PD.TVZ.HR
POLYTECHNIC & DESIGN Vol. 9, No. 2, 2021.
Mrežna inačica: ISSN 2459-6302
Tiskana inačica: ISSN 1849-1995
DOI: 10.19279/TVZ.PD.2021-9-2-03

PREGLED RAZVOJA TEHNOLOGIJE AUTOMATSKOG STROJNOG PREVOĐENJA

REVIEW OF THE EVOLUTION OF THE TECHNOLOGY OF AUTOMATIC MACHINE TRANSLATION

Ivan Dunđer

Filozofski fakultet Sveučilišta u Zagrebu, Ivana Lučića 3, Zagreb

SAŽETAK

Automatsko strojno prevođenje postalo je nezamjenjiv dio velikog broja organizacija koje posluju u međunarodnom okruženju i koje imaju potrebu generirati velike količine prijevoda za svoju dokumentaciju. Strojno prevođenje danas se smatra jednom od neizostavnih disruptivnih tehnologija koja uvelike doprinose cjelovitoj transformaciji poslovnih procesa u segmentu prevođenja tekstova napisanih na prirodnom jeziku. Ideja iza strojnog prevođenje je omogućiti automatizaciju barem dijela procesa prevođenja, posebno kada je riječ o velikoj količini podataka, ne bi li se ubrzalo cjelokupno poslovanje jedne organizacije i time se ostvarila konkurentska prednost na tržištu koje se brzo mijenja i kojemu se brzo treba prilagoditi. No, razvoj tehnologije automatskog strojnog prevođenja nije tekao tako glatko. Naime, razvoj je popraćen nizom uspona i padova, a upravo je cilj ovog znanstvenog rada dati kritičan i sistematiziran pregled svih ključnih faza razvoja navedene tehnologije, i to u kontekstu svjetskih, ali i domaćih istraživanja u tom području.

Ključne riječi: *automatsko strojno prevođenje, pristupi strojnom prevođenju, jezične tehnologije, informacijske i komunikacijske znanosti*

ABSTRACT

Automatic machine translation has become a truly irreplaceable part of a large number of organizations that operate in an international environment and in need of generating large amounts of translations for their documentation. Today, machine translation is considered one of the indispensable disruptive technologies that greatly contribute to the complete transformation of business processes in the segment of translating texts written in natural language. The idea behind machine translation is to enable the automation of at least part of the translation process, especially when it comes to a large amount of data, in order to speed up the overall business of an organization and thus gain a competitive advantage in a rapidly changing market, to which one needs to adapt quickly. But the development of automatic machine translation technology did not go so smoothly. Namely, the development is accompanied by a series of ups and downs, and the aim of this very research paper is to give a critical and systematic overview of all key stages of development of this technology, in the context of global and domestic research in this area.

Keywords: *automatic machine translation, machine translation approaches, language technologies, information and communication sciences*

1. UVOD
1. INTRODUCTION

Informacija predstavlja najvrjedniji resurs u informacijskom dobu. No, informacija mora biti na raspolaganju pravovremeno, stoga i ne čudi što su dostupnost informacije, mogućnosti pretraživanja i indeksiranja informacija pa čak i na stranom jeziku neki od imperativa suvremenog poslovanja. U današnjem globaliziranom svijetu povećava se potreba za višejezičnom komunikacijom.

Članak se nalazi u bazama: ● http://hrcak.srce.hr ● https://scholar.google.hr

Naime, bez komunikacije, bila ona interna ili ona s klijentima i partnerima, poslovanje se ne može odvijati [1], a s obzirom da se danas u svijetu govori više od 7.000 jezika [2], ne iznenađuje potreba za raznim jezičnim tehnologijama. Takve tehnologije omogućuju učinkovitije upravljanje resursima, kvalitetniju razmjenu znanja te „recikliranje" već prevedenih dokumenata i pripadajućih prijevoda. Pored toga, one nastoje povećati konzistentnost prijevoda i efikasnost rada te time smanjuju troškove prevođenja. Iako uvođenje jezičnih tehnologija zahtijeva određene financijske izdatke, njihovom kvalitetnom implementacijom se uložena sredstva relativno brzo mogu povratiti [1, 3]. Jedan od mogućih pristupa osiguravanja informacija na stranom jeziku, posebno važan za manje govorene jezike, jest primjena tehnologije za automatsko strojno prevođenje. Pod pojmom strojno prevođenje (eng. *machine translation*) podrazumijeva se primjena računala za automatiziranje (dijela) procesa prevođenja s jednog jezika na drugi [4]. Cilj specijaliziranih sustava za strojno prevođenje je što brže računalno generirati velik broj prijevoda prihvatljive kvalitete uz minimalan trošak. Razvoj takvih sustava od izuzetne je važnosti za svakodnevno poslovanje, za akademsku suradnju te napredak u gospodarstvu i industriji. Međutim, izgradnja i testiranje takvih sustava vrlo su zahtjevni zadatci [5], što se očituje i kroz brojne uspone i padove kroz koje je prošla tehnologija automatskog strojnog prevođenja. Naime, razvoj ove tehnologije prožet je različitim fazama u kojima su uspjesi i neuspjesi uvelike određeni neumornim provođenjem istraživanja entuzijastičnih znanstvenika, sve do današnjih dana. Cilj ovog znanstvenog rada jest dati kritičan i sistematiziran pregled svih relevantnih etapa razvoja tehnologije automatskog strojnog prevođenja, i to posebno u kontekstu svjetskih, ali i s osvrtom na domaća istraživanja u tom području.

2. FAZE RAZVOJA STROJNOG PREVOĐENJA
2. PHASES OF DEVELOPMENT OF MACHINE TRANSLATION

U nastavku ovoga poglavlja bit će prikazane različite faze razvoja tehnologije automatskog strojnog prevođenja, budući da je povijest strojnog prevođenja puna uspona i padova [6].

2.1. RANI RAZVOJ
2.1. EARLY DEVELOPMENT

1930-ih godina Georges Artsrouni prijavljuje prve patente za „strojeve koji prevode" (eng. *translation machines*) [7]. Radilo se o automatskom dvojezičnom rječniku zapisanom na bušenim trakama. Ruski istraživač Petr Smirnov-Troyanskij iste godine, neovisno o istraživanju Georgesa Artsrounija, prijavljuje opsežniji patent za također dvojezični rječnik koji se, međutim, oslanja i na metodu prepoznavanja gramatičkih uloga u raznim jezicima [8]. Za vrijeme Drugog svjetskog rata, matematičar Alan Turing radi na kripto-analizi Enigme, pritom koristeći elektromehaničke strojeve koji su se upotrebljavali za dekodiranje znakova. Prvo programibilno računalo ENIAC (*Electronic Numerical Integrator And Computer*) predstavljeno je javnosti 1946. te ubrzo nakon toga započinju istraživanja na području strojnog prevođenja. Među pionirima svakako treba spomenuti Warrena Weavera koji je predložio da se prevođenju pristupa računalno. 1947. u korespondenciji s kibernetičarom Norbertom Wienerom predložio je upotrebu digitalnog računala za prevođenje dokumenata između dva prirodna (ljudska) jezika [7]. Na nagovor svojih kolega, Weaver počinje intenzivno istraživati područje strojnog prevođenja, a dvije godine kasnije objavljuje svoje rezultate istraživanja u memorandumu „Translation" [9]. U memorandumu se iznose ciljevi strojnog prevođenja te potencijali digitalnih računala u prevođenju prirodnih jezika. Weaver razmatra razne pristupe strojnom prevođenju, pa tako primjerice predlaže da se kriptografske i kriptoanalitičke metode primijene u procesu prevođenja, što uvelike doprinosi stvaranju pozitivne i financijski stabilne istraživačke klime [6]. Godine 1948. Claude Shannon objavljuje rad „A Mathematical Theory of Communication", što se smatra početkom razvoja teorije informacije [10]. U središtu navedene teorije jest problematika prijenosa informacije kroz kanal sa šumom [11]. Naime, za pouzdanu komunikaciju kanalom sa šumom, komunikacija se treba odvijati u

skladu s tzv. kapacitetom kanala. Nadalje, prema Shannonu, informacija se može kvantificirati u bitove potrebne za opisivanje ishoda neizvjesnog događaja. Potreban broj bitova opisan je svojom entropijom [12]. Stoga se kao ključna mjera informacije i danas upotrebljava entropija koja mjeri količinu neizvjesnosti [6] i koja se definira kao prosječan broj bitova potrebnih za pohranjivanje ili komunikaciju jednog znaka u poruci. Treba naglasiti da su neki od ranih principa u strojnom prevođenju uspostavljenih još 1940-ih godina i danas vrlo aktualni [6]. Pa se tako primjerice još uvijek govori o dekodiranju izvornog jezika u ciljni jezik primjenom kanala sa šumom kao metodom modeliranja složenih sustava.

2.2. ZATIŠJE U RAZVOJU STROJNOG PREVOĐENJA
2.2. LULL IN THE MACHINE TRANSLATION DEVELOPMENT

Yehoshua Bar-Hillel, znanstvenik s MIT-a (*Massachusetts Institute of Technology*), prvi je istraživač koji je u punom radnom vremenu istraživao strojno prevođenje potpomognuto rječnicima. Godine 1952. organizirao je i prvu međunarodnu konferenciju o strojnom prevođenju. Kasnijih je godina izrazio sumnju u budućnost i mogućnosti strojnog prevođenja [13]. 1954. održan je Georgetown eksperiment u uredima korporacije IBM (*International Business Machines Corporation*) u New Yorku [14]. Radilo se zapravo o prvom javnom predstavljaju sustava za strojno prevođenje koji je prevodio s ruskog jezika na engleski. Međutim, radilo se o vrlo jednostavnom sustavu s ograničenim vokabularom od 250 riječi, 6 gramatičkih pravila i 49 odabranih rečenica iz područja kemije. Sam eksperiment je bio vrlo dobro medijski popraćen, što je privuklo dodatne investitore. 1950-ih i 1960-ih godina Noam Chomsky istražuje mogućnosti modeliranja znanja i jezika primjenom formalnih gramatika. Općenito, navedeno razdoblje obilježili su sustavi za strojno prevođenje temeljeni na dvojezičnim rječnicima i pravilima za održavanje ispravnog poretka riječi u rečenici. S obzirom na metode i pristupe prevođenju, tadašnji sustavi za strojno prevođenje rabili su: direktnu metodu (eng. *direct method*) – koja pomoću jednostavnih pravila uparuje riječi; metodu transfera (eng. *transfer method*) – sofisticiranija metoda koja koristi morfološku i sintaktičku analizu, te metodu međujezika (eng. *interlingua*) – koja koristi apstraktnu reprezentaciju značenja [6, 10, 15]. Godine 1964. vlada Sjedinjenih Američkih Država naredila je ispitivanje stanja i brzinu napretka područja strojnog prevođenja, s obzirom na velike financijske izdatke. Odbor ALPAC (*Automatic Language Processing Advisory Committee*) tada sastavljen od sedmero znanstvenika objavio je 1966. izvješće o stanju u području strojnog prevođenja [16]. Zaključak izvješća bio je da je strojno prevođenje skuplje, manje precizno i sporije od klasičnog ljudskog prevođenja te da nije izgledno da strojno prevođenje u skoroj budućnosti i s povećanjem financijske potpore dosegne razinu kvalitete ljudskog prevoditelja. Izvješće je bilo intonirano vrlo negativno, što je zaustavilo razvoj strojnog prevođenja ne samo u Sjedinjenim Američkim Državama, već i u Sovjetskom Savezu te Velikoj Britaniji. Istraživanja su, međutim, nastavljena u Kanadi, Francuskoj i Njemačkoj. Unatoč napuštanju strojnog prevođenja kao znanstvene discipline, u SAD-u s radom nastavljaju Peter Toma, osnivača tvrtke „Systran“ (1968.) te Bernard Scott, osnivač tvrtke „Logos“ (1970.). Od 1970. sustav Systran koristilo je američko ministarstvo obrane, a od 1976. Komisija Europske Zajednice. Iste godine se pojavljuje kanadski sustav „TAUM Météo“ (*Traduction Automatique à l'Université de Montréal*) razvijen na montrealskom sveučilištu, koji prevodi vremenske prognoze s engleskog na francuski [7]. Mogao je prevesti 80.000 riječi dnevno, tj. oko 30 milijuna riječi godišnje. Sustav TAUM Météo bio je u operativnoj upotrebi sve do 2011. godine. 1980-ih godina izlaze i drugi komercijalni sustavi za strojno prevođenje, npr. „Logos“ i „METAL“ [6]. Razdoblje 1970-ih i ranih 1980-ih godina pripada strojnom prevođenju temeljenom na pravilima.

2.3. NOVI OPTIMIZAM
2.3. NEW OPTIMISM

1980-ih i 1990-ih godina u središtu istraživanja je metoda međujezika koja nastoji formalno opisati i reprezentirati značenje neovisno o određenom jeziku.

Sustavi temeljeni na metodi međujezika bili su „CATALYST“, za prevođenje tehničkih priručnika tvrtke „Caterpillar“ te sustav „Pangloss“ [6]. U sklopu njemačkog projekta „Verbmobil“ također su razvijeni sustavi temeljeni na metodi međujezika. U to vrijeme brojne japanske tvrtke razvijale su vlastite sustave za strojno prevođenje: Brother, Fujitsu, Hitachi, Mitsubishi, NEC, Panasonic, Sanyo, Sharp, Toshiba itd. [9]. S obzirom da je prevođenje jezika vrlo teško strogo formalizirati pravilima, 1980-ih godina se pojavila potreba za novim metodama strojnog prevođenja. Umjesto da se jezik striktno opiše pravilima, postavlja se pitanje kako „učiti“ iz već prevedenih tekstova, tj. iz već viđenih primjera tekstova i pripadajućih prijevoda? Istraživanja na tom području dovela su do novih rješenja, uglavnom temeljenih na podatcima (eng. *data-driven methods*). Rani pokušaji da se iskoriste već prevedeni dijelovi rečenica rezultirali su strojnim prevođenjem temeljenim na primjerima (eng. *example-based machine translation*) [17]. Iako ova metoda tada nije uspjela zainteresirati velik broj istraživača, danas se ovaj pristup itekako primjenjuje [6]. Naime, jezična tehnologija koja ima vrlo široku primjenu u računalno-potpomognutom prevođenju jest prijevodna memorija (eng. *translation memory*), koja pohranjuje segmente ili rečenice na jednom jeziku te već ranije (ljudski) prevedene semantičke prijevodne ekvivalente na drugom jeziku [18, 19]. Kada prevoditelj prevodi novi tekst, sustav prijevodne memorije pretražit će postojeće prijevode u bazi podataka te prevoditelju ponuditi adekvatne prijevode ukoliko se novi tekst do određene razine podudara s tekstom koji se već nalazi u prijevodnoj memoriji. Kasnih 1980-ih godina dolazi do velikog skoka u razvoju strojnog prevođenja [7]. U sklopu istraživačkog projekta „CANDIDE“ (1988.), usmjerenog na prepoznavanje govora, postavljeni su matematički temelji za daljnji razvoj strojnog prevođenja. Naime, IBM-ov istraživački tim prepoznavanje govora shvaća kao statistički problem kojemu se može pristupati promatrajući veliku količinu tekstualnih podataka, tj. korpusa. U sklopu projekta razvijen je i sustav za strojno prevođenje englesko-francuskog jezičnog para [20]. U isto vrijeme, broj digitalnih tekstualnih korpusa raste, prvenstveno zbog sve većeg broja stvaratelja digitalnih dokumenata (posebno na internetu), a dijelom i zbog provođenja postupka digitalizacije fizičke dokumentacije. 1990-ih godina cijena računala opada, a procesorska snaga i računalne performanse potrebne za statističku obradu podataka rastu. Nadalje, koncept središnjeg računala (eng. *mainframe computer*) zastarijeva, te se razvoj računala preusmjerava prema osobnim računalima i radnim stanicama (eng. *workstations*), a time potencijalno raste i broj korisnika strojnog prevođenja. 1990-ih se pojavljuje i besplatni internetski prevodilački servis „BabelFish“ na mrežnoj stranici „AltaVista“, koji se temelji na Systranovoj tehnologiji strojnog prevođenja temeljeno na pravilima. Sustav je dobio ime prema „Babel fishu“, fiktivnom žutom biću koje potječe iz bestselera Douglasa Adamsa „Vodič kroz galaksiju za autostopere“ (*The Hitchhiker's Guide to the Galaxy*). Radi se o ribici koja se postavlja u uho, nakon čega za čovjeka simultano prevodi bilo koji jezik. 1999. sudionici radionice koja se održavala na Sveučilištu Johns Hopkins reimplementirali su IBM-ove metode, a novonastali programski alati postali su javno dostupni [6, 21]. Organizacija DARPA (*Defense Advanced Research Projects Agency*) prepoznaje važnost ponovno implementiranih IBM-ovih metoda te potom financira istraživačke projekte TIDES (*Translingual Information Detection Extraction and Summarization*) i GALE (*Global Autonomous Language Exploitation*).

2.4. DANAŠNJI UBRZANI RAZVOJ
2.4. TODAY'S FAST DEVELOPMENT

2000-e su godine velikog optimizma u području strojnog prevođenja. To je razdoblje sustava za statističko strojno prevođenje koje se temelji na empirijskim opažanjima [6]. Takvi sustavi se u pravilu izgrađuju za određenu domenu, tj. karakteristično područje s ograničenim vokabularom i specifičnim rečenicama, s obzirom da za uže područje namjene generiraju kvalitetnije strojne prijevode [22]. Prednosti takvog pristupa su relativno jeftina izgradnja sustava za statističko strojno prevođenje, jednostavno dodavanje novih jezika te automatizirano ugađanje sustava.

Nadalje, takvi strojni prijevodi su vrlo tečni, tj. fluentni. S druge pak strane, statističko strojno prevođenje ima mnoge poteškoće s gramatičkim aspektima jezika (vrijeme, broj, padež, slaganje itd.), a samo ugađanje sustava ovisi o brojnim faktorima i stoga nije uvijek precizno [3]. Pored toga, statistički strojni prijevodi vrlo su nepredvidivi, a ispuštanje ili neprevođenje riječi je vrlo često. Za statističko strojno prevođenje potrebne su velike količine radne memorije i jake računalne performanse [23, 24]. Veliku zaslugu u razvoju empirijskih sustava imaju i automatske metrike za evaluaciju kvalitete strojnog prijevoda [25, 26]. Danas se statističkim strojnim prevođenjem bave brojne akademske institucije, istraživački centri i privatne organizacije, poput kompanija SDL, Systran, Asia Online, IBM, Google i Microsoft. Zadnjih godina se istražuju i hibridni pristupi strojnom prevođenju: npr. u sustave za statističko strojno prevođenje ugrađuju se drugi izvori jezičnog (sintaktičkog i morfološkog) znanja ili se kombiniraju s pristupom temeljenim na pravilima [15, 27-29]. Današnji popularni sustavi temeljeni na pravilima su Apertium [30] i Systran [31]. Takvi sustavi se dobro nose sa svim gramatičkim aspektima jezika te generiraju predvidive prijevode, a uz to ne zahtijevaju veliku računalnu snagu [3]. Nadalje, prednosti takvih sustava danas su i mogućnost vrlo preciznog ugađanja gramatike te malen broj ispuštanja riječi u strojnom prijevodu. Međutim, razvijanje takvih sustava je skupo i vremenski vrlo zahtjevno, a i dodavanje novih jezika također je vrlo složeno. Ugađanje se vrši ručno, što povećava mogućnosti pogrešaka, a i strojni prijevodi su na kraju manje fluentni u odnosu na statističke prijevode [3]. Danas se istražuju i mogućnosti strojnog prevođenja tipa govor-u-govor [32, 33] te mogućnosti integriranja sustava za strojno prevođenje u radni tijek organizacije [34, 35], što može biti izrazito praktično, jer se razni poslovni procesi na taj način mogu ubrzati i optimizirati. Integracija prijevodnih memorija u sustave za strojno prevođenje također je od istraživačkog interesa [3, 36, 37], jer se na taj način može povećati razina konzistentnosti korištene terminologije u prijevodima. Budući da je za izgradnju sustava za statističko strojno prevođenje neizbježna velika količina dvojezičnih tekstova, istražuju se mogućnosti prikupljanja i izrade paralelnih korpusa pomoću crowdsourcinga kao jedne od mogućnosti oslanjanja na mnoštvo [38]. Komercijalne platforme crowdsourcinga, kao npr. mehanički Turčin (eng. *mechanical Turk*), koriste se i za potrebe evaluacije kvalitete strojnog prijevoda [39, 40] – na takvoj platformi ljudi su plaćeni za odrađeni mikro-posao, poput ljudske provjere kvalitete strojnog prijevoda. Unatoč tome što danas postoji nekoliko različitih pristupa strojnom prevođenju, uz statističko strojno prevođenje sve popularnije je i neuralno strojno prevođenje, koje svojom kvalitetnom može nadmašiti ostale pristupe. To je najnoviji pristup automatskom prevođenju i podrazumijeva strojno prevođenje temeljeno na umjetnoj inteligenciji i tehnici dubokog učenja (eng. *deep learning*) [41]. Takav pristup koristi umjetnu neuralnu mrežu (eng. *artificial neural network*) da bi predvidio nizove riječi te ne zahtijeva zasebno strojno učenje značajki modela kao što je to slučaj kod statističkog strojnog prevođenja [42].

3. PERCEPCIJA KVALITETE STROJNOG PRIJEVODA
3. PERCEPTION OF MACHINE TRANSLATION QUALITY

Glavne kritike na račun strojnog prevođenja danas su usmjerene prema opasnostima zamjene ljudskih prevoditelja te kvaliteti i (domenskim) ograničenjima strojnog prevođenja [3]. Međutim, ideja strojnog prevođenja nije zamijeniti čovjeka, već asistirati mu pri prevođenju. Ono što se mijenja su samo brzina te tijek i aktivnosti unutar procesa prevođenja. S obzirom da je danas količina teksta koju treba prevesti prevelika za čovjeka, računalo može generirati prijevode koje zatim čovjek samo dorađuje, umjesto da ih prevodi ispočetka. Mnogi su izrazito skeptični kada govore o kvaliteti strojnih prijevoda. Prosudbe se vrlo često donose na temelju opaženih performansi raznih prevodilačkih servisa na internetu. No, milijuni ljudi svakodnevno koriste strojno prevođenje, najčešće kako bi dobili osnovnu informaciju o sadržaju koji ne razumiju [6]. Pored toga, takvi servisi su u pravilu besplatni i stoga dostupni zainteresiranoj javnosti. Svakako ovdje treba naglasiti da su sustavi za strojno prevođenje koji se po narudžbi izgrađuju za određenu organizaciju znatno drugačiji i često daleko kvalitetniji.

Naime, takvi sustavi su posebno prilagođeni radnom okruženju i potrebama organizacije te stoga postižu daleko bolje i kvalitetnije strojne prijevode. No, kritičari su u pravu kada tvrde da jedino čovjek može „ispravno" razumjeti profinjenost i suptilnost jezika i kulture, te stoga ispravno prevesti razne oblike teksta (barem za sada). Kompleksnost jezika izrazito otežava strojno prevođenje, međutim, ispostavilo se da postoji vrlo velik broj tipova uobičajenih rečenica i tekstova s kojim računalo ima manje poteškoća pri prevođenju [3]. Svakako se pritom ne misli na poeziju ili sofisticirane prijevode, već na tekstove s ustaljenim vokabularom, poretkom riječi i jasnim te jednoznačnim jezikom bez metaforičkog značenja. To se primjerice može odnositi na brošure ili korisničke upute za korištenje raznih proizvoda (pogodno za lokalizaciju), prevođenje dinamičkog sadržaja na internetu, filmskih titlova, vremenskih prognoza, televizijskih ili radijskih vijesti, korisničkog sadržaja (komentari na mrežnim stranicama, SMS-ovi, poruke na chatu ili društvenim mrežama), na praćenje informacija iz stranih izvora ili na generiranje grubih prijevoda za osnovno razumijevanje informacije [3].

4. RAZVOJ STROJNOG PREVOĐENJA ZA HRVASKI JEZIK

4. DEVELOPMENT OF MACHINE TRANSLATION FOR THE CROATIAN LANGUAGE

Za potrebe hrvatskog jezika razvijeno je više sustava za strojno prevođenje te je provedeno opsežno istraživanje u području metoda računalne adaptacije domene [43]. Strojno prevođenje istraženo je za različite svrhe i s osvrtom na različite domene prema kojima se sustav treba optimizirati, poput domene poezije [44] koja obiluje vrlo specifičnim problemima, poput polisemije, složenih anafora, pjesničkog izražavanja i sl. Evaluacija domene poezije je također izvršena s posebnim osvrtnom na ljudsku evaluaciju kvalitete strojnog prijevoda [45], pri čemu je uočeno da relativno slobodan poredak referentnih prijevoda i stil prevođenja mogu otežati proces evaluacije kvalitete. Specijaliziran sustav za domenu industrije također je nedavno razvijen [46] uz pomoć suboptimalnog računalnog korpusa s relativno velikom količinom šuma, a koji je sadržavao slobodne ili neodgovarajuće prijevode, manjak prijevoda na jednoj strani paralelnog korpusa, nepodudarna pisma i/ili jezike i sl. Analizirane su i mogućnosti osiguranja kvalitete prijevoda generiranih pomoću alata za računalno-potpomognuto prevođenje u poslovnom okruženju [47], pri čemu su ispitani metodološki okvir za analizu osiguranja kvalitete te opravdanost uvođenja samog procesa osiguravanja kvalitete u CAT procesu, i to na temelju komparativne evaluacija različitih alata za osiguranje kvalitete te primjene tzv. MQM (Multidimensional Quality Metrics) okvira za kategoriju pogrešaka. Automatska evaluacija kvalitete strojnog prijevoda ispitana je u domeni sociologije, filozofije i religioznosti [48] primjenom metrika BLEU, NIST, METEOR i GTM, pri čemu je potvrđeno da je kvaliteta strojnih prijevoda bolja u slučaju prevođenja s jezika veće morfološke složenosti na jezik manje složenosti. Razlog leži prvenstveno u duljim rečenicama, morfološkom bogatstvu jezika, relativnom slobodnom poretku riječi i gramatičkim slaganjem padeža. Automatska evaluacija je izvršena i za različite jezične parove [49] – u ovom konkretnom slučaju je potvrđeno da je kvaliteta strojnog prijevoda bolja za srodne jezike, poput hrvatskog-ruskog jezičnog para, što je koreliralo i s provedenom ljudskom evaluacijom kvalitete. Ljudska evaluacija prijevoda za hrvatski jezik i primjena online servisa za strojno prevođenje također su ispitani u jednom istraživanju [50], pri čemu je utvrđeno da se lakše postiže adekvatnost nego fluentnost strojnog prijevoda, a to je potvrđeno Cronbach alphom, tj. mjerom unutarnje konzistentnosti evaluatora kojom se utvrđuje razina (ne)slaganja evaluatora. Integracija strojnog prevođenja i automatskog prepoznavanja govora također je analizirana za hrvatski jezik u domeni poslovne korespondencije [51]. Potvrđeno je da se sustavi za automatsko prepoznavanje govora znatno optimiziraju u slučaju da se treniraju za posebnu domenu i svrhu, a metrika PER (Position-independent word Error Rate) se pokazala prikladnijom u odnosu na metriku WER (Word Error Rate) za hrvatski jezik koji kao takav dopušta relativno slobodan poredak riječi u prijevodima.

Istražene su i brojne mogućnosti za pripremu digitalnih resursa potrebnih za razvoj sustava za strojno prevođenje, poput digitalizacije dokumenata klasičnim stolnim skenerom i optičkog prepoznavanja znakova te naknadnog post-editiranja pogrešno prikazanih znakova [52], ili primjene crowdsourcinga [53] u nastavi jezičnih tehnologija ili računalne obrade prirodnog jezika, pri čemu se studenti uključuju u rad na posebno izrađenoj gamifikacijskoj platformi koja nagrađuje angažman studenata prilikom prikupljanja odgovarajućih resursa, kao što su jednojezični ili paralelni korpusi.

5. ZAKLJUČAK
5. CONCLUSION

Strojno prevođenje podrazumijeva primjenu računala radi automatiziranja (dijela) procesa prevođenja s jednog prirodnog jezika na drugi. Namjera sustava za strojnog prevođenje je omogućiti generiranje velikog broja prijevoda prihvatljive kvalitete s malo troška. No, takva tehnologija vrlo je složena i može se danas razviti i implementirati različitim pristupima. U najnovije vrijeme oslanja se na različite oblike umjetne inteligencije i metode strojnog učenja. Brojne faze razvoja tehnologije strojnog prevođenja prožete su različitim problemima, entuzijazmom i razočarenjima. Upravo su one bile predmet kritičkog istraživanja u sklopu ovog znanstvenog rada. Automatsko strojno prevođenje danas je općeprihvaćena tehnologija u suvremenim organizacijama, i poslovanje bi bilo nezamislivo bez takvog oblika jezične tehnologije. Važno je stoga nastaviti istraživati ovu tehnologiju te posebno poduprijeti razvoj alata i resursa za hrvatski jezik.

6. REFERENCE
6. REFERENCES

[1.] Dillinger M.; Introduction to MT / tutorial documentation; The Ninth Conference of the Association for Machine Translation in the Americas, p. 29, 2010.

[2.] Stüker S.; Waibel A.; Towards human translations guided language discovery for ASR systems; The first International Workshop on Spoken Languages Technologies for Under-resourced languages (SLTU - 2008), pp. 76-79, 2008.

[3.] Dillinger M.; Marciano J.; Introduction to MT / tutorial documentation; The Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2012), p. 47, 2012.

[4.] Folajimi Y. O.; Omonayin I.; Using Statistical Machine Translation (SMT) as a Language Translation Tool for Understanding Yoruba Language; EIE's 86 2nd Intl' Conf. Comp., Energy, Net., Robotics and Telecom. (eieCon2012), pp. 86-91, 2012. DOI: 10.13140/2.1.3522.8485.

[5.] Unnikrishnan P.; Antony P. J.; Soman K. P.; A Novel Approach for English to South Dravidian Language Statistical Machine Translation System; International Journal on Computer Science and Engineering (IJCSE), Vol. 2, No. 8, pp. 2749-2759, 2010. e-ISSN: 0975–3397.

[6.] Koehn P.; Statistical Machine Translation; Cambridge University Press. ISBN-13: 978-0521874151, 2010.

[7.] Hutchins J.; Machine translation over fifty years; Histoire, Epistémologie, Langage: Le traitement automatique des langues, Vol. 23, No. 1, pp. 7-31, 2001. e-ISSN: 1638-1580.

[8.] Hutchins J.; The Georgetown-IBM experiment demonstrated in January 1954. Machine translation: from real users to research; 6th conference of the Association for Machine Translation in the Americas (AMTA 2004), pp. 102-114, 2004. DOI: 10.1007/978-3-540-30194-3_12.

[9.] Chéragui M.; A Theoretical Overview of Machine translation; Proceedings of the 4th International Conference on Web and Information Technologies (ICWIT 2012), pp. 160-169, 2012.

[10.] Shannon C. E.; A Mathematical Theory of Communication; Bell System Technical Journal, Vol. 27, No. 3, pp. 379-423. DOI:10.1002/j.1538-7305.1948.tb01338.x.

[11.] Specia L.; Fundamental and New Approaches to Statistical Machine Translation / tutorial documentation; Propor 2010 - International Conference on Computational Processing of the Portuguese Language, p. 32, 2010.

[12.] Manning C. D.; Schütze H.; Foundations of Statistical Natural Language Processing; The MIT Press, p. 620, 1999. ISBN-13: 978-0262133609.

[13.] Hutchins J.; Machine Translation and Human Translation: In Competition or in Complementation?; International Journal of Translation, Vol.13, No. 1-2, pp. 5-20, 2001.

[14.] Hutchins J.; Has machine translation improved? some historical comparisons; Proceedings of the MT Summit IX, Association for Machine Translation in the Americas, p. 8, 2003.

[15.] Stein D.; Machine Translation - Past, Present, and Future; Translation: Computation, Corpora, Cognition (TC3), Vol. 3, No. 1, pp. V-XII, 2013. ISSN: 2193-6986.

[16.] Hutchins J.; ALPAC: the (in)famous report; MT News International, No. 14, pp. 9-12, 1996.

[17.] Hutchins J.; Towards a definition of example-based machine translation; Proceedings of the Tenth Machine Translation Summit (MT Summit X), Asia-Pacific Association for Machine Translation, Thai Computational Linguistics Laboratory (NICT), p. 8, 2005.

[18.] Reinke U.; State of the Art in Translation Memory Technology; Translation: Computation, Corpora, Cognition. Special Issue on Language Technologies for a Multilingual Europe, Vol. 3, No. 1, pp. 27-48, 2013. ISSN: 2193-6986.

[19.] Baldwin T.; Translation Memory Engines: A Look under the Hood and Road Test; Proceedings of the 15th International Japanese/English Translation Conference, p. 19, 2004.

[20.] Berger A. L.; Brown P. F.; Della Pietra S. A.; Della Pietra V. J. ; Gillett J. R.; Lafferty J. D.; Mercer R. L.; Printz H.; Ureš L.; The Candide System for Machine Translation; Proceedings of the HLT '94 Workshop on Human Language Technology, ACL, pp. 157-162, 1994.

[21.] Al-Onaizan Y.; Curin J.; Jahr M.; Knight K.; Lafferty J.; Melamed D.; Och F.-J.; Purdy D.; Smith N. A.; Yarowsky D.; Statistical Machine Translation; Final Report of the JHU Summer Workshop, p. 42, 1999.

[22.] Koehn P.; Haddow B.; Towards Effective Use of Training Data in Statistical Machine Translation; Proceedings of the 7th Workshop on Statistical Machine Translation, ACL, pp. 317-321, 2012.

[23.] Turchi M.; De Bie T.; Goutte C.; Cristianini N.; Learning to Translate: A Statistical and Computational Analysis; Advances in Artificial Intelligence, Vol. 2012, p. 15, 2012. DOI: 10.1155/2012/484580.

[24.] Turchi M.; De Bie T.; Cristianini N.; Learning Performance of a Machine Translation System: a Statistical and Computational Analysis; Proceedings of the Third Workshop on Statistical Machine Translation, ACL, pp. 35-43, 2008. DOI: 10.3115/1626394.1626399.

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

[26.] Giménez J.; Màrquez L.; Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation; The Prague Bulletin of Mathematical Linguistics, No. 94, pp. 77-86, 2010. DOI: 10.2478/v10108-010-0022-6.

[27.] Okpor M. D.; Machine Translation Approaches: Issues and Challenges; IJCSI International Journal of Computer Science Issues, Vol. 11, No. 5/2, pp. 159-165, 2014. ISSN: 1694-081.

[28.] Costa-jussà M. R.; Banchs R. E.; Rapp R.; Lambert P.; Eberle K.; Babych B.; Workshop on Hybrid Approaches to Translation: Overview and Developments; Proceedings of the 2nd HyTra Workshop, ACL, pp. 1-6, 2013.

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

[30.] Tyers F. M.; Sánchez-Martínez F.; Sánchez-Martínez O.; Forcada M. L.; Free/Open-Source Resources in the Apertium Platform for Machine Translation Research and Development; The Prague Bulletin of Mathematical Linguistics, No. 93 pp. 67-76, 2010. DOI: 10.2478/v10108-010-0015-5.

[31.] Hutchins J.; The Development and Use of Machine Translation Systems and Computer-based Translation Tools; International Journal Of Translation, Vol. 15, No. 1, pp. 5-26, 2003.

[32.] Hutchins J.; Multiple Uses of Machine Translation and Computerised Translation Tools; Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages (ISMTCL 2009), pp. 13-20, 2009.

[33.] Black A. W.; Brown R. D.; Frederking R.; Singh R.; Moody J.; Steinbrecher E.; TONGUES: rapid development of a speech-to-speech translation system; Proceedings of the second international conference on Human Language Technology Research (HLT '02), pp. 183-186, 2002.

[34.] Vilar D.; Schneider M.; Burchardt A.; Wedde T.; Towards the Integration of MT into a LSP Translation Workflow; Proceedings of the 16th EAMT Conference, pp. 73-76, 2012.

[35.] Sun Y.; Liu J.; Li Y.; Deploying MT into a Localisation Workflow: Pains and Gains; Proceedings of the 13th Machine Translation Summit, pp. 236-243, 2011.

[36.] Wang K.; Zong C.; Su K.-Y.; Integrating Translation Memory into Phrase-Based Machine Translation during Decoding; Proceedings of the 51st Annual Meeting of the ACL, ACL, pp. 11-21, 2013.

[37.] Kanavos P.; Kartsaklis D.; Integrating Machine Translation with Translation Memory: A Practical Approach; Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry" (JEC '10), pp. 11-20, 2010.

[38.] Post M.; Callison-Burch C.; Osborne M.; Constructing Parallel Corpora for Six Indian Languages via Crowdsourcing; Proceedings of the 7th Workshop on Statistical Machine Translation, ACL, pp. 401-409, 2012.

[39.] Bojar O.; Buck C.; Callison-Burch C.; Federmann C.; Haddow B.; Koehn P.; Monz C.; Post M.; Soricut R.; Specia L.; Findings of the 2013 Workshop on Statistical Machine Translation; Proceedings of the Eighth Workshop on Statistical Machine Translation, ACL, pp. 1-44, 2013.

[40.] Callison-Burch C.; Koehn P.; Monz C.; Post M.; Soricut R.; Specia L.; Findings of the 2012 Workshop on Statistical Machine Translation; Proceedings of the 7th Workshop on Statistical Machine Translation, ACL, pp. 10-51, 2012. ISBN: 978-1-937284-20-6 / 1-937284-20-4.

[41.] Kamath U.; Liu J.; Whitaker J.; Deep Learning for NLP and Speech Recognition; Springer, p. 621, 2019. ISBN: 978-3-030-14596-5.

[42.] Cho K.; van Merriënboer B.; Bahdanau D.; Bengio Y.; On the Properties of Neural Machine Translation: Encoder–Decoder Approaches; Proceedings of the SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pp. 103-111, 2014. DOI: 10.3115/v1/W14-4012

[43.] Dunđer I.; Statistical Machine Translation System and Computational Domain Adaptation / doctoral dissertation; University of Zagreb, Zagreb, 2015.

[44.] Dunđer I.; Seljan S.; Pavlovski M.; Automatic Machine Translation of Poetry and a Low-Resource Language Pair; 43rd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2020), pp. 1034-1039, 2020. DOI:10.23919/MIPRO48935.2020.9245342.

[45.] Seljan S.; Dunđer I.; Pavlovski M.; Human Quality Evaluation of Machine-Translated Poetry; 43rd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2020), pp. 1040-1045, 2020. DOI: 10.23919/MIPRO48935.2020.9245436.

[46.] Dunđer I.; Machine Translation System for the Industry Domain and Croatian Language; Journal of Information and Organizational Sciences (JIOS), Vol. 44, No. 1, 2020. DOI: 10.31341/jios.44.1.2.

[47.] Seljan S.; Škof Erdelja N.; Kučiš V.; Dunđer I.; Pejić Bach M.; Quality Assurance in Computer-Assisted Translation in Business Environments; Natural Language Processing for Global and Local Business, IGI Global, p. 22, 2020.
DOI: 10.4018/978-1-7998-4240-8.ch011.

[48.] Seljan S.; Dunđer I.; Automatic Quality Evaluation of MachineTranslated Output in Sociological-Philosophical-Spiritual Domain; Proceedings of the10th Iberian Conference on Information Systems and Technologies (CISTI'2015), Vol. 2, pp. 128-131, 2015.
ISBN: 978-989-98434-5-5.

[49.] Seljan S.; Dunđer I.; Machine Translation and Automatic Evaluation of English/Russian-Croatian; Proceedings of the International Conference "Corpus Linguistics – 2015" (CORPORA 2015), pp. 72-79, 2015. ISBN: 978-5-8465-1498-0.

[50.] Seljan S.; Tucaković M.; Dunđer I.; 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, 2015. DOI:10.1007/978-3-319-16486-1_108

[51.] Seljan S.; Dunđer I.; 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, 2014. DOI: 10.5281/zenodo.1096693

[52.] Seljan S.; Dunđer I.; Gašpar A.; 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, 2013. DOI: 10.13140/RG.2.1.1883.9521.

[53.] Jaworski R.; Seljan S.; Dunđer I.; 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, 2017.
ISBN: 978-83-64864-94-0.

AUTOR · *AUTHOR*

● Ivan Dunđer
Inženjer je informacijske tehnologije i docent zaposlen na Katedri za informatiku Odsjeka za informacijske i komunikacijske znanosti pri Filozofskom fakultetu Sveučilišta u Zagrebu. Znanstveno-istraživački je angažiran na području obrade prirodnog jezika, strojnog prevođenja i evaluacije, jezičnih i govornih tehnologija, upravljanja znanjem te modeliranja i razvoja baza podataka, aplikacija i informacijskih sustava. Izlagao je na brojnim međunarodnim konferencijama s međunarodnom recenzijom te se znanstveno i stručno usavršavao na seminarima, certificiranim programima edukacije, javnim tribinama i izlaganjima, okruglim stolovima, na konferencijama i radionicama u Hrvatskoj i inozemstvu. Objavio je više od 60 znanstvenih radova te sudjelovao u brojnim znanstveno-istraživačkim projektima. Aktivno participira u radu znanstvenih i stručnih udruženja te govori njemački, engleski i francuski jezik.

Korespondencija · *Correspondence*
ivandunder@gmail.com