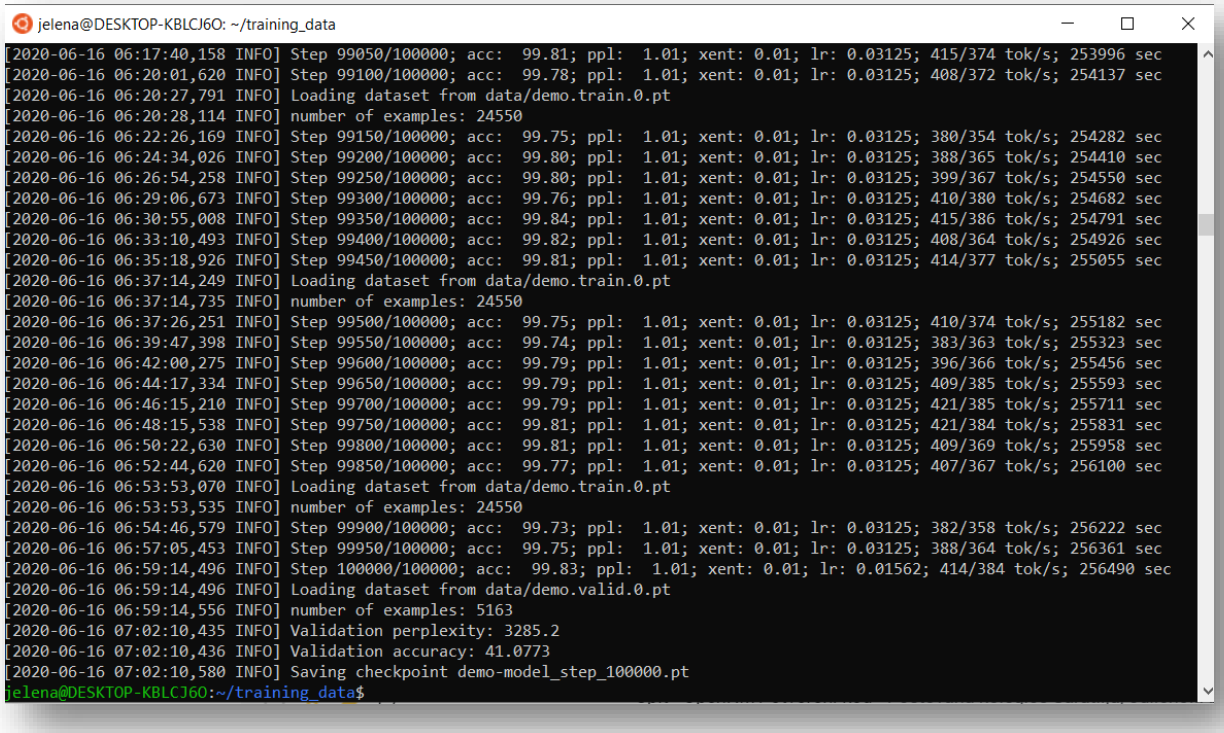
Svakih 5.000 koraka sustav prema zadanom kodu sprema podatak o zadnjem dostignutom koraku. Taj postupak prikazan je na slici 27. U tom slučaju, ako je treniranje sustava bilo prekinuto, pri sljedećem pokretanju treniranja, nastavit će se od zadnjeg spremljenog koraka.

[125] Luong, T. i sur. op. cit.



Slika 27. Spremljene etape treniranog sustava

Kada je treniranje sustava završeno (slika 28), sustav je spreman za prevođenje tekstova.

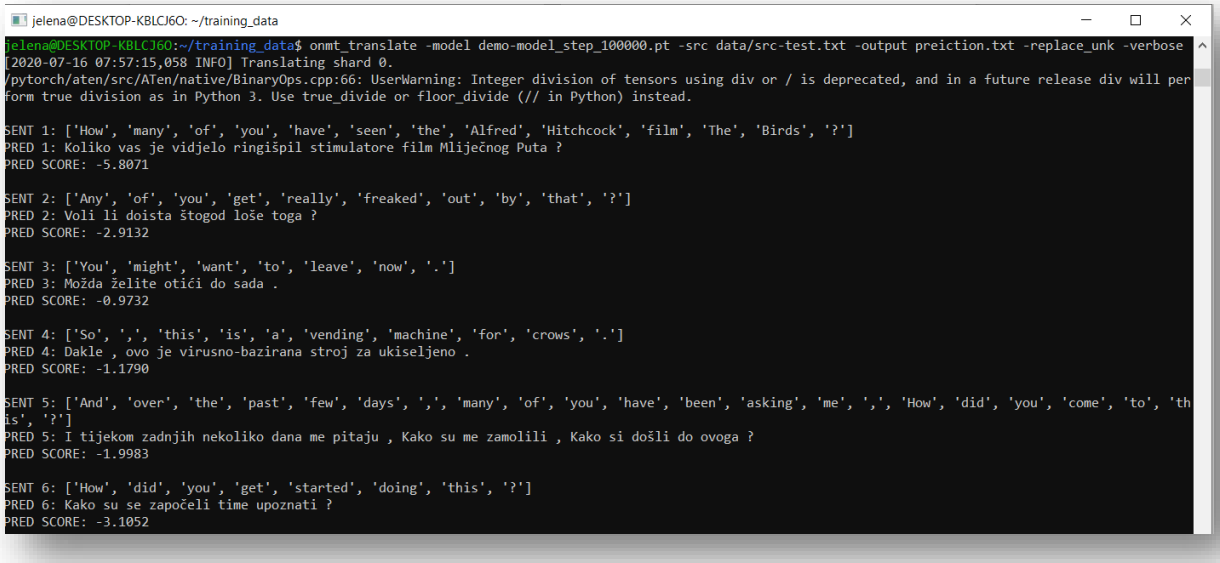


Slika 28. Završetak procesa treniranja sustava

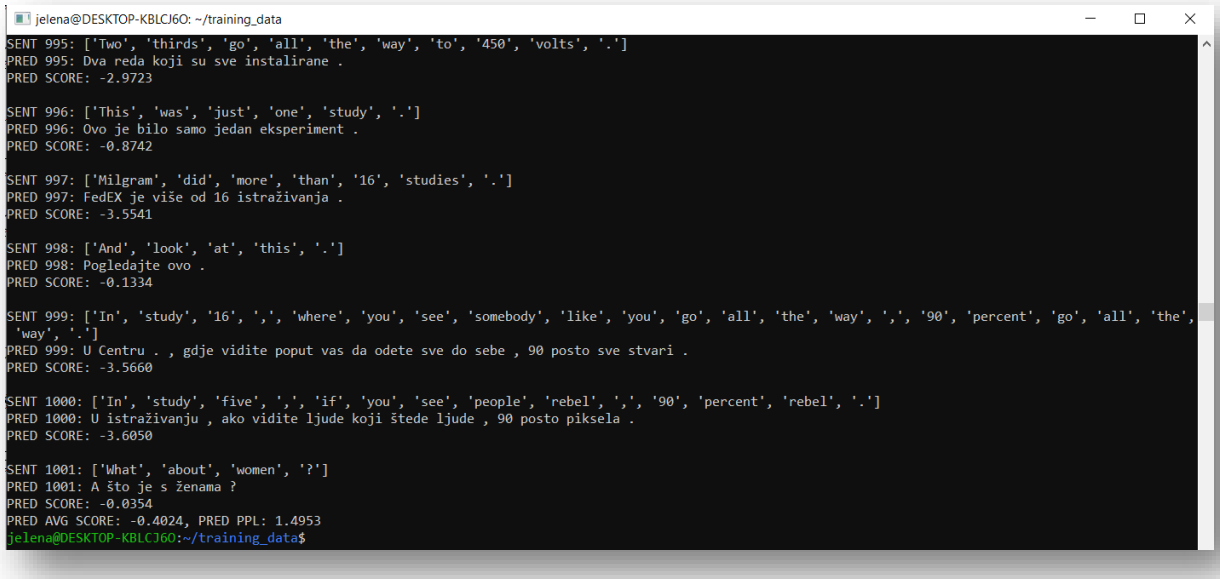
Budući da je sustav treniran na malom setu podataka, nije realno očekivati potpuno točne i smislene prijevode. Potrebni su korpusi s više stotina tisuća ili milijuna usporednih rečenica kako bi se izgradio kvalitetan sustav za neuralno strojno prevođenje.

Kako je već navedeno, cilj ovog istraživanja bio je pokazati način i mogućnost izgradnje sustava za neuralno strojno prevođenje za englesko-hrvatski jezični par. Iz korpusa korištenog za treniranje sustava, TedTalks korpusa, preuzeto je 1001 rečenica koja ne čini dio korpusa za treniranje sustava niti za validaciju, te su spremljene u posebnu datoteku.

Budući da datoteka koja se prevodi sadrži 1001 rečenicu nije ih moguće sve prikazati u radu, te su na slikama u nastavku (slike 29 i 30) moguće vidjeti prvih i zadnjih nekoliko prevedenih rečenica. Svakoj rečenici dodijeljen je određeni "PRED SCORE", a na samom kraju postupka izračunat je "PRED AVG SCORE" koji iznosi -0.4024 I "PRED PPL" koji iznosi 1.4953.



Slika 29. Primjeri prijevoda 1



Slika 30. Primjeri prijevoda 2

3.6. Automatska evaluacija kvalitete strojnog prijevoda

Automatska evaluacija strojnog prijevoda označava mjeru sličnosti strojnog i jednog ili više referentnih ljudskih prijevoda. Najkvalitetnija evaluacija strojnog prijevoda bila bi ona ljudska, međutim takva evaluacija je skupa i dugotrajna. Može trajati mjesecima, a za svaki pojedinačni prijevod potrebno je krenuti ispočetka. Glavne prednosti automatskih metrika za evaluaciju kvalitete strojnog prijevoda u odnosu na ljudsku evaluaciju su objektivnost, brzina i

ponovna iskoristivost (Dunđer, 2015[126]). Neke od poznatih metoda evaluacije su WER (eng. *Word-error rate*), PER (eng. *Position-independent word error rate*), TER (eng. *Translation error/edit rate*), NIST (eng. *National Institute of Standards and Technology*), ROUGE (eng. *Recall-Oriented Understudy for Gisting Evaluation*), GTM (eng. *General Text Matcher*), METEOR (eng. *Metric for Evaluation of Translation with Explicit ORdering*) te BLEU (eng. *BiLingual Evaluation Understudy*) metoda automatske evaluacije koja je korištena za evaluaciju prijevoda u ovom radu.

Neki od primjera automatske evaluacije za hrvatski jezik uključuju različite metrike za različite jezične parove i domene. U istraživanju Seljan i Dunđer (2014[127]) provedena je evaluacija sustava za englesko-hrvatski i i hrvatsko-engleski jezik nakon procesa automatskog prepoznavanja govora i strojnog prevođenja. Evaluacija je provedena primjenom WER i PER metrike i uspoređena s ljudskom evaluacijom. Brkić i sur. (2011[128]) koristili su automatske metrike BLEU, F-mjeru i NIST za evaluaciju online sustava za strojno prevođenje za englesko-hrvatski i hrvatsko-engleski smjer za četiri različite domene (pravna, tehnička, sport i opća) te ih usporedili s ljudskom evaluacijom. U radu Seljan i Dunđer (2015a[129]) analiziran je rusko-hrvatski i englesko-hrvatski strojni prijevod iz opće domene primjenom metrika BLEU, NIST, METEOR i GTM za dva online sustava. U radu Seljan i Dunđer (2015b[130]) provedena je automatska evaluacija iz sociološko-filozofsko-duhovne domene, nastale procesom digitalizacije za hrvatsko-engleski i englesko-hrvatski jezični smjer, pri čemu su korištene BLEU, NIST, METEOR i GTM metrike. U radu Seljan i sur. (2012[131]) provedena je automatska evaluacija primjenom BLEU metode za strojne prijevode iz domene prava, ali primjenom više referentnih skupova. Dunđer i sur. (2020[132]) koristili su automatsku BLEU metriku za evaluaciju strojno prevedenih tekstova poezije za hrvatsko-njemački i njemačko-hrvatski jezični par primjenom metrika BLEU, METEOR, RIBES i CharacTER.

[126] Dunđer, I. op. cit.
[127] Seljan, S.; Dunđer, I. (2014). Combined Automatic Speech Recognition and Machine Translation in Business Correspondence Domain for English-Croatian. Int. Journal of Computer, Information, Systems and Control Engineering, WASET 8 (11), 1069.
[128] Brkić, M.; Seljan, S.; Vičić, T. (2011). Machine translation evaluation for croatian-english and english-croatian language pairs. NLPCS Workshop: Human-Machine Interaction in Translation. Copenhagen: Copenhagen Business School, 93-104.
[129] Seljan, S.; Dunđer, I. (2015). Information Systems and Technologies (CISTI 2015), 1-4.
[130] Seljan, S.; Dunđer, I. (2015). Automatic quality evaluation of machine-translated output in sociological-philosophical-spiritual domain. Information Systems and Technologies (CISTI 2015), 1-4.
[131] Seljan, S.; Brkić, M.; Vičić, T. (2012). BLEU Evaluation of Machine-Translated English-Croatian Legislation.. LREC, 2143-2148.
[132] Dunđer, I.; Seljan, S.; Pavlovski, M. (2020). Automatic Machine Translation of Poetry and a Low-Resource Language Pair, MIPRO.

3.6.1. BLEU metoda evaluacije

Papineni i sur. (2002[133]) predlažu BLEU (*Bilingual evaluation understudy*) – automatsku metodu evaluacije strojnog prijevoda koja je brza, ekonomična, neovisna o jeziku, odnosno primjenjiva na sve jezične parove i visoko korelira s ljudskom evaluacijom. Prema Dunđer (2015[134]) BLEU predstavlja standardnu metriku koja se koristi za evaluaciju strojnog prijevoda, a izračun se temelji na preklapanju n-grama između strojnog prijevoda i jednog ili više referentnih prijevoda, a orijentirana je i prema preciznosti. Potreba za automatskom evaluacijom strojnog prijevoda prvenstveno je rezultat potrebe programera koji razvijaju sustave za strojno prevođenje. Oni su uvidjeli potrebu mogućnosti praćenja efekta dnevnih promjena koje naprave u sustavima kako bi mogli izuzeti loše ideje od dobrih (Papineni i sur., 2002[135]). Vjeruje se kako je napredak sustava za strojno prevođenje omogućen upravo primjenom metoda automatske evaluacije kojom se na vrijeme može uočiti vodi li implementacija neke ideje u krivom smjeru, odnosno hoće li unazaditi kvalitetu sustava. Jednako relevantan primjer bila bi usporedba s projektom razvoja softvera u kojem je testiranje sustava započeto na samom početku. Ukoliko se greške otkrivaju pravovremeno, na samom kraju ćemo imati kvalitetniji sustav s puno manje grešaka i nepredviđenih scenarija korištenja koji bi mogli dovesti do kasnijih problema u korištenju sustava. Međutim, ako se sustav ne testira pravovremeno, postoji veliki rizik da će u krajnjoj aplikaciji biti grešaka koje bi u potpunosti mogle onemogućiti izvršavanje procesa te na čiji će se ispravak utrošiti mnogo više vremena i resursa nego što bi se utrošilo na pravovremeno testiranje.

Glavna ideja koja se nalazi u pozadini koncepta automatske evaluacije strojnog prevođenja jest: *„Što je strojni prijevod bliži profesionalnom, ljudskom prijevodu, to je bolji“* (Papineni i sur., 2002[136]).

Primarni zadatak BLEU metrike je usporediti n-grame kandidata za prijevod s n-gramima referentnog prijevoda te zbrojiti preklapanja koja su neovisna o poziciji u rečenici. Što je veći broj preklapanja, kandidat za prijevod je bolji. Temeljni princip ove metrike je mjera

[133] Papineni K.; Roukos S.; Ward T.; Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pages 311–318
[134] Dunđer, I. op. cit.
[135] Papineni, K. i sur. op. cit.
[136] Ibid.

preciznosti (eng. *precision measure*) koja se računa brojem unigrama u prijevodu kandidatu koji se nalaze i u referentom prijevodu i zatim se dijeli s brojem riječi u prijevodu kandidatu. Međutim, prema ovakvom izračunu, ako se neka riječ iz prijevoda kandidata u njemu pojavljuje više puta, a u referentnom prijevodu ne, sustav bi ju uključio u izračun i greškom bi se dobio nerealno visok rezultat. Prema Papineni i sur. (2002[137]) ovdje je važno naglasiti koncept modificirane unigramske preciznosti prema kojem se jedna riječ iz prijevoda kandidata poništava s riječju u referentnom prijevodu čime se izbjegava prethodno navedeni problem. Razlika između izračuna unigramske i modificirane unigramske preciznosti može se proučiti na primjerima u tablici 1.

Kandidat prijevod:	Ona i pjeva i i i i.
Referentni prijevod:	Ona i pjeva i pleše i svira.
Standardna unigramska preciznost:	8/8
Modificirana unigramska preciznost:	6/8

Tablica 1. Izračun unigramske preciznosti

Modificirana n-gramska preciznost jednako se izračunava za bilo koji *n*-gram. Tako bi za rečenice u gornjem primjeru izračun modificiranih n-gramski preciznosti bio sljedeći:

modificirana bigramska preciznost: 3/7,

modificirana 3-gramska preciznost: 2/6,

modificirana 4-gramska preciznost: 1/5.

Modificirana unigramska preciznost utječe na adekvatnost prijevoda, a n-gramska preciznost na fluentnost (Seljan, Dunđer, 2015a[138]). Prema Dunđer (2015[139]) adekvatnost označava mjeru u kojoj je značenje u ciljnom jeziku sačuvano u potpunosti ili mjeru u kojoj je dio informacija izgubljen, izmijenjen ili dodan, a fluentnost vrednuje gramatički oblik rečenice, izbor riječi i pridržavanje jezičnih pravila.

[137] Papineni, K. i sur. op. cit.
[138] Dunđer, I. op. cit.
[139] Ibid.

Među nedostacima BLEU metode Koehn (2010[140]) navodi kao problem što se ne uzima u obzir relativna relevantnost riječi, sveukupna gramatičku koherentnost, neintuitivnost te oslanjanje na cjelokupni testni set podataka. Ova metrika prepoznaje jedino potpuna podudaranja, a ignorira riječi koje nisu u istom obliku riječi, odnosno u istom broju, rodu, deklinacijskom ili konjugacijskom obliku.

Nadalje, prema Seljan i Dunđer (2015a[141]) BLEU isto tako dodjeljuje i kaznu za kratkoću automatskim prijevodima koji su kraći od referentnih prijevoda. Rečenica kandidat prijevoda ne bi smjela biti niti preduga niti prekratka u odnosu na referentni prijevod. One rečenice duže od referentnih prijevoda su zapravo već kažnjene mjerom modificirane n-gramske preciznosti, no za one kraće uveden je novi koncept, multiplikativni kazneni faktor kratkoće (eng. *multipliative brevity penalty factor*). Prema Papineni i sur. (2002[142]) idealna vrijednost faktora kratkoće je 1.0 što znači da je duljina rečenice kandidata prijevoda ista kao duljina referentnog prijevoda.

Kako bi se dobio izračun BLEU vrijednosti, uzima se geometrijska sredina svih izračunatih vrijednosti modificirane n-gramske preciznosti te se taj rezultat množi s eksponencijalnim kaznenim faktorom kratkoće.

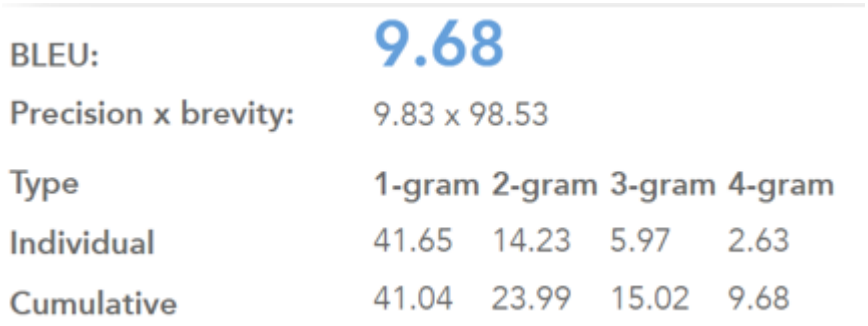
Prema Dunđer (2015[143]) BLEU vrijednosti kreću se od 0 koja označava da ne postoje preklapanja riječi u strojnom i referentnom prijevodu do vrijednosti 1 koja označava potpuno podudaranje strojnog i referentnog prijevoda. BLEU vrijednost veće od 0.3 u pravilu odražavaju razumljive strojne prijevode, a BLEU rezultat veći od 0.5 upućuje na dobre strojne prijevode (Dunđer, 2015[144], Lavie i sur., 2010[145]).

3.6.2. Analiza BLEU rezultata

Tekst preveden u izgrađenom OpenNMT sustavu za strojno prevođenje evaluiran je BLEU metodom automatske evaluacije. Ukoliko se želi usporediti BLEU rezultat dvaju sustava za strojno prevođenje, moguće je dodati još jednu datoteku koja je rezultat strojnog prijevoda istog izvornog teksta. U ovom radu cilj istraživanja je bio analizirati rezultat treniranog sustava.

[140] Koehn, P. op. cit.
[141] Seljan, S. i Dunđer, I. (2015a). op. cit.
[142] Papineni, K. i sur. op. cit.
[143] Dunđer, I. op. cit.
[144] Ibid.
[145] Lavie, A. i sur. op. cit.

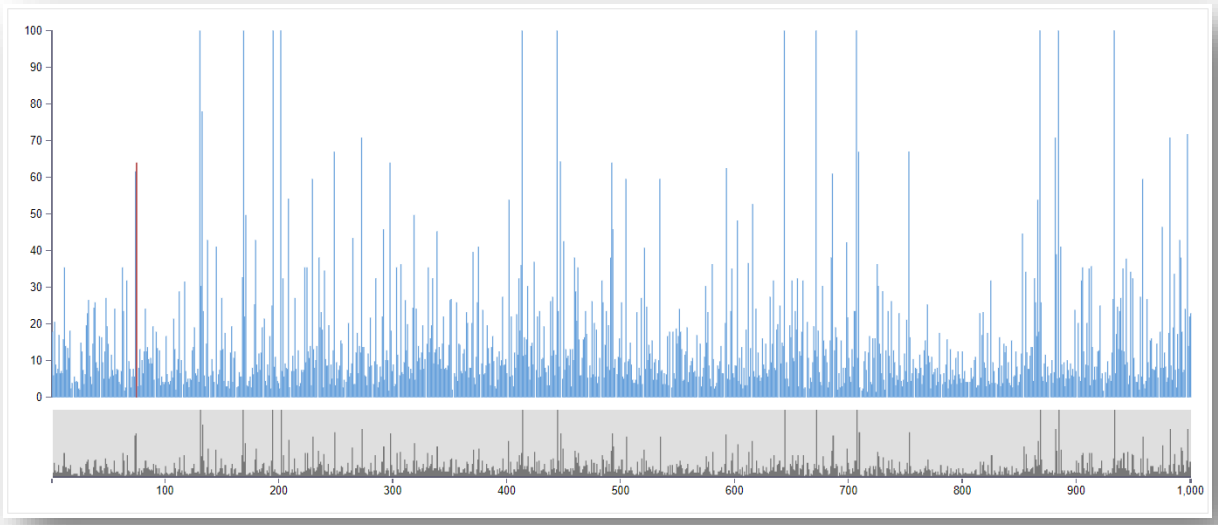
Jedan od čimbenika koji može utjecati na realniji i bolji BLEU rezultat sustava za strojno prevođenje jest mogućnost tokenizacije i zapisa malim slovima.



Slika 31. Rezultat evaluacije sustava na temelju analiziranih prijevoda

BLEU se prilikom evaluacije kvalitete strojnog prijevoda u pravilu računa na razini cijelog korpusa. Odnosno, prema Dunđer (2015)[146] i Madnani (2011)[147] pri izračunu BLEU-a na razini cijelog testnog skupa istovremeno se uzimaju sve rečenice i računaju preklapanja n-grama preko svih rečenica pa iz toga proizlazi da BLEU na razini korpusa ne odgovara aritmetičkoj sredini na razini rečenica. S obzirom na to i na činjenicu da evaluirani prijevod sadrži 1001 rečenicu, istovremeno su uzete sve rečenice i izračunata preklapanja n-grama preko svih rečenica. Zatim je izračunata preciznost i kratkoća te su vrijednosti pomnožene da bi se dobio BLEU rezultat evaluacije sustava prikazan na slici 31. Slika 32 daje grafički prikaz postignutih BLEU rezultata za svaku rečenicu na razini cijelog korpusa od 1001 rečenice.

[146] Dunđer, I. op. cit.
[147] Madnani, N. (2011). iBLEU Interactively Debugging and Scoring Statistical Machine Translation Systems. 2011 Fifth IEEE International Conference on Semantic Computing (ICSC), pp. 213-214.



Slika 32. Grafički prikaz BLEU rezultata evaluacije nad cijelom datotekom

Analizom grafa sa slike 32, utvrđeno je da je velika većina rečenica postigla BLEU rezultat do 30.0, dok je 12 rečenica postiglo maksimalni BLEU rezultat 100.0, a dvije rečenice imale su BLEU niži od 1.0. U tablici 2 u nastavku prikazane su odabrane rečenice s maksimalnim i minimalnim vrijednostima.

Br.	

1.

Sentence 131	BLEU	Length ratio	
Source	-	-	thanks very much .
Human	100.00	1.00	hvala vam puno .
Machine	100.00	1.00	hvala vam puno .

2.

Sentence 445	BLEU	Length ratio	
Source	-	-	it was hard .
Human	100.00	1.00	bilo je teško .
Machine	100.00	1.00	bilo je teško .

3.

Sentence 644	BLEU	Length ratio	
Source	-	-	you know all about big brains .
Human	100.00	1.00	znate sve o velikim mozgovima .
Machine	100.00	1.00	znate sve o velikim mozgovima .

4.

Sentence 885	BLEU	Length ratio	
Source	-	-	and that tells us several things .
Human	100.00	1.00	a to nam govori nekoliko stvari .
Machine	100.00	1.00	a to nam govori nekoliko stvari .

5.

Sentence 934	BLEU	Length ratio	
Source	-	-	there are three ways .
Human	100.00	1.00	postoje tri načina .
Machine	100.00	1.00	postoje tri načina .

6.

Sentence 144	BLEU	Length ratio	Text
Source	-	-	since every scientist in the world now believes this , and even president bush has seen the light , or pretends to , we can take this is a given .
Human	100.00	1.00	budući da sada svaki znanstvenik na svijetu u to vjeruje , i čak je i predjednik bush vidio svjetlo , ili se samo pravi , možemo to uzeti zdravo za gotovo .
Machine	0.13	0.19	od svakog je jučer sada ?

7.

Sentence 483	BLEU	Length ratio	Text
Source	-	-	audience : no . susan blackmore : someone says no , very loudly , from over there .
Human	100.00	1.00	publika : ne . netko je rekao ne .. pa , ja kažem da , i da postoji , dala bih nagradu darwinu .
Machine	0.06	0.13	kad ne .

Tablica 2. Rečenice s maksimalnim i minimalnim BLEU rezultatom

S obzirom na veličinu korpusa, u radu nije analizirana svaka rečenica zasebno. U nastavku će prema kriterijima koje koristi Ljubas (2017[148]), a definiraju (Kučiš i sur., 2009[149], Simeon, 2010[150], Seljan i sur., 2011[151], Seljan i sur., 2012[152], Brkić i sur., 2013[153], Kučiš i Seljan, 2014[154], Seljan i sur., 2015[155]) u svom radu, biti analizirane 3 rečenice:

- neprevedene riječi (riječi koje su zadržane na izvornom jeziku),
- izostavljene riječi,
- umetnute riječi (riječi iz ciljnog jezika koje su nepravedno umetnute),
- leksičke pogreške (riječi koje se semantički znatno razlikuju od riječi u izvornome tekstu),

[148] Ljubas, S. (2017). Analiza pogrešaka u strojnim prijevodima sa švedskog na hrvatski. Hieronymus 4, pp. 28-64.
[149] Kučiš, V.; Seljan, S.; Klasnić, K. (2009). Evaluation of electronic translation tools through quality parameters. INFuture 2009: Digital Resources and Knowledge Sharing, 341-351
[150] Simeon, I. (2008). Vrednovanje strojnoga prevođenja. Doktorska disertacija. Zagreb: Filozofski fakultet.
[151] Seljan, S.; Brkić, M.; Kučiš, V. (2011). Evaluation of free online machine translations for Croatian-English and English-Croatian language pairs. INFuture2011: Information Sciences and e-Society, 331-344.
[152] Seljan, S.; Brkić, M.; Vičić, T. (2012). BLEU Evaluation of Machine-Translated English-Croatian Legislation. LREC, 2143-2148.
[153] Brkić, M.; Seljan, S.; Vičić, T. (2012). Automatic and human evaluation on english-croatian legislative test set. Intelligent Text Processing and Computational Linguistics, Springer, 311-317.
[154] Kučiš, V.; Seljan, S. (2014). The role of online translation tools in language education. Babel 60 (3).
[155] Seljan, S.; Tucaković, M.; Dunđer, I. (2015). Human evaluation of online machine translation services for english/russian-croatian. New Contributions in Information Systems and Technologies, 1089-1098.

- pravopisne pogreške (uporaba velikih i malih slova u ovom slučaju neće biti evaluirana jer su u cilju postizanja boljeg rezultata datoteke pripremljene na način da su svi znakovi zapisani malim slovima),
- morfosintaktičke pogreške (pogrešni oblici, nesročnost između subjekta i predikata, pogreške u članovima, pogrešno odabrane funkcionalne riječi itd.),
- stilističke pogreške (nespretno sročene sintagme, prijevodni ekvivalenti koji su semantički bliski polaznoj riječi, ali mogu dovesti do nesporazuma ili zbuniti čitatelja),
- pogreške u redu riječi.

Primjer 1:

Konzola

SENT 9: ['And', 'he', 'was', 'telling', 'me', 'that', 'really', ',', 'we', 'ought', 'to', 'try', 'and', 'eradicate', 'these', 'things', '.']

PRED 9: I on mi je pričao kako bismo trebali pokušati i preokrenuti te stvari .

PRED SCORE: -3.5422

Sentence 9	BLEU	Length ratio	Text
Source	-	-	and he was telling me that really , we ought to try and eradicate these things .
Human	100.00	1.00	i rekao mi je da bismo morali probati istrijebiti tu vrstu .
Machine	7.14	1.17	i on mi je pričao kako bismo trebali pokušati i preokrenuti te stvari .

Analiza

U ovom strojnom prijevodu mjera modificirane unigramske preciznosti iznosi 4/14, a bigramske 1/13. Strojni prijevod (14 tokena) duži je od referentnog prijevoda (12 tokena). U odnosu na referentni prijevod, u strojnom prijevodu ne pronalaze se neprevedene, izostavljene riječi i pravopisne greške. U strojnom prijevodu umetnut je veznik „i“ između glagola „pokušati“ i „preokrenuti“ te zamjenica „on“ koja ne postoju u referentnom prijevodu jer je po prirodi hrvatskog jezika moguće zaključiti o kojem se licu i broju radi prema obliku u kojem se glagol nalazi. Nadalje, u strojnom prijevodu pronalaze se leksičke greške poput glagola „preokrenuti“ koji je korišten umjesto „istrijebiti“ što pridonosi nejasnoći prijevoda. Morfosintaktički gledano, ne postoje nelogičnosti u vidu neslaganja subjekta i predikata niti su korišteni pogrešni oblici, no postoje stilističke pogreške koje pridonose nerazumljivosti prijevoda. Formulacija rečenice „i on mi je pričao kako bismo…“ gdje je umetnuta zamjenica „on“

	naglasak stavlja na činjenicu da *mu je još jedna osoba* rekla istu stvar, dok u referentnom prijevodu naglasak nije stavljen na tu činjenicu. Nadalje, u strojnom prijevodu je korišten veznik namjerne rečenice „kako“ dok je u referentnom prijevodu korišten veznik „da“. Pronađen je i prijevodni ekvivalent, sinonim, koji ne narušava shvaćanje rečenice kao što je glagol „pokušati“ umjesto „probati“, ali isto tako postoje i prijevodni ekvivalenti koji utječu na krivo shvaćanje rečenice. Korišteni su glagol „trebali“ umjesto „morali“, „preokrenuti“ umjesto „istrijebiti“ i imenica „stvari“ umjesto „vrstu“. Međutim, proučavanjem rečenice izvornog jezika, može se utvrditi da umetnuta riječ „on“ te imenica „stvari“ uistinu postoje u izvorniku, ali su prema ljudskom prijevodu zbog suvišnosti i boljeg shvaćanja značenja rečenice izbačene („on“) i zamijenjene („vrstu“ umjesto „stvari“).

Tablica 3. Analiza rezultata 1

Primjer 2:

Konzola	**SENT 17: ['But', 'part', 'of', 'the', 'reason', 'that', 'I', 'found', 'this', 'interesting', 'is', 'that', 'I', 'started', 'noticing', 'that', 'we', 'are', 'very', 'aware', 'of', 'all', 'the', 'species', 'that', 'are', 'going', 'extinct', 'on', 'the', 'planet', 'as', 'a', 'result', 'of', 'human', 'habitation', 'expansion', ',', 'and', 'no', 'one', 'seems', 'to', 'be', 'paying', 'attention', 'to', 'all', 'the', 'species', 'that', 'are', 'actually', 'living', '--', 'that', 'are', 'surviving', '.']** **PRED 17: Ali dio razloga zbog kojih sam pronašao to zanimljivo je da sam počeo primjećivati kako smo vrlo svjesni svih vrsta koji će izumrijeti na planetu kao rezultat ljudske pomisli , i nitko ne polaže djeci svim vrstama vode koji su ustvari vruće -- to upravljaju prikaz .** **PRED SCORE: -14.0774**

Sentence 17	BLEU	Length ratio	Text
Source	-	-	but part of the reason that i found this interesting is that i started noticing that we are very aware of all the species that are going extinct on the planet as a result of human habitation expansion , and no one seems to be paying attention to all the species that are actually living -- that are surviving .
Human	100.00	1.00	razlog zašto mi je to bilo zanimljivo je dijelom to što sam počeo primjećivati da smo poprilično svjesni svih vrsta koje će izumrijeti na planetu kao rezultat proširivanja ljudskog prebivališta , a nitko ne obraća pažnju na sve vrste koji zapravo žive – koje preživljavaju .
Machine	18.06	1.02	ali dio razloga zbog kojih sam pronašao to zanimljivo je da sam počeo primjećivati kako smo vrlo svjesni svih vrsta koji će izumrijeti na planetu kao rezultat ljudske pomisli , i nitko ne polaže djeci svim vrstama vode koji su ustvari vruće -- to upravljaju prikaz .

Analiza	U ovom strojnom prijevodu mjera modificirane unigramske preciznosti iznosi 21/47, a bigramske 11/46, 3-gramske 6/45 i 4-gramske 3/44. Strojni prijevod (47

	tokena) duži je od referentnog prijevoda (46 tokena). U odnosu na referentni prijevod, u strojnom prijevodu ne pronalaze se neprevedene riječi. Od pravopisnih grešaka, pogrešno je korištena engleska inačica spojnice „--„ umjesto „-„. Nadalje, pojavljuje se mnogo umetnutih riječi. Sama rečenica strojnog prijevoda započinje riječima „ali dio razloga zbog kojih …“ koje se ne pojavljuju u referentnom prijevodu, štoviše, strojni prijevod započinje suprotnim veznikom „ali“ što prema hrvatskom pravopisu nije moguće. Tu se uočava i prva morfosintaktička greška jer je riječ „razloga“ interpretirana u množini, a ne kao imenica deklinirana u genitivu te zbog toga dolazi do krivog slaganja s riječju „kojih“. U daljnjem prijevodu može se uočiti da je kao veznik korištena riječ „da“ umjesto „što“ te „kako smo“ umjesto „da smo“ što ne narušava znatno shvaćanje rečenice no svakako utječe na modificiranu n-gramsku preciznost. Važno je istaknuti kako je u jednoj ovako dugačkoj rečenici pronađen i jedan 10-gram: „svjesni svih vrsta koji će izumrijeti na planetu kao rezultat“. Ovaj niz prekinut je zančajnom leksičkom pogreškom gdje je izraz iz referentnog prijevoda „proširivanja ljudskog prebivališta“ interpretiran kao „ljudske pomisli“ u strojnom prijevodu. Strojni prijevod ostatka rečenice („i nitko ne polaže djeci svim vrstama vode koji su ustvari vruće – to upravljaju prikaz“) obiluje stilističkim, leksičkim i morfosintaktičkim (npr. „upravljaju prikaz“) greškama te je značenje potpuno krivo preneseno. Umetnute su riječi koje ne postoje u referentnom prijevodu, kao „polaže“, „djeci“, „vode“, „upravljaju“, „prikaz“, što je dovelo do potpunog gubitka informacija iz posljednjeg dijela rečenice.

Tablica 4. Analiza rezultata 2

Primjer 3:

Konzola	**SENT 593: ['So', ',', 'this', 'is', 'a', 'view', 'of', 'what', 'humans', 'are', '.']** **PRED 593: Ovo je pogled na ono što ljudi .** **PRED SCORE: -1.9607**

	Sentence 593	BLEU	Length ratio	
	Source	-	-	so , this is a view of what humans are .
	Human	100.00	1.00	to je pogled na ono što ljudi jesu .
	Machine	62.40	0.89	ovo je pogled na ono što ljudi .
Analiza	U posljednjem analiziranom primjeru strojnog prijevoda mjera modificirane unigramske preciznosti iznosi 7/8, bigramske 5/7, 3-gramske 4/6 i 4-gramske 3/5. Strojni prijevod (8 tokena) kraći je od referentng prijevoda (9 tokena). U odnosu na referentni prijevod, u strojnom prijevodu ne postoje neprevedene niti umetnute riječi te je red riječi u rečenici točan. Međutim, u odnosu na referentni prijevod, izostavljen je glagol „jesu“ što narušava razumijevanje rečenice te ona djeluje kao nedovršena. Osim navedene stilističke greške, riječ „to“ kojom započinje rečenica u strojnom prijevodu zamijenjena je riječju „ovo“ što pak ne narušava shvaćanje rečenice strojnog prijevoda.			

Tablica 5. Analiza rezultata 3

Na temelju provedenog istraživanja i analizom rezultata prikazanih u tablicama 3, 4, i 5, može se zaključiti da je postojanje sustava za neuralno strojno prevođenje, unaprjeđivanje metrika za evaluaciju te ulaganje u njihov razvoj i istraživanje od važnosti za komunikaciju u današnjem dobu. Ljudi žele imati pristup informacijama iz bilo kojeg kraja svijeta, u bilo kojem trenutku na materinskom jeziku. S obzirom na količinu informacija koja se proizvede u svakom trenutku, automatizacija prevođenja dobiva sve veći značaj. Neuralni pristup strojnom prevođenju već je uzeo maha, te je implementiran u sustave svjetski poznatih i najznačajnijih tvrtki u industriji poput Google-a, Microsoft-a, IBM-a i slično. Analizom rezultata ovoga istraživanja može se zaključiti da je za izgradnju sustava za neuralno strojno prevođenje važno osigurati računalo adekvatne snage koje će moći procesuirati milijune podataka te prikupiti što veći i kvalitetniji korpus koji sadrži setove paralelnih rečenica za pojedini jezični par. Ovim istraživanjem prikazan je način pripreme i primjene jednog sustava za neuralno strojno prevođenje korištenjem korpusa od 25.000 paralelnih rečenica za englesko-hrvatski jezični par. Sustav je testiran na setu od 1001 rečenice te je BLEU metodom automatske evaluacije izračunato da je sustav postigao rezultat od 9.68. Uzimajući u obzir kako je za izgradnju kvalitetnog sustava potrebno koristiti korpuse od stotina tisuća, ili nekoliko milijuna paralelnih

rečenica, bilo je realno za očekivati kako se treniranjem sustava korpusom od 25.000 rečenica korištenim u ovom projektu neće postići visoka kvaliteta prijevoda. Neuralni pristup strojnom prevođenju zasigurno može ostvariti kvalitetne prijevode, no i ono trenutno ima nedostataka. Prema Luong i sur. (2016[156]) potrebno je proširiti pokrivenost vokabulara sustava, ostvariti bolji prijevod dugih rečenica, omogućiti jednako dobre prijevode raznih jezičnih varijanti i koristiti više izvora podataka.

[156] Luong, T. i sur. op. cit.

4. Zaključak

Neuralno strojno prevođenje predstavlja najnoviji pristup strojnom prevođenju čijom se implementacijom u određenim segmentima postižu bolji prijevodi u odnosu na one postignute pristupima koji su mu prethodili, poput statističkog strojnog prevođenja i strojnog prevođenja temeljenog na pravilima, iako i statističko strojno prevođenje ima svojih prednosti. Ovaj pristup postao je široko primjenjiva metoda strojnog prevođenja jednako kao i efikasan pristup drugim zadacima koji koriste obradu prirodnog jezika poput dialoga, parsiranja i sažimanja.

U ovom radu korištena je OpenNMTplatforma. Pokretanjem te implementacijom i pripremom potrebnih podataka izgrađen je sustav za neuralno strojno prevođenje s engleskog na hrvatski jezik, gdje je korišteno po 25.000 segmenata englesko-hrvatskih parova i za testiranje sustava set od 1001 rečenice. Evaluacija je provedena automatskom BLEU metrikom te je provedena ručna analiza nad nekoliko odabranih rečenica. S obzirom da je primjena NMT-a relativno nova metoda izgradnje sustava za strojno prevođenje, postoji još mnogo neistraženih aspekata i načina na koje bi se ovakvi sustavi mogli još više unaprijediti.

U današnjim vremenima sasvim je normalno u bilo kojem trenutku moći saznati najnovije vijesti iz svakog kraja svijeta, bez obzira na jezik kojim se govori. Razmisli li se o veličini planeta, broju zemalja, stanovnika, raznovrsnosti jezika te o potrebi svakog pojedinca da u bilo kojem trentku dođe do određene informacije, jasno je kako je potreba za kvalitetnim sustavima za strojno prevođenje koji mogu obraditi velike količine podataka od neizmjerne važnosti. Budući da, osim strojnog prevođenja, postoji mnogo zadataka koji se oslanjaju na primjenu umjetne inteligencije i dubokog učenja, u narednim je godinama moguće očekivati znatan napredak i pronalazak rješenja za trenutne poteškoće koji će zatim rezultirati mogućnošću izgradnje sve kvalitetnijih sustava.

5. Literatura

1) Bahdanau, D.; Cho, K.; Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Proceedings of the International Conference on Learning Representations (ICLR). Dostupno na: https://arxiv.org/pdf/1409.0473.pdf (Datum pristupa: 5.9.2020.)
2) Božić-Štulić, D. (2017). Semantička segmentacija slika metodama dubokog učenja. Kvalifikacijski ispit. Split: Sveučilište u Splitu, Fakultet elektrotehnike, strojarstva i brodogradnje
3) Bhattacharyya, P. (2015). Machine Translation. CRC PRESS Taylor & Francis Group, xix str (Preface)
4) Brkić, M.; Seljan, S.; Bašić Mikulić, B. (2009). Using translation memory to speed up translation process. INFuture 2009 : Digital resources and knowledge sharing, 353-363
5) Brkić, M.; Seljan, S.; Vičić, T. (2009). Evaluation of the statistical machine translation service for Croatian-English. INFuture 2009: Digital resources and knowledge sharing, 319-322
6) Brkić, M.; Seljan, S.; Vičić, T. (2012). Automatic and human evaluation on english-croatian legislative test set. Intelligent Text Processing and Computational Linguistics, Springer, 311-317
7) Brkić, M.; Seljan, S.; Vičić, T. (2011). Machine translation evaluation for croatian-english and english-croatian language pairs. NLPCS Workshop: Human-Machine Interaction in Translation. Copenhagen: Copenhagen Business School, 93-104.
8) Brkić, M.; Seljan, S.; Vičić, T. (2013). Automatic and human evaluation on english-croatian legislative test set. Intelligent Text Processing and Computational Linguistics, Springer, 311-317
9) Chéragui, M. A. (2012). Theoretical Overview of Machine translation. Proceedings of the 4th International Conference on Web and Information Technologies (ICWIT 2012), pp. 160-169
10) CSA Research. (2017). Zero-Shot Translation is Both More and Less Important than you think. Dostupno na: https://csa-research.com/Insights/ArticleID/90/Zero-Shot-Translation-is-Both-More-and-Less-Important-than-you-think. (Datum pristupa: 20.04.2020.)

11) CSA Research. (2019). Why buy CAT tools when NMT rules? Dostupno na: https://csa-research.com/Insights/ArticleID/604/global-content-translation-volume (Datum pristupa: 21.4.2020.)
12) Čanić, J. (2018). Komparativna analiza strojnog prijevoda sa švedskog na hrvatski jezik. Diplomski rad. Zagreb: Filozofski fakultet
13) DeepAI. (2020). Softmax Function. Dostupno na: https://deepai.org/machine-learning-glossary-and-terms/softmax-layer. (Datum pristupa: 25.7.2020.)
14) Dillinger, M.; Marciano, J. (2012). Introduction to MT. The Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2012)
15) Dovedan, Z.; Seljan, S.; Vučković, K. (2002). Strojno prevođenje kao pomoć u procesu komunikacije. Informatologija 35 (4), 283-291
16) Dunđer, I. (2020). Machine Translation System for the Industry Domain and Croatian Language. Journal of information and organizational sciences. 44. 33-50
17) Dunđer, I. (2015). Sustav za statističko strojno prevođenje i računalna adaptacija domene. Doktorska disertacija. Sveučilište u Zagrebu
18) Dunđer, I.; Seljan, S.; Pavlovski, M. (2020). Automatic Machine Translation of Poetry and a Low-Resource Language Pair. MIPRO 2020
19) Google AI Blog. (2016). Zero-Shot Translation with Google's Multilingual Neural Machine Translation System. Dostupno na: https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html (Datum pristupa: 23.7.2020.)
20) Hutchins, J.; Somers, H. L. (1992). An introduction to machine translation. London: Academic Press. Dostupno na: http://www.hutchinsweb.me.uk/IntroMT-TOC.htm (Datum pristupa: 25.7.2020.)
21) Hutchins, J. (1995). Concise history of the language sciences: from the Sumerians to the cognitivists. U E.F.K. Koerner i R.E. Asher, (431-445), Pergamon Press, Oxford. Dostupno na: http://hutchinsweb.me.uk/ConcHistoryLangSci-1995.pdf (Datum pristupa: 25.7.2020.)
22) Hutchins, J. (2001). Machine translation over fifty years. Histoire, Epistémologie, Langage: Le traitement automatique des langues, vol. 23, no. 1, pp. 7-31. Dostupno na: http://www.hutchinsweb.me.uk/HEL-2001.pdf (Datum pristupa: 25.7.2020.)

23) Hutchins, J. (2004). Two precursors of machine translation: Artsrouni and Trojanskij. International Journal of Translation, vol. 16, no. 1, pp. 11-31. Dostupno na: http://www.hutchinsweb.me.uk/IJT-2004.pdf (Datum pristupa: 25.7.2020.)
24) Jaworski, R.; Seljan, S.; Dunđer, I. (2017). Towards educating and motivating the crowd – a crowdsourcing platform for harvesting the fruits of NLP students' labour. Human Language Technologies as a Challenge for Computer Science and Linguistics 1, 332-336
25) Johnson, M.; Schuster, M., Le, Q. V.; Krikun, M.; Wu, Y.; Chen, Z., … & Hughes, M. (2016). Google's multilingual neural machine translation system: enabling zero-shot translation. arXiv preprint arXiv:1611.04558. Dostupno na: https://arxiv.org/pdf/1611.04558.pdf (Datum pristupa: 5.9.2020.)
26) Klein G.; Kim Y.; Deng Y.; Nguyen V.; Senellart J.; Rush A. M. (2018). OpenNMT: Neural Machine Translation Toolkit. arXiv:1805.11462 [cs.CL]. Dostupno na: https://arxiv.org/pdf/1805.11462.pdf (Datum pristupa: 10.6.2020.)
27) Koehn, P. (2010), Statistical Machine Translation. Cambridge: University Press.
28) Koehn, P.; Haddow, B. (2012). Interpolated backoff for factored translation models. Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA). Dostupno na: http://www.mt-archive.info/AMTA-2012-Koehn.pdf (Datum pristupa: 10.6.2020.)
29) Kranjčić, D. (2016). Analiza sustava za evaluaciju strojnih prijevoda primjenom Dinamičkog okvira i Višedimenzionalne metrike. Diplomski rad. Zagreb: Filozofski fakultet. Odsjek za informacijske i komunikacijske znanosti
30) Kučiš, V.; Seljan, S. (2014). The role of online translation tools in language education. Babel 60 (3)
31) Kučiš, V.; Seljan, S.; Klasnić, K. (2009). Evaluation of electronic translation tools through quality parameters. INFuture 2009: Digital Resources and Knowledge Sharing, 341-351
32) Kunchukuttan, A. (2018). An introduction to Machine Translation. Center for Indian Language Technology, Indian Institute of Technology Bombay. Ninth IIT-H Advanced Summer School on NLP, 27th June 2018. Prezentacija. Dostupno na: https://slideplayer.com/slide/15632076/ (Datum pristupa: 8.6.2020.)
33) Lavie, A. (2010). Evaluating the Output of Machine Translation Systems. The Ninth Conference of the Association for Machine Translation in the Americas (AMTA

2010), tutorial documentation, p. 86. Dostupno na: http://www.mt-archive.info/AMTA-2010-Lavie.pdf (Datum pristupa: 20.7.2020.)

34) Luong, T.; Cho, K.; Manning, C. (2016). Neural machine translation. NYU, Stanford. Prezentacija. Dostupno na: https://nlp.stanford.edu/projects/nmt/Luong-Cho-Manning-NMT-ACL2016-v4.pdf (Datum pristupa: 25.5.2020.)

35) Luong, T.; Brevdo, E.; Zhao, R. (2017). Neural Machine Translation (seq2seq Tutorial). Dostupno na: https://github.com/tensorflow/nmt (Datum pristupa: 25.5.2020.)

36) Luong, T.; Pham, H.; Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1412–1421

37) Ljubas, S. (2017). Analiza pogrešaka u strojnim prijevodima sa švedskog na hrvatski. Hieronymus 4, pp. 28-64. Dostupno na: http://www.ffzg.unizg.hr/hieronymus/wp-content/uploads/2018/01/H4-2017_2_Ljubas.pdf (Datum pristupa: 13.7.2020.)

38) Madnani, N. (2011). iBLEU Interactively Debugging and Scoring Statistical Machine Translation Systems. 2011 Fifth IEEE International Conference on Semantic Computing (ICSC), pp. 213-214

39) Medium. (2017). History and Frontier of the Neural Machine Translation. Dostupno na: https://medium.com/syncedreview/history-and-frontier-of-the-neural-machine-translation-dc981d25422d (Datum pristupa: 19.4.2020.)

40) Neubig, G. (2017). Neural Machine Translation and Sequence-to-sequence Models: A Tutorial. Language Technologies Institute, Carengie Mellon University. arXiv:1703.01619v1 [cs.CL]. Dostupno na: https://arxiv.org/pdf/1703.01619.pdf (Datum pristupa: 3.6.2020.)

41) OpenNMT-py. (2017). OpenNMT-py. Dostupno na: https://opennmt.net/OpenNMT-py/quickstart.html (Datum pristupa 15.6.2020.)

42) Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pages 311–318

43) Pathmind. (2019). A Beginner's Guide to Neural Networks and Deep Learning. Dostupno na: https://pathmind.com/wiki/neural-network (Datum pristupa: 26.5.2020.)

44) PCChip. (2016). Koja je razlika između APU, CPU i GPU procesora? Dostupno na: https://pcchip.hr/helpdesk/koja-je-razlika-izmedu-apu-cpu-i-gpu-procesora/ (Datum pristupa: 14.7.2020.)
45) Popadić, D. (2017). Usporedna analiza alata za strojno prevođenje. Diplomski rad: Sveučilište Josipa Jurja Strossmayera u Osijeku, Fakultet elektrotehnike, računarstva i informacijskih tehnologija, Osijek
46) Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in te brain. Psychol. Rev., vol. 65, no. 6, pp. 386–408. Dostupno na: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.335.3398&rep=rep1&type=pdf (Datum pristupa: 25.7.2020.)
47) Seljan, S. (2000). Sublanguage in Machine Translation. Mipro 2000
48) Seljan, S. (2011). Translation technology in education and business. Informatologia 44 (4), 279-286
49) Seljan, S.; Brkić, M.; Vičić, T. (2012). BLEU Evaluation of Machine-Translated English-Croatian Legislation.. LREC, 2143-2148
50) Seljan, S.; Brkić; M.; Kučiš, V. (2011). Evaluation of free online machine translations for Croatian-English and English-Croatian language pairs. INFuture2011: Information Sciences and e-Society, 331-344
51) Seljan, S. (2018a). Total Quality Management Practice in Croatian Language Service Provider Companies. EntreNova 18, 4 (1), 461-469. INFuture2015: e-Institutions–Openness, Accessibility, and Preservation
52) Seljan, S. (2018b). Quality Assurance (QA) of Terminology in a Translation Quality Management System (QMS) in the Business Environment. European Parliament: Translation Services in the Digital World, 92-145
53) Seljan, S.; Dunđer, I. (2014). Combined Automatic Speech Recognition and Machine Translation in Business Correspondence Domain for English-Croatian. Int. Journal of Computer, Information, Systems and Control Engineering, WASET 8 (11), 1069
54) Seljan, S.; Dunđer, I. (2015a). Machine Translation and Automatic Evaluation of English/Russian-Croatian. Proceedings of the International Conference "Corpus Linguistics - 2015". St. Petersburg, Rusija: St. Petersburg State University. Pp. 72-79

55) Seljan, S.; Dunđer, I. (2015b). Automatic quality evaluation of machine-translated output in sociological-philosophical-spiritual domain. Central European Conference on Information and Intelligent Systems (CECIIS), 318
56) Seljan, S.; Dunđer, I.; Pavrlovski, M. (2020). Human Quality Evaluation of Machine-Translated Poetry. MIPRO 2020
57) Seljan, S.; Gašpar, A. (2009). Primjena prevoditeljskih alata u EU i potreba za hrvatskim tehnologijama. Jezična politika i jezična stvarnost, 617-625
58) Seljan, S.; Klasnić, K.: Stojanac, M.; Pešorda, B.; Mikelić Preradović, N. (2015). Information Transfer through Online Summarizating and Translation Technology. INFuture2015: e-Institutions–Openness, Accessibility, and Preservation
59) Seljan, S.; Pavuna, D. (2006). Translation Memory Database in the Translation Proces.. Proceedings of Information and Intelligent Systems IIS, 327-332
60) Seljan, S.; Škof Erdelja, N.; Kučiš, V.; Dunđer, I.; Pejić Bach, M. (2020). Quality Assurance in Computer-Assisted Translation in Business Environments. Natural Language Processing for Global and Local Business. IGI-Global, 247-270
61) Seljan, S.; Tucaković, M.; Dunđer, I. (2015). Human evaluation of online machine translation services for english/russian-croatian. New Contributions in Information Systems and Technologies, 1089-1098
62) Simeon, I. (2008). Vrednovanje strojnoga prevođenja. Doktorska disertacija. Zagreb: Filozofski fakultet
63) StatLect. (2020). Score Vector. Dostupno na: https://www.statlect.com/glossary/score-vector (Datum pristupa: 29.7.2020.)
64) Sutskever, I.; Vinyals, O.; Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104–3112). Dostupno na: https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf (Datum pristupa: 15.6.2020.)
65) Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)
66) Towards Data Science. (2020). A Beginner's Guide to Deep Learning. Dostupno na: https://towardsdatascience.com/a-beginners-guide-to-deep-learning-ed41ac75c4e5 (Datum pristupa: 26.7.2020.)

67) Van Veen F. (2016). The Asimov Institute: The Neural Network Zoo. Dostupno na: https://www.asimovinstitute.org/author/fjodorvanveen/ (Datum pristupa: 27.7.2020.)
68) Vasić, D. (2018). Analiza i primjena metoda automatskog semantičkog označavanja teksta. Split: Sveučilište u Splitu. Fakultet elektrotehnike, strojarstva i brodogradnje. Dostupno na: https://data.fesb.unist.hr/public/news/KDI_Vasic.pdf (Datum pristupa: 20.7.2020.)

Popis slika

Slika 1. Važnost prevođenja (CSA, Research, 2019) 2
Slika 2. Pristupi izgradnji sustava za strojno prevođenje (Kunchukuttan, 2018) 10
Slika 3. Prikaz jednoslojnog perceptrona (Vasić, 2018) 11
Slika 4. Određivanje izlaza neurona (Vasić, 2018) 12
Slika 5. Struktura duboke neuralne mreže (Pathmind, 2019) 12
Slika 6. Razine kompleksnosti značajki (Pathmind, 2019) 14
Slika 7. Opći prikaz enkoder-dekoder arhitekture (Luong i sur., 2017) 16
Slika 8. Graf izračuna vjerojatnosti za 3-gramski jednosmjerni neuralni jezični model (Neubig, 2017) 18
Slika 9. Primjer rekurentne neuralne mreže (Van Veen, 2016) 20
Slika 10. Izračun trenutnog skrivenog stanja h_t (Neubig, 2017) 20
Slika 11. Prikaz arhitekture LSTM-a (Neubig, 2017) 22
Slika 12. Prikaz enkoder-dekoder modela 24
Slika 13. Greedy metoda dekodiranja (Neubig, 2017) 26
Slika 14. Beam metoda dekodiranja (Neubig, 2017) 27
Slika 15. Globalni model mehanizma pažnje (Luong i sur., 2015) 28
Slika 16. Lokalni model mehanizma pažnje (Luong i sur., 2015) 29
Slika 17. Usporedba pivot i "zero-shot" prijevoda (CSA Research, 2017) 32
Slika 18. Usporedba karakteristika NMT-a i SMT-a (Medium, 2017) 34
Slika 19. OPUS korpusi za EN-HR jezični par 36
Slika 20. Prikaz koncepta neuralnog strojnog prevođenja (Klein i sur., 2018) 38
Slika 21. SRC datoteka: engleski korpus za treniranje sustava 39
Slika 22. TGT datoteka: hrvatski korpus za treniranje sustava 40
Slika 23. Ubuntu konzola 41
Slika 24. Pretprocesiranje podataka 42
Slika 25. Treniranje sustava 43
Slika 26. Proces treniranja sustava 44
Slika 27. Spremljene etape treniranog sustava 45
Slika 28. Završetak procesa treniranja sustava 46
Slika 29. Primjeri prijevoda 1 47
Slika 30. Primjeri prijevoda 2 47

Slika 31. Rezultat evaluacije sustava na temelju analiziranih prijevoda 52
Slika 32. Grafički prikaz BLEU rezultata evaluacije nad cijelom datotekom 53

Popis tablica

Tablica 1. Izračun unigramske preciznosti 50
Tablica 2. Rečenice s maksimalnim i minimalnim BLEU rezultatom 55
Tablica 3. Analiza rezultata 1 57
Tablica 4. Analiza rezultata 2 58
Tablica 5. Analiza rezultata 3 59

Izgradnja sustava za neuralno strojno prevođenje

Sažetak

Konstantni razvoj tehnologije omogućava napredak u svim područjima njezine primjene. Jedno od tih područja je i strojno prevođenje čija je važnost neizmjerna u vremenima globalizacije. Potrebno je osigurati neometan protok informacija koje će svima biti jasne i razumljive. U svrhu postizanja tog cilja važno je raditi na optimiziranju sustava za strojno prevođenje. U radu će biti prikazani teorijski i praktični aspekti izgradnje sustava za neuralno strojno prevođenje te će se usporediti statistički i neuralni model strojnog prevođenja. Sustav će biti izgrađen u odabranoj domeni nad pripremljenim skupom podataka u kojoj će se zatim testirati i evaluirati.

Ključne riječi: strojno prevođenje, neuralne mreže, duboko učenje, neuralno strojno prevođenje, umjetna inteligencija

Development of a neural machine translation system

Summary

A continuous technological change improves every aspect of our lives where it is being implemented. One of those aspects is also machine learning whose importance is immense in times of globalization. Nowadays it is necessary to ensure a consistent flow of information that is understandable and comprehensible for everyone. In order to achieve that goal, it's important to work on optimizing machine translation systems and implementing the newest technologies. This thesis is intended to represent theoretical and practical aspects of development of a neural machine traslation system and to compare statistical and neural model of machine translation. A machine translation system will be built in a chosen domain with a prepared set of data in which it will be also tested and evaluated.

Ključne riječi: machine translation, neural networks, deep learning, neural machine translation, artificial intelligence