Sveučilište u Rijeci
University of Rijeka
http://www.uniri.hr

Polytechnic: Journal of Technology Education, Volume X, Number Y (2021)
Politehnika: Časopis za tehnički odgoj i obrazovanje, Volumen X, Broj Y (2021)

Politehnika
Polytechnica
http://www.politehnika.hr/journal
cte@uniri.hr

DOI: https://doi.org/10.36978/cte.5.2.4

Stručni članak
Professional paper
UDK: 81'322.4
81'25:004.42

Analiza modela sustava za automatsko statističko strojno prevođenje

Ivan Dunđer
Filozofski fakultet
Sveučilište u Zagrebu
Ivana Lučića 3
ivandunder@gmail.com, idundjer@ffzg.hr

Sažetak

Automatsko strojno prevođenje sve je popularnija istraživačka tema u znanosti i raznim znanstvenim disciplinama, kao što su informacijske i komunikacijske znanosti, računarstvo, računalna lingvistika i sl. Razlog tome je prvenstveno to što danas omogućuje nezaobilaznu komunikaciju i brz prijenos informacija između različitih prirodnih jezika. To je posebno bitno za manje govorene jezike poput hrvatskoga, za koji još uvijek ne postoji dovoljan broj softverskih alata i digitalnih resursa potrebnih za razvoj specijaliziranih i kvalitetnih sustava za strojno prevođenje koji bi bili optimizirani za upotrebu u jednom specifičnom području. Sve brži rast količine podataka i sve veća potreba raznih dionika u sektorima industrije, gospodarstvu, znanosti, ali i u svakodnevnom životu ljudi impliciraju motivaciju za sistematiziranim i organiziranim razvojem te naknadnom prilagodbom sustava za automatsko strojno prevođenje za različite jezične parove. Budući da strojni prijevodi nisu savršeni, važno je primijeniti metode za računalno generiranje prijevoda prihvatljive razine kvalitete koja ovisi o samom zadatku i području primjene sustava za strojno provođenje. U ovom radu analiziran je model sustava za automatsko statističko strojno prevođenje, njegove komponente te uloga i značaj pojedinih elemenata unutar modela.

Ključne riječi: *automatsko strojno prevođenje; model statističkog strojnog prevođenja; jezične tehnologije; računalna obrada prirodnog jezika; informacijske i komunikacijske znanosti.*

1 Uvod

Tehnologija automatskog strojnog prevođenja danas je jedna od neizostavnih disruptivnih tehnologija koja uvelike doprinosi cjelovitoj transformaciji poslovnih procesa u segmentu prevođenja tekstova sastavljenih na prirodnom jeziku.

Glavna ideja iza primjene tehnologije strojnog prevođenja je razvoj i prilagodba specijaliziranih sustava za automatsko prevođenje koji bi bili u stanju automatizirati barem jedan veći dio procesa prevođenja – procesa koji se smatra skupim i vrlo zahtjevnim poslom, posebno u kontekstu velike količine podataka. To je višestruko izraženo u donašanjem informacijskom dobu i eri sveopće dominacije informacija koje predstavljaju najvrjedniji resurs, ali koje kao takve moraju biti pravovremeno na raspolaganju.

Primjenom sustava za automatsko strojno prevođenje značajno se ubrzava i potencijalno optimizira poslovanje jedne organizacije, što omogućuje toliko željenu konkurentsku prednost na

globalnom tržištu koje se neprekidno i sve brže mijenja.

No, razvoj sustava za automatsko strojno prevođenje nije jednostavan zadatak, budući da se prirodni jezici vrlo teško mogu formalizirati jezičnim pravilima i kodirati međusobnim odnosima, a upravo takav pristup koristi se u strojnom prevođenju temeljenom na (jezičnim) pravilima (eng. Rule-based machine translation, RBMT).

Jedan od pristupa strojnom prevođenju koji se pak oslanja na statistiku, frekvencije i distribuciju riječi u tekstu je tzv. statističko strojno prevođenje (eng. Statistical machine translation, SMT). To je jedan od popularnih i dominantnih pristupa danas, uz neuralno strojno prevođenje (eng. Neural machine translation, NMT) (Kamath i sur., 2019).

Neuralno strojno prevođenje u svom standardnom modelu izravno optimizira uvjetne vjerojatnosti ciljne rečenice na temelju dane izvorne rečenice (Cho i sur., 2014). Vjerojatnosti se definiraju na umjetnoj neuronskoj mreži koja se temelji na (en)koder-dekoder arhitekturi (Sutskever i sur., 2014), pri čemu (en)koder obuhvaća izvornu rečenicu i preslikava ju u niz odgovarajućih reprezentacija, a dekoder potom generira riječi u ciljnom jeziku na temelju tih reprezentacija (Cho i sur., 2014). Takva arhitektura se uobičajeno izvodi pomoću povratne neuronske mreže (RNN), konvolucijske neuronske mreže (CNN) ili transformera. Model transformera danas je najkorišteniji pristup neuralnom strojnom prevođenju.

Ovaj rad predstavlja nastavak ranijeg opsežnog istraživanja (Dunđer, 2015), a cilj mu je dati kratak pregled porijekla statističkog strojnog prevođenja, specifičnog modela za statističko strojno prevođenje temeljenog na frazama, analizirati faze razvoja sustava, prikazati komponente takvog modela te istražiti ulogu i značaj elemenata u modelu sustava za strojno prevođenje. Ovim radom se naglašavaju izazovi i aktivnosti u razvoju sustava za statističko strojno prevođenje te se prezentiraju recentnija istraživanja na području tehnologija strojnog prevođenja za hrvatski jezik. Nadalje, u radu je iskazano mišljenje autora o primjeni različitih pristupa strojnom prevođenju te mogućnostima unaprjeđenja automatskog strojnog prevođenja.

2 Porijeklo statističkog strojnog prevođenja

Jedan od pristupa automatskom strojnom prevođenju jest pristup upravljan podatcima (eng. data-driven approach). Takav pristup uključuje i model statističkog strojnog prevođenja (Brown i sur., 1993) koji je neovisan o jeziku te stoga u pravilu ne zahtijeva posebno lingvističko znanje (Koehn, 2010).

Istraživanja na području statističkog strojnog prevođenja započela su kasnih 1980-ih na IBM-ovom projektu „CANDIDE“ koji je za cilj imao izgradnju sustava za automatsko prevođenje s francuskog na engleski jezik (Koehn, 2010).

Izvorna IBM-ova istraživanja statističkog strojnog prevođenja rezultirala su definiranjem tzv. IBM modela 1-5, koji pripadaju generativnoj metodi modeliranja. Naime, proces generiranja podataka dijele u manje korake koji se zatim zasebno modeliraju i statistički opisuju, a konačan rezultat nastaje kombiniranjem svih pojedinačnih koraka. IBM modeli polaze od uparivanja pojedinih riječi u izvornom i ciljnom jeziku te dozvoljavaju umetanje i ispuštanje riječi (Koehn, 2010).

3 Statističko strojno prevođenje temeljeno na frazama

Statističko strojno prevođenje posljednjih je godina od velikog istraživačkog interesa, osobito zbog mogućnosti izgradnje sustava za automatsko strojno prevođenje primjenom velike količine jezičnih resursa u obliku paralelnih i jednojezičnih korpusa te jezično neovisnih alata.

Paralelni korpusi su podatkovni skupovi koji se sastoje od tekstova na izvornom i ciljnom jeziku, odnosno ciljnom i izvornom jeziku (Klaper i sur., 2013). Takvi korpusi predstavljaju osnovno sredstvo rada sustava za statističko strojno prevođenje (Wetzel i Bond, 2012).

Jedan od mogućih pristupa statističkom strojnom prevođenju jest primjena prijevodnog modela temeljenog na frazama (eng. phrase-based translation model) (Koehn, 2010).

Osnova ideja statističkog strojnog prevođenja temeljenog na frazama jest segmentirati skup rečenica izvornog jezika u fraze, odnosno nizove riječi koji se zatim prevode u ciljni jezik. Cjelovite rečenice u ciljnom jeziku nastaju sastavljanjem prijevoda fraza, a nazivaju se prijevodni kandidati (eng. candidate translations).

U modelu statističkog strojnog prevođenja riječi izvornog i ciljnog jezika se uparuju (sravnjuju), a same fraze ekstrahiraju se upravo iz sravnjenosti riječi u izvornom i ciljnom jeziku, pri čemu sve riječi iz fraznog para trebaju biti međusobno sravnjene (Koehn, 2010).

4 Faze u statističkom strojnom prevođenju

Prema Koehnu (2010), tri su ključne faze u statističkom strojnom prevođenju temeljenom na frazama: sravnjivanje riječi, ekstrakcija fraznih parova

i izračun vjerojatnosti za svakih frazni par. Nadalje, tri komponente modela sustava za strojno prevođenje temeljeno na frazama izravno utječu na kvalitetu statističkog strojnog prijevoda (Koehn, 2010; Koehn, 2008), a to su:

- tablica prijevoda fraza, tj. fraznih struktura (eng. phrase translation table),
- model preslagivanja, tj. premještanja redoslijeda (poretka) riječi (eng. reordering model), i
- jezični model (eng. language model).

Tim komponentama modela, tj. značajkama modela (eng. features) se u postupku učenja, tj. treniranja modela pridružuju određene vrijednosti težina (eng. weights) koje utječu na logiku modela statističkog strojnog prevođenja temeljenog na frazama (Koehn, 2010).

Značajke i težine su u modelu statističkog strojnog prevođenja temeljenog na frazama implementirani u obliku log-linearnog modela (Jurafsky i Martin, 2013). Log-linearni model je matematički model koji preuzima oblik funkcije čiji je logaritam jednak linearnoj kombinaciji parametara, odnosno značajki modela.

S obzirom na to da se izračuni vrijednosti težina pojedinih (pod)modela (prijevodni model, jezični model itd.) odvijaju u zasebnim koracima, tj. procesima, modeli nemaju optimalne vrijednosti, tj. parametre za dekodiranje i generiranje strojnog prijevoda. Stoga, primjenom višestrukog, ali ograničenog broja ponavljanja postupka ugađanja (eng. tuning), sustav postupno usklađuje vrijednosti pojedinih značajki modela za statističko strojno prevođenje (Koehn, 2010). Svakom prijevodnom kandidatu pridružen je skup pripadajućih vrijednosti težina značajki modela.

Podatkovni skup za treniranje n-gramskog jezičnog modela (eng. monolingual training set) čini jednojezični korpus ciljnog jezika (Koehn, 2015), pri čemu n-gram predstavlja niz uzastopnih riječi (Koehn, 2010). N-gramski jezični model je vjerojatnosni jezični model koji se koristi za predviđanje idućeg elementa u nizu riječi pomoću Markovljevog stohastičkog modela.

S obzirom na to da n-gramski jezični model ne može pokriti sve varijacije n-grama, postoji opasnost da se određenim nizovima riječi u postupku učenja jezičnog modela pridruže vjerojatnosti jednake nuli (Madnani, 2010).

Stoga se pribjegava metodama izglađivanja (eng. smoothing) koje do tada neviđenim n-gramima pridružuju vjerojatnosti veće od nule, s obzirom na to da vjerojatnost nula izrazito loše utječe na konačnu procjenu vjerojatnosti niza riječi. Distribucija vjerojatnosti tako postaje glađa (eng. smoother), a zadatak izglađivanja je od viđenih n-grama oduzeti djelić vjerojatnosti te ga distribuirati među neviđenim n-gramima (Koehn, 2010).

Nadalje, bolji jezični modeli mogu primjerice kombinirati procjene vjerojatnosti n-grama iz više različitih jezičnih modela (Madnani, 2010). Pod pojmom „bolji“ podrazumijevaju se jezični modeli koji pri intrinzičnoj evaluaciji postižu manju perpleksnost, a pri vanjskoj evaluaciji veću razinu fluentnosti u generiranom ciljnom prijevodu (Dunđer, 2015) – drugim riječima, generiraju se kvalitetniji i tečniji strojni prijevodi.

Ulazni podatkovni skup koji se koristi za treniranje prijevodnog modela sadrži rečenice, odnosno segmente teksta na izvornom jeziku te semantičke prijevodne ekvivalente na ciljnom jeziku, ekstrahirane iz nekog paralelnog korpusa (Koehn, 2005).

Ukoliko je na raspolaganju ograničena količina kvalitetnog paralelnog korpusa, sustav za strojno prevođenje treba podesiti na način da veću težinu pridaje jezičnom modelu ciljnog jezika (Mauser i sur., 2008), a manje prijevodnom modelu.

A upravo evaluacija kvalitete strojnih prijevoda automatskim metrikama poput BLEU, NIST, METEOR, GTM, ROUGE, WER, TER, PER i sl. (Dunđer, 2015) može dati korisne smjernice na koji način treba podesiti težine pojedinih modela (Gonzàlez, 2014).

5 Komponente modela za statističko strojno prevođenje

Model sustava za statističko strojno prevođenje temeljeno na frazama, za razliku od ostalih pristupa strojnom prevođenju, posebno se obazire na leksičke i morfološke varijacije riječi, koje su osobito važne kod morfološki bogatih jezika kao što je hrvatski, budući da to rezultira kvalitetnijim strojnim prijevodima.

Učinkovito upravljanje leksičkim varijacijama riječi predstavlja snagu modela sustava za statističko strojno prevođenje temeljeno na frazama (Eisele i sur., 2008). Takav model favorizira ili penalizira određene prijevode i šumove te se ne ograničava na fraze u lingvističkom smislu, kao sintaktički motivirane skupine riječi (Koehn, 2010).

Naprotiv, ograničavanje na sintaktički motivirane fraze umanjuje kvalitetu statističkog strojnog prijevoda (Koehn i sur., 2003). Naime, model funkcionira na statističkoj razini koristeći Bayesov teorem, Markovljev lanac i druge statističke paradigme za učenje iz paralelnih korpusa (Koehn, 2010; Manning i Schütze, 1999), a sam statistički strojni prijevod rezultat je niza ulančanih i uvjetovanih odluka koje se donose s određenom vjerojatnošću (Ueffing i sur., 2007).

Standardni model statističkog strojnog prevođenja sastoji se od tri ključne komponente: jezični model (eng. language model), prijevodni model (eng. translation model) i dekoder (eng. decoder) (Jurafsky i Martin, 2013).

Bayesovim teoremom opisuje se vjerojatnost prevođenja segmenta iz izvornog *f* u ciljni jezik *e*, tj. *p(e|f)*, što je prikazano jednadžbom u nastavku (Knight, 1999).

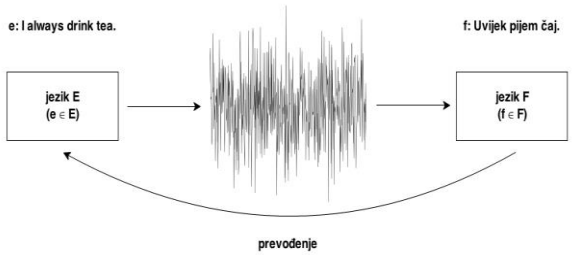
$$p(e|f) = \frac{p(e)p(f|e)}{p(f)}$$

S obzirom na to da je segment f konstantan u odnosu na sve moguće pripadajuće prijevode *e*, *p(f)* kao neovisna vjerojatnost od f se može zanemariti, što rezultira jednadžbom u nastavku (España-Bonet i Gonzàlez, 2014; Manning i Schütze, 1999).

$$e = argmax_e p(e|f) = argmax_e p(e)p(f|e)$$

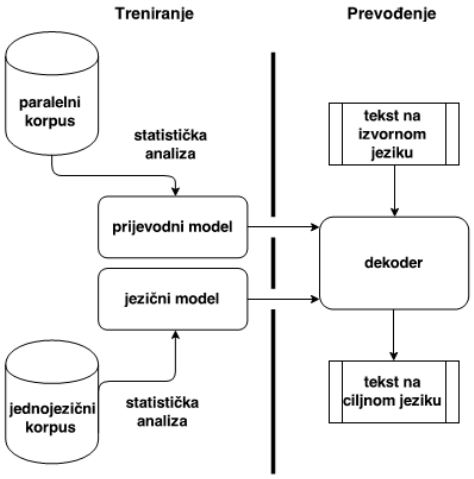
Takav pristup omogućuje razdvajanje i kombinaciju jezičnog modela *p(e)* i prijevodnog modela *p(f|e)*, što ga svrstava u klasu modela kanala sa šumom (eng. noisy-channel model) (Koehn, 2010). U modelu kanala sa šumom, *f* predstavlja poruku, tj. rečenicu na izvornom jeziku koju treba prevesti u poruku, tj. rečenicu na ciljnom jeziku *e*.

Prijevodni model definira kako se ostvaruje prijevod poruke, tj. rečenice, dok jezični model procjenjuje koje su rečenice u ciljnom jeziku vjerojatne. Model kanala sa šumom prikazan je na jednom primjeru u nastavku (Slika 1) (Dunđer, 2015, adaptirano prema España-Bonet i Gonzàlez, 2014).



Slika 1. Prikaz modela kanala sa šumom.

Slika 2 prikazuje model sustava za statističko strojno prevođenje (Dunđer, 2015, adaptirano prema Koehnu, 2004, odnosno Wayu i Hassanu, 2009).



Slika 2. Model statističkog strojnog prevođenja s obzirom na fazu treniranja statističkih modela i prevođenja novog teksta.

Izvorni IBM modeli (Brown i sur., 1993) nalažu da se proces strojnog prevođenja treba dekomponirati u manje korake, pri čemu se svaki korak oslanja na riječi kao atomarne prijevodne jedinice (eng. atomic translation units) (Koehn, 2010).

Na kvalitetu strojnog prijevoda utječu i razlike između izvornoga i ciljnoga jezika. Naime, model statističkog strojnog prevođenja lakše prevodi s morfološki složenijeg jezika na manje složen jezik (Koehn, 2006).

Nadalje, ispostavilo se, da se udvostručavanjem podatkovnih skupova, odnosno paralelnog korpusa za treniranje prijevodnog modela ili jednojezičnog korpusa za treniranje jezičnog modela, kontinuirano poboljšava i kvaliteta strojnog prijevoda, što se reflektira u rezultatima automatskih metrika (Turchi i sur., 2012).

6 Izazovi i aktivnosti u razvoju sustava za statističko strojno prevođenje

Treba spomenuti da statistički pristup strojnom prevođenju ima i brojne nedostatke te izazove. Na ortografskoj, tj. leksičkoj razini i najmanja pravopisna ili tipografska pogreška povećava vokabular i onemogućuje prevođenje riječi s obzirom na to da za vrijeme treniranja modela takva riječ nije viđena, a stoga ni statistički opisana.

Zbog toga se prije same izgradnje sustava za strojno prevođenje podatkovni skupovi trebaju pretprocesirati (eng. data preprocessing) – to je postupak kojim se podatci odgovarajuće obrađuju i pripremaju za naredne aktivnosti (Koehn, 2015; Buck i sur., 2014). To znači da se riječi iz podatkovnih skupova prije treniranja modela pretvaraju u riječi zapisane malim slovima (eng. lowercasing) kako bi se

izbjegle ortografske neusklađenosti među jednakim riječima.

Ipak, ispravan zapis riječi treba sačuvati. Proces pohranjivanja ispravnog zapisa početnog znaka riječi i postupak naknadnog pretvaranja početnog znaka riječi u odgovarajuće malo ili veliko slovo (eng. truecasing) može se primijeniti nakon procesa generiranja strojnog prijevoda, tj. dekodiranja novog podatkovnog skupa (eng. recasing/capitalisation).

Osim toga, često se postupkom normalizacije uklanjaju riječi koje su semantički jednake, a treba izbjegavati i dvoznačne ili višeznačne konstrukcije.

Transliteracija predstavlja pretvaranje niza znakova iz jednog pravopisa u drugi, i time se može sačuvati fonetika u oba jezika (Manning i Schütze, 1999). Transliteracija odnosi se u pravilu na imena i brojeve (Koehn, 2010).

Tokenizacija (eng. tokenisation) je postupak rastavljanja na pojavnice i time se zapravo tekst dijeli u riječi, tj. umeću se razmaci između riječi i interpunkcije (u jezicima s latinskim alfabetom) (Manning i Schütze, 1999). Detokenizacija (eng. detokenisation) je suprotan proces kojim se tokenizirani tekst pretvara u prirodan/ispravan oblik.

Podatkovne skupove treba i prethodno očistiti (eng. corpus cleaning), tj. treba ukloniti predugačke, prekratke, neuparene, nekompatibilne i prazne rečenice, odnosno segmente te suvišne razmake među riječima (Koehn, 2015).

Morfološki bogati jezici, poput hrvatskoga koji obiluje velikim brojem različitih morfema, također znatno povećavaju kompleksnost sustava za statističko strojno prevođenje.

Sintaksa definira načela i pravila konstrukcije rečenica u prirodnom jeziku, međutim, ona nije integrirana u klasičnom modelu statističkog strojnog prevođenja temeljenog na frazama, s obzirom na to da su fraze obični nizovi riječi bez podataka o samoj strukturi niza riječi (Koehn, 2010). Zbog toga su problemi s redoslijedom riječi vrlo česti u statističkom strojnom prevođenju, pogotovo kada se prevodi s jezika s relativno slobodnim poretkom riječi u jezik s relativno fiksnim poretkom riječi kao što je engleski, ali i obrnuto.

Jednako tako, ukoliko izvorni i ciljni jezik imaju različite strukture (npr. subjekt–glagol–objekt i subjekt–objekt–glagol) problemi s ispravnim redoslijedom riječi vrlo često su neizbježni (Reddy i Hanumanthappa, 2013), naročito ukoliko se premještanje riječi unutar jedne rečenice treba izvršiti preko velike udaljenosti (eng. long-range reordering). Na semantičkoj razini, značenje jedne rečenice izravno ovisi o dijelovima i odnosima unutar rečenice. Stoga se riječi koje čine jednu rečenicu, tj. neposredni kontekst, opisuju u jezičnom modelu ciljnoga jezika.

7 Tehnologija strojnog prevođenja za hrvatski jezik

Tehnologija strojnog prevođenja istražena je i u kontekstu hrvatskoga jezika. U tu svrhu razvijeno je više sustava za strojno prevođenje te je provedeno opsežno istraživanje u području metoda adaptacije domene (Dunđer, 2015).

Strojno prevođenje istraženo je u raznim domenama, poput poezije (Dunđer i sur., 2020) s posebnim osvrtnom na ljudsku evaluaciju kvalitete strojnog prijevoda (Seljan i sur., 2020). Specijaliziran sustav za automatsko strojno prevođenje za domenu industrije također je razvijen za hrvatski jezik (Dunđer, 2020).

Analizirane su i mogućnosti osiguranja kvalitete prijevoda generiranih pomoću alata za računalno-potpomognuto prevođenje u poslovnom okruženju (Seljan i sur., 2020).

Automatska evaluacija kvalitete strojnog prijevoda ispitana je u domeni sociologije, filozofije i religioznosti (Seljan i Dunđer, 2015a) te za različite jezične parove, uključujući i hrvatski jezik (Seljan i Dunđer, 2015b).

Ljudska evaluacija prijevoda za hrvatski jezik i primjena *online* servisa za strojno prevođenje također su analizirani u jednom istraživanju u kojemu su ispitane mjere fluentnosti i adekvatnosti strojnih prijevoda te interna razina slaganja/neslaganja među ljudskim evaluatorima (Seljan i sur., 2015).

Za ljudsku evaluaciju adekvatnosti i fluentnosti te pripadajuću konzistentnost korišten je i hi-kvadrat test (Brkić i sur., 2009). Izvršena je i analiza pogrešaka u strojnim prijevodima na razini segmenata teksta uz primjenu metrika za leksičku sličnost, i to za njemačko-hrvatski jezični par (Brkic Bakaric i sur., 2020).

Komparativna analiza pogrešaka u strojnim prijevodima njemačko-engleski-hrvatski jezični smjer dodatno je potkrijepljena analizom razine (ne)slaganja ljudskih evaluatora (Brkic Bakaric i sur., 2017).

Za potrebe analize utjecaja idioma na kvalitetu englesko-hrvatskih strojnih prijevoda razvijena je posebna taksonomija pogrešaka (Manojlović i sur., 2017).

Integracija strojnog prevođenja i automatskog prepoznavanja govora također je analizirana za hrvatski jezik u domeni poslovne korespondencije (Seljan i Dunđer, 2014).

Pored toga, istražene su i brojne mogućnosti za pripremu digitalnih resursa koji su prijeko potrebni za razvoj sustava za strojno prevođenje, poput digitalizacije dokumenata i optičkog prepoznavanja znakova (Seljan i sur., 2013) ili primjene

crowdsourcinga kao jedne od mogućnosti oslanjanja na fenomen mnoštva (Jaworski i sur., 2017).

8 Diskusija i mogući pravci razvoja

Proteklih nekoliko desetljeća strojno prevođenje kao istraživačko područje obilježeno je razvojem nekoliko ključnih pristupa strojnom prevođenju. Radi se o strojnom prevođenju temeljenom na pravilima, na statistici te umjetnim neuronskim mrežama. Pojava i razvoj svakog od navedenih pristupa revolucionirali su cjelokupno područje strojnog prevođenja i značajno unaprijedili efikasnost sustava za automatsko strojno prevođenje.

U proteklih nekoliko godina neuralno strojno prevođenje razvilo se do te mjere da može nadmašiti sve ranije pristupe. Takav pristup strojnom prevođenju izgrađuje jednu veliku umjetnu neuronsku mrežu pomoću koje se izvodi cijeli proces automatskog prevođenja (Sutskever i sur., 2014).

No, velik broj problema u strojnom prevođenju i dalje nije otklonjen. Stoga, snaga strojnog prevođenja po mišljenju autora ovoga rada leži u kombinaciji više različitih pristupa strojnom prevođenju, budući da se problemu prevođenja tada može pristupati s više različitih stajališta, a time se ujedno rješavaju i različiti problemi u prevođenju prirodnih jezika.

Model statističkog strojnog prevođenja se može proširiti i dodatnim značajkama koje uključuju jezično znanje te druge raspoložive jezične resurse. Pa unatoč tome što je područje strojnog prevođenja trenutno usmjereno prema razvoju i optimizaciji sustava za neuralno strojno prevođenje, budućnost strojnog prevođenja će vjerojatno biti obilježena modeliranjem hibridnih sustava koji u svojim arhitekturama objedinjuju različite pristupe strojnom prevođenju pa tako i statistički pristup.

9 Zaključak

Automatsko strojno prevođenje privlači sve više pozornosti u brojnim segmentima istraživačke zajednice koji se bave različitim temama u rasponu od umjetne inteligencije, računalne obrade prirodnog jezika, računalne lingvistike, strojnog učenja i znanosti o podacima.

Strojno prevođenje vrlo je složen zadatak u kojemu se stroj, tj. računalo priprema i koristi za potrebe prevođenja tekstualnog podatkovnog skupa s izvornog na jedan ili više ciljnih jezika, i to bez ljudskog sudjelovanja ili uz minimalne intervencije. Strojno prevođenje danas ima veliku primjenu – u gospodarstvu, industriji, znanosti, državnim tijelima, ali i u svakodnevnom životu običnih ljudi.

Iako postoji nekoliko pristupa strojnom prevođenju, danas uz statističko strojno prevođenje dominira pristup koji se oslanja na umjetne neuronske mreže. Statistički pristup oslanja se na matematički opis jezika i koristi više značajki modela koji se moraju zasebno trenirati i prilagođavati određenom zadatku i domeni. S druge strane neuralno strojno prevođenje ne zahtijeva zasebno strojno učenje značajki modela.

Cilj ovoga rada bio je analizirati model sustava za automatsko statističko strojno prevođenje te time istražiti pristup strojnom prevođenju koji je neovisan o lingvističkom znanju o nekom konkretnom jeziku.

Porijeklo statističkog strojnog prevođenja je također prikazano u ovom radu, kao i jedna posebna klasa statističkog strojnog prevođenja koja se oslanja na statistički motivirane fraze.

Objašnjene su faze u razvoju sustava za statističko strojno prevođenje te prikazane ključne komponente modela takvog jednog sustava. Budući da je tehnologija strojnog prevođenja vrlo složena, u radu su analizirani izazovi s kojima se susrećemo prilikom razvoja i prilagodbe sustava za strojno prevođenje.

Nadalje, pojašnjene su aktivnosti koje treba poduzeti kako bi se ublažila kompleksnost modela sustava, a budući da je tehnologija strojnog prevođenja važna za sve države u informacijskom dobu, u radu se autor osvrće i na recentnija istraživanja koja uključuju hrvatski jezik u tom istraživačkom području.

Znanstveni doprinos ovoga rada leži i u kritičkom promišljanju budućnosti i mogućim pravcima razvoja područja strojnog prevođenja. Naime, prema mišljenju autora ovoga rada, činjenicu da se strojnom prevođenju može pristupiti pomoću više pristupa treba iskoristiti za modeliranje hibridnih sustava koji kombiniraju različite algoritme i arhitekture te koji sveobuhvatnom problemu prevođenja pristupaju s različitih polazišnih točaka, a time ujedno rješavaju i različite vrste probleme u procesu automatskog strojnog prevođenja.

Literatura

Brkic Bakaric, M., Babic, N., Dajak, L., Manojlovic, M. (2017). A comparative error analysis of English and German MT from and into Croatian. *Proceedings of the INFuture2017: Integrating ICT in Society Conference (INFuture 2017)* (pp. 31-41).

Brkic Bakaric, M., Tonkovic, K., Nacinovic Prskalo, L. (2020). Clash between Segment-level MT Error Analysis and Selected Lexical Similarity Metrics. *International Journal of Advanced Computer Science and Applications*, 11 (5), 35–42.

Brkić, M., Vičić, T., Seljan, S. (2009). Evaluation of the Statistical Machine Translation Service for Croatian-English. *Proceedings of the 2nd*

International Conference The future of information sciences: Digital resources and knowledge sharing (INFuture 2009) (pp. 319-332).

Brown, P. F., Della Pietra, V. J., Della Pietra, S. A., Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics - Special issue on using large corpora II*, 19(2), pp. 263–311.

Buck, C., Heafield, K., van Ooyen, B. (2014). N-gram Counts and Language Models from the Common Crawl. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)* (pp. 3579-3584). Language Resources and Evaluation Conference.

Cho, K., van Merriënboer, B., Bahdanau, D., Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)* (pp. 103-111).

Dunđer, I. (2015). Sustav za statističko strojno prevođenje i računalna adaptacija domene (Statistical Machine Translation System and Computational Domain Adaptation) / doctoral dissertation. University of Zagreb.

Dunđer, I. (2020). Machine Translation System for the Industry Domain and Croatian Language. *Journal of Information and Organizational Science (JIOS)*, 44(1), 33–50.

Dunđer, I., Seljan, S., Pavlovski, M. (2020). Automatic Machine Translation of Poetry and a Low-Resource Language Pair. *Proceedings of the 43rd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2020)* (pp. 1034-1039).

Eisele, A., Christian, F., Uszkoreit, H., Saint-Amand, H., Kay, M., Jellinghaus, M., Hunsicker, S., Herrmann, T., Chen, Y. (2008). Hybrid Architectures for Multi-Engine Machine Translation. *Translating and the Computer*, 30, 12.

España-Bonet, C., Gonzàlez, M. (2014). Statistical Machine Translation and Automatic Evaluation / tutorial documentation. *The 9th edition of the Language Resources and Evaluation Conference (LREC 2014)* (p. 308). Language Resources and Evaluation Conference.

Gonzàlez, M. (2014). Automatic MT Evaluation / tutorial documentation. *The 9th edition of the Language Resources and Evaluation Conference (LREC 2014)* (p. 76). Language Resources and Evaluation Conference.

Jaworski, R., Seljan, S., Dunđer I. (2017). Towards educating and motivating the crowd – a crowdsourcing platform for harvesting the fruits of NLP students' labour. *Proceedings of the 8th Language & Technology Conference – Human Language Technologies as a Challenge for Computer Science and Linguistics* (pp. 332-336).

Jurafsky, D., Martin, J. (2013). *Speech and Language Processing*. Pearson New International Edition. Pearson Education Limited, 2nd edition, 2013.

Kamath, U., Liu, J., Whitaker, J. (2019*). Deep Learning for NLP and Speech Recognition*. Berlin: Springer, p. 621.

Klaper, D., Ebling, S., Volk, M. (2013). Building a German/Simple German Parallel Corpus for Automatic Text Simplification. *Proceedings of the ACL 2013 Conference: The Second Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2013)* (pp. 11-19). Association for Computational Linguistics.

Knight, K. A. (1999). Statistical MT Tutorial Workbook. *JHU summer workshop* (p. 36).

Koehn, P. (2004). Challenges in Statistical Machine Translation. Presentation at PARC, Google, ISI, MITRE, BBN, University of Montreal (p. 51).

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *AAMT: The Tenth Machine Translation Summit* (pp. 79-86).

Koehn, P. (2006). Statistical Machine Translation: the basic, the novel, and the speculative / tutorial documentation. *EACL: 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)* (p. 81). European Chapter of the Association for Computational Linguistics.

Koehn, P. (2008). Introduction to Statistical Machine Translation / tutorial documentation. *Chinese Workshop for Machine Translation* (p. 214).

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Koehn, P. (2015). *Moses - Statistical Machine Translation System: User Manual and Code Guide*. University of Edinburgh.

Koehn, P., Och, F. J., Marcu, D. (2003). Statistical Phrase-Based Translation. *Proceedings of the 2003 Human Language technology Conference - North American Chapter of the Association for Computational Linguistics (HLT/NAACL 2003)* (p.

7). North American Chapter of the Association for Computational Linguistics.

Madnani, N. (2010). Language Models / course material. INFM718G/CMSC838G course on Data-Intensive Information Processing Applications (Lin, J.; Madnani, N.), p. 63. University of Maryland.

Manning, C. D., Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.

Manojlović, M., Dajak, L., Brkić Bakarić, M. (2017). Idioms in state-of-the-art Croatian-English and English-Croatian SMT systems. *Proceedings of the 40th Jubilee International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2017)* (pp. 1798-1802).

Mauser, A., Hasan, S., Ney, H. (2008). Automatic Evaluation Measures for Statistical Machine Translation System Optimization. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)* (pp. 3089-3092). Language Resources and Evaluation Conference.

Reddy, M. V., Hanumanthappa, M. (2013). NLP challenges for machine translation from English to Indian languages. *International Journal of Computer Science and Informatics*, 3(1), p. 35.

Seljan, S., Dunđer, I. (2014). Combined Automatic Speech Recognition and Machine Translation in Business Correspondence Domain for English-Croatian. *Proceedings of the International Conference on Embedded Systems and Intelligent Technology (ICESIT 2014) – International Journal of Computer, Information, Systems and Control Engineering*, vol. 8 (pp. 1069-1075).

Seljan, S., Dunđer, I. (2015a). Automatic Quality Evaluation of Machine-Translated Output in Sociological-Philosophical-Spiritual Domain. *Proceedings of the 10th Iberian Conference on Information Systems and Technologies (CISTI'2015)*, vol. 2 (pp. 128-131).

Seljan, S., Dunđer, I. (2015b). Machine Translation and Automatic Evaluation of English/Russian-Croatian. *Proceedings of the International Conference "Corpus Linguistics – 2015" (CORPORA 2015)* (pp. 72-79).

Seljan, S., Dunđer, I., Gašpar, A. (2013). From Digitisation Process to Terminological Digital Resources. *Proceedings of the 36th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2013)* (pp. 1329-1334).

Seljan, S., Dunđer, I., Pavlovski, M. (2020). Human Quality Evaluation of Machine-Translated Poetry. *Proceedings of the 43rd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2020)* (pp. 1040-1045).

Seljan, S., Škof Erdelja, N., Kučiš, V., Dunđer, I., Pejić Bach, M. (2020). Quality Assurance in ComputerAssisted Translation in Business Environments. *Natural Language Processing for Global and Local Business*, IGI Global, p. 22.

Seljan, S., Tucaković, M., Dunđer, I. (2015). Human Evaluation of Online Machine Translation Services for English/Russian-Croatian. *Proceedings of the WorldCIST'15 – 3rd World Conference on Information Systems and Technologies (Advances in Intelligent Systems and Computing – New Contributions in Information Systems and Technologies)* (pp. 1089-1098).

Sutskever, I., Vinyals, O., Le, Q. V. Sequence to Sequence Learning with Neural Networks. *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14)*, vol. 2, pp. (3104-3112).

Turchi, M., Goutte, C., Cristianini, N. (2012). Learning Machine Translation from In-domain and Out-of-domain Data. *Proceedings of 16th Annual Conference of the European Association for Machine Translation* (pp. 305-312). European Association for Machine Translation.

Ueffing, N., Haffari, G., Sarkar, A. (2007). Semi-supervised model adaptation for statistical machine translation. *Machine Translation Journal*, 21(2), 77–94.

Way, A., Hassan, H. (2009). Statistical Machine Translation: Trends & Challenges / tutorial documentation. *Second International Conference on Arabic Language Resources & Tools* (p. 174).

Wetzel, D., Bond, F. (2012). Enriching parallel corpora for statistical machine translation with semantic negation rephrasing. *Proceedings of the ACL 2012 Conference: Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6 '12)* (pp. 20-29). Association for Computational Linguistics.

Analysis of the automatic statistical machine translation system model

Abstract

Automatic machine translation is an increasingly popular research topic in science and various scientific disciplines, such as information and communication sciences, computer science, computational linguistics etc. The reason for this is primarily that today it enables unavoidable communication and fast transfer of information between different natural languages. This is especially important for less spoken languages such as Croatian, for which there are still not enough software tools and digital resources that are needed for developing specialized and quality machine translation systems that would be optimized for use in one specific area. The evermore faster growth of data and the growing need of various stakeholders in the sectors of industry, economy, science, but also in peoples' everyday life imply the motivation for the systematic and organized development and subsequent adaptation of automatic machine translation systems for different language pairs. Since machine translations are not perfect, it is important to apply methods for computationally generating translations of an acceptable level of quality that depends on the task itself and the scope of implementation of the machine translation system. In this paper, the model of a system for automatic statistical machine translation, its components and the role and significance of individual elements within the model are analyzed.

Keywords: *automatic machine translation; statistical machine translation model; language technologies; natural language processing; information and communication sciences.*