Pregledni rad
https://doi.org/10.31784/zvr.9.1.23
Datum primitka rada: 17. 11. 2020.
Datum prihvaćanja rada: 20. 1. 2021.

PREGLED METODA OBRADE PRIRODNIH JEZIKA I STROJNOG PREVOĐENJA

Sabrina Šuman
Dr. sc., viša predavačica, Veleučilište u Rijeci, Vukovarska 58, 51 000 Rijeka, Hrvatska;
e-mail: ssuman@veleri.hr

SAŽETAK

U radu je dan pregled područja povezanog s procesiranjem prirodnih jezika i njihova međusobnog odnosa, počevši od šire domene kao što je umjetna inteligencija, putem strojnog učenja, računalne lingvistike, metoda strojnog prevođenja te posebice onih zasnovanim na dubokom učenju. Opisane su karakteristike, primjene, faze i glavni problemi obrade prirodnih jezika s leksičke, sintaktičke, semantičke, govorne i pragmatičke perspektive. Opisane su faze prepoznavanja i analize prirodnog jezika kao i faza generiranja prirodnih jezika. Postupci pre-editinga i post-editinga uz korištenje kontroliranih prirodnih jezika dani su kao primjeri prakse kojom se povećava točnost i kvaliteta automatskog prevođenja i općenito procesiranja teksta. Poseban je fokus stavljen na strojno prevođenje te metode strojnog prevođenja. Pristupi strojnom prevođenju kao statistički, temeljen na pravilima, hibridni i pristup temeljen na dubokom učenju opisani su i predstavljeni s obzirom na njihove prednosti i nedostatke i prikladnu primjenu u praksi. Na kraju su dani još uvijek neriješeni izazovi kao smjer daljnjih istraživanja vezanih uz obradu prirodnih jezika te značaj razvoja pristupa temeljenog na dubokom učenju.

Ključne riječi: *obrada prirodnog jezika, strojno prevođenje, računalna lingvistika, duboko učenje, umjetna inteligencija*

1. UVOD

Umjetna inteligencija (engl. *Artificial intelligence* – AI u daljnjem tekstu) može se definirati kao znanstveno područje koje istražuje načine kako postići da se računalo inteligentno ponaša. AI istražuje računalno modeliranje ljudske inteligencije koja obuhvaća mnoge sposobnosti među kojima se ističe sposobnost opažanja i sposobnost razumijevanja i generiranja jezika (Silva, Fonseca, 2019). Inteligentno ponašanje uvjetovano je znanjem koje se može opisati kao ljudsko djelovanje temeljeno na onome što se zna (ili vjeruje) o svijetu (Jordan, Russell, 2001; Brachman, Levesque, 2004). Ljudsko je znanje raspoređeno i pohranjeno na različite načine – tekst, zvuk, slike ili bilo što čime se može pohraniti znanje (Gottschalk-Mazouz, 2007), a tijekom procesa učenja ono postaje osobno, ugrađeno i pohranjeno u ljudski um. Različitim kognitivnim potrebama i funkcijama (npr.

razmišljanje, učenje, motivacija, emocije) pokreću se određena pohranjena znanja kojima se tada nadograđuju prošla ili kreiraju nova znanja koja se koriste za procese zaključivanja i donošenja odluka. Cilj je i težnja da i neki AI sustav ima slične karakteristike i način funkcioniranja – da bude sposoban naučiti znanje pohranjeno u različitim nositeljima znanja. Tako *naučeno* znanje postaje personalizirano, pohranjeno i dostupno umjetnom kognitivnom procesu.

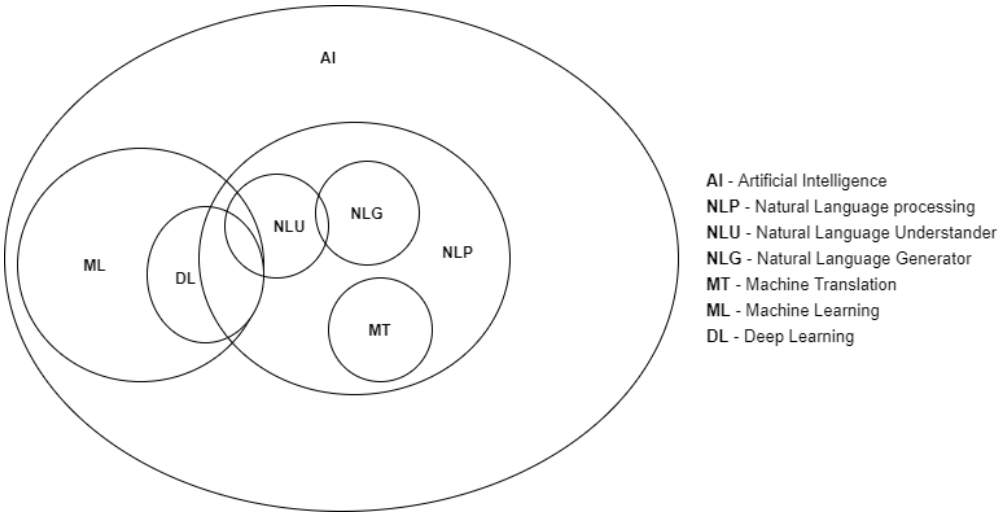
Procesiranje ili obrada prirodnih jezika (engl. *Natural language processing*, NLP u daljnjem tekstu) dio je umjetne inteligencije te predstavlja fokus ovog rada.

Cilj ovoga rada je dati pregled područja povezanih sa NLP-om te njihovih međusobnih odnosa. Svrha rada je predstavljanje metoda koje se koriste u obradi prirodnih jezika kroz specifikaciju njihovih prednosti i nedostataka kao i opisa mogućih primjena. Posebno je dan naglasak na faze i metode strojnog prevođenja (engl. *Machine translation*, MT u daljnjem tekstu).

Motivacija za ovo istraživanje proizašla je iz težnje da se sintetski i sistematski predstave temeljna povezana područja NLP-a budući da postoji veliki broj radova u kojem se detaljno opisuju pojedine AI metode i algoritmi a nema puno recentnih istraživanja koja obuhvaćaju pregled povezanih područja. Predstavljene su glavne značajke i razine obrade prirodnih jezika – od leksičke do pragmatičke razine, kao i temeljna problematika ovog područja.

U nastavku je dan pregled problematike kod različitih faza procesiranja prirodnih jezika kao i pregled metoda i tehnika koje se koriste u području strojnog prevođenja. Kako će se obuhvatiti različiti pojmovi tijekom ovog pregleda, radi boljeg razumijevanja, na slici 1 dan je kontekst i odnos različitih sastavnica NLP područja.

Slika 1. Prikaz odnosa pojmova povezanih s NLP-om



Izvor: autorica

2. OBRADA PRIRODNOG JEZIKA – NLP

NLP dio je područja umjetne inteligencije koje je povezano s lingvistikom. Budući da računalo „ne razumije" prirodne jezike, postoji potreba za obradom prirodnog jezika (Manning, Schutze, 1999). NLP istražuje načine uporabe računala za obradu ili razumijevanje ljudskih – prirodnih jezika. Upotrebljava se za pretvaranje ili prevođenje podataka s prirodnog jezika na računalu razumljiv jezik – strojno razumljiv format. Nakon što procesiraju prirodni jezik, računala mogu komunicirati jezikom koji upotrebljavaju ljudi.

Gledano sa znanstvene perspektive, cilj NLP-a jest modeliranje kognitivnih mehanizama na kojima se temelji razumijevanje i produkcija ljudskih jezika. Iz inženjerske perspektive, NLP obuhvaća razvoj različitih praktičnih aplikacija za olakšavanje interakcije između računala i ljudskih jezika. Uobičajene aplikacije u NLP-u uključuju raspoznavanje govora (engl. *Speech recognition*), leksičku i semantičku analizu jezika, strojno prevođenje (engl. *Machine translation*, MT), automatsku sumarizaciju (engl. *Automatic summarization*), analizu mišljenja (engl. *Sentiment analysis*), dohvaćanje informacija (engl. *Information retrieval*), odgovaranje na pitanja i dr. (Deng, Liu, 2018).

Pojam računalna lingvistika (engl. *Computational linguistics*, CL u daljnjem tekstu) često se smatra sinonim za NLP iako se razlikuju po predmetu istraživanja. CL je znanstvena disciplina koja proučava lingvističke procese iz računalne perspektive dok NLP koristi računala za stvaranje korisnih aplikacija s jezikom (Johnson, 2012).

Istraživački su napori kod CL-a usmjereni na otkrivanje kako su nizovi riječi povezani s njihovim značenjima te na razvoj formalizama pomoću kojih su opisane te povezanosti. Fokus je kod CL-a na jeziku, a računalni su algoritmi u funkciji potpore.

Kod NLP-a fokusira se na analizu, dizajn i primjenu računalnih algoritama i načina reprezentiranja za obradu prirodnih jezika (Eisenstein, 2019). Ciljevi su u NLP-u realizacija efikasnih algoritama, analiziranje rečenične strukture i/ili značenja zadane rečenice (Tsujii, 2011). NLP istražuje mogućnost primjene računalnih algoritama na zadaće kao što su ekstrakcija informacija iz teksta (*Information extraction*), prevođenje jednog jezika na drugi, mogućnost automatizirana odgovora na pitanja, razumijevanja govornih ili tekstualnih naredbi, vođenja razgovora na prirodnom jeziku i sl (Šuman, 2019).

Današnji pristupi NLP-u često koriste strojno učenje (engl. *Machine learning*, u daljnjem tekstu ML), kojim se omogućava stvaranje složenih programa na temelju ulazno-izlaznih parova primjera. U slučaju NLP-a strojno učenje, pojednostavljeno rečeno, različitim tehnikama, niz riječi-tokena iz jednog rječnika pretvara u niz riječi-tokena iz drugog rječnika, što se tada naziva prevođenjem jednog jezika na drugi (engl. *Translation*) (Eisenstein, 2019).

2. 1 Faze i razine obrade

Problem i najveći izazov kod NLP-a jest dvosmislenost na više razina: značenju riječi, morfologiji, sintaktičkim svojstvima i ulogama i vezama između dijelova teksta. Moguće se dvosmislenosti događaju kod pojedinih riječi, dijelova rečenica i cijelih rečenica, a pogotovo kada se trebaju obraditi duže, gramatički složene rečenice (Poibeau, 2017). Ljudi rješavaju problem dvosmislenosti

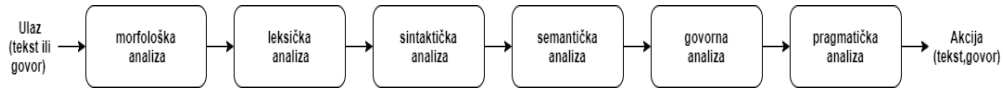
uzimajući u obzir širi kontekst, iskustvo i prethodno znanje, ali ni tada ne mogu izbjeći probleme u komunikaciji. Računala problem dvosmislenih značenja također rješavaju slično kao ljudi: uzimajući širi kontekst oko riječi i zaključivanje na temelju prošlih slučajeva (Šuman, 2019).

Stoga se prevođenje i/ili interpretiranje temelje na vjerojatnosnu zaključivanju o pravom značenju koje nije direktno sadržano u izvornom tekstu, a uključuje znanje iz konteksta i domene. Kako bi uspješno izveo svoje zadaće, NLP proces izvodi se u nekoliko koraka/faza u procesu obrade u ovisnosti o razinama jezika (Liddy, 2003) – od morfologije riječi do leksičkih, sintaktičkih i semantičkih aspekata teksta i pragmatičnih svojstava teksta prirodnog jezika. Neki su sustavi fokusiraniji na niže razine obrade, a drugi na više ili sve razine. Što više razina postoji, sustav je složeniji, ali i podložniji pogreškama. Opći model razumijevanja prirodnih jezika prikazan je na slici 2.

Sljedeće razine NLP područja izdvajaju se te međusobno isprepliću, a prikazane su i na slici 2:

- *morfološka razina* – ako se radi o pisanom jeziku-tekstu, u općem, tradicionalnom modelu razumijevanja prirodnog jezika (Dovedan, 1993), prva faza analize jest morfološka analiza. Morfem je najmanja jedinica riječi koja nosi neko značenje. Prema Jurafsky, Martin, 2008, pojmovi i afiksi tipovi su morfema. *Stem* je primarni dio riječi te daje neko značenje (npr. igra). Dio riječi koje joj daje dodatno značenje jest afiks. Afiksi su najčešće sufiksi (npr. *-nje* u riječi igranje) ili prefiksi (na primjer, *do-* u riječi doigravanje);
- *leksička razina* **–** na ovoj se razini provodi analiza u kojoj se riječ svrstava u leksičke kategorije (na primjer imenica, glagol, pridjev i sl.) i određuju obilježja kao što je, na primjer, jednina ili množina.
- *sintaktička razina* – sintaksa je znanost o rečenicama, a sintaktička analiza dijeli rečenice na sastavne dijelove. Sintaksa je primarno usredotočena na pravila i uvjete koji se definiraju gramatikom te oblikuju rečenicu;
- *semantička razina* – sintaktička je analiza povezana sa strukturom rečenica, dok semantička razina otkriva značenje u tim rečenicama (Feldman, 1999; Rajesh, Reddy Lokanatha 2009). Sa semantičkog gledišta većina riječi ima različita značenja i određivanje točnog značenja za neku riječ ovisi o kontekstu, odnosno poziciji riječi u rečenici. Semantičkom se analizom rečenica prevodi u oblik pogodan za donošenje zaključaka. Sintaktička i semantička razina nedjeljive su, međusobno djeluju jedna na drugu te su u obradi prirodnog jezika te dvije razine duboko isprepletene, u identifikaciji i upotrebi (Rajesh, Reddy i Lokanatha, 2009);
- *govorna razina* – govorna razina upotrijebljena je u NLP-u za određivanje značenja rečenica, u ovisnosti o ostalim rečenicama iz određenog teksta ili odlomka nekog dokumenta (Feldman, 1999; Rajesh, Reddy i Lokanatha, 2009);
- *pragmatička razina* – ova razina istražuje rečeničnu analizu i mogućnosti različitog iskorištavanja u različitim situacijama – jasno određuje kontekst rečenice, kao što je odnos između vremena kada je rečenica izgovorena i vremena na koje se odnosi (Rajesh, Reddy i Lokanatha, 2009).

Slika 2. Opći model razumijevanja prirodnih jezika



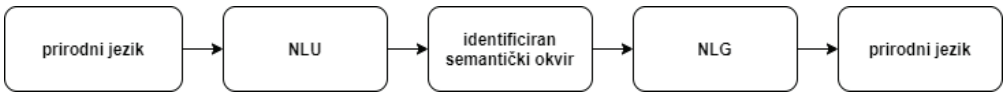
Izvor: autorica

2. 2 Prepoznavači i generatori prirodnih jezika

NLU (engl. Natural Language Understander), odnosno prepoznavač prirodnih jezika i NLG (engl. Natural Language Generator) ili generator prirodnih jezika dijelovi su domene i istraživačke tematike NLP-a. Predstavljaju različite dijelove procesa kojim se nastoji reproducirati ljudska komunikacija i ponašanje putem računala. NLU, kao dio NLP-a, rastavlja jezik (tekst ili govor) na manje dijelove, tzv. part of speech, ili još manje, morpheme, te nastoji "razumjeti" pravo značenje teksta/govora analizirajući ga sintaktički i semantički na temelju gramatike, leksika te različitih algoritama i mehanizama za izvođenje tih analiza. Nakon što je računalo/uređaj „razumjelo" tekst ili govor (što predstavlja neki ulaz), potrebno je da ono stvori ljudima „razumljiv" odgovor. NLG proces je kreiranja smislenih fraza i rečenica na prirodnom jeziku. Može se reći da NLU nastoji „pročitati" tekst ili „čuti" govor s razumijevanjem dok NLG nastoji ispisati govor, odgovoriti na njega ili ga reproducirati na ljudski razumljivu jeziku. NLU pretvara tekst ili govor (koji pripadaju nestrukturiranim tipovima podataka) u strukturirane podatke (semantički okvir) dok NLG kreira tekst ili govor na temelju strukturiranih podataka (slika 3). NLU nastoji pronaći semantičko značenje u danom tekstu, a NLG ima za cilj stvoriti odgovarajuće rečenice na temelju dane semantike (Su i sur., 2018; Su i sur., 2019).

NLG primjenu ostvaruje u različitim područjima kao što su sumarizacija i personaliziranje izvještaja, *chatbotovi*, personalizirana komunikacija s klijentima, ispis statusa uređaja u IoT okruženju (engl. *Internet of things*) (Sciforce, 2019).

Slika 3. Ciklus razumijevanja i generiranja prirodnog jezika



Izvor: obrada autorice na temelju Su i sur., 2019.

3. STROJNO PREVOĐENJE – MT

Strojno prevođenje (engl. *Machine translation*, MT u daljnjem tekstu) istraživačko je područje koje obuhvaća računalnu lingvistiku (CL) i procesiranje prirodnih jezika (NLP). Strojno je prevođenje vrlo slično procesu razgovora-dijaloga čovjeka i stroja na nekom prirodnom jeziku budući da generiranje odgovora nekog stroja na prirodnom jeziku uključuje također neku vrstu prevođenja. Strojno prevođenje kao i strojno generiran dijalog mapiraju jedan niz znakova u drugi: kod

prevođenja to je npr. niz riječi/znakova iz engleskog jezika u npr. niz riječi na hrvatskom jeziku, a kod strojnog dijaloga niz riječi znakova prirodnog jezika u očekivani niz znakova strojno generiranog odgovora (Lane i sur., 2019).

Glavna je zadaća strojnog prevođenja (MT) prevođenje s izvornog jezika na ciljani jezik uz očuvanje originalnog značenja bez „tragova". Izazov kvalitetnog prijevoda ogleda se u zadržavanju značenja originalnog teksta u ciljnom jeziku, kao i činjenice da taj novi tekst treba biti pravilno napisan („u duhu jezika") u ciljnom jeziku, a ne doslovno preveden riječ po riječ (Skadiņa i sur., 2019)

Kako problem prevođenja zahtijeva više razina apstrakcije (riječi, klauze, složene klauze, POS, morfologija riječi, sintaksa, semantika, diskurs, veza između klauza) nad prirodnim jezikom, ali i puno stručnog znanja i zaključivanja, to MT čini jednim od težih problema kod umjetne inteligencije (Koehn, 2020). Također, budući da se jedna rečenica može prevesti na više način i procjena kvalitete prevođenja predstavlja dodatni izazov (Majcunić, Matetić, Brkić Bakarić,2019).

3. 1 Prededitiranje, posteditiranje i kontrolirani jezik

Kako bi se smanjile pogreške automatskog prevođenja i olakšalo kreiranje algoritma za strojno prevođenje, često se koriste prilagodba i uređivanje ulaznih tekstova, naknadno uređivanje rezultata strojnog prijevoda te korištenje reduciranog, kontroliranog prirodnog jezika. Prededitiranje ili uređivanje postupak je pripreme i transformacije ulaznih tekstova (ili govora) koja uključuje izmjenu tekstova prije prevođenja – najčešće se odnosi na ispravljanje uglavnom gramatičkih pogrešaka u tekstu uz uklanjanje nejasnoća i suvišnih dijelova te zadržavanje semantike (Gerlach i sur., 2013). Cilj je prethodnom detekcijom i prilagodbom/modifikacijom problematičnih elemenata stvoriti izvorni tekst koji stroj lakše može prevesti. Pravila za *pre-editing* može čovjek primijeniti ručno ili automatski (Nitzke, 2019).

Post-editing ili postuređivanje proces je kojim stručnjaci dodatno pregledavaju i ispravljaju izlaz procesa strojnog prevođenja - ispravljanje sintaktičkih, semantičkih pogrešaka prijevoda, uklanjanje redundancije i izmjenu rečenične strukture (*Pre-editing and post-editing*).

Kontrolirani prirodni jezik (engl. *Controlled natural language*, CNL u daljnjem tekstu) definira se kao umjetno stvoren jezik koji se temelji na nekom prirodnom jeziku, zadržava većinu njegovih svojstava, ali je ograničen u leksičkom, sintaktičkom i semantičkom smislu (Kuhn, 2014). CNL moguće je definirati odabirom i/ili pojednostavljenjem leksičkih elemenata i kategorija i/ili lingvističkih pravila ili dodajući neke sintaktičke ili semantičke bilješke na originalan tekst u prirodnom jeziku. Korisnik CNL-a može biti čovjek ili stroj: pisac, urednik, prevoditelj, znanstvenik, student, softver za obradu teksta, softver za razumijevanje teksta, za rudarenje teksta i sl. (Choi, Isahara, 2014).

Formulacija kontroliranog jezika (Arendse, Gdaniec, 2001; Miyata i sur., 2015) olakšava postupak *pre-editinga* ili ga u potpunosti eliminira, a također olakšava *post-editing* (O'Brien, Roturier, 2007; Aikawa i sur., 2007; *The Machine translation Odyssey*, 2018; Miyata, Fujita, 2017).

Budući da postoje različite metodologije i pristupi za automatiziranje postupka strojnog prevođenja, automatizirani MT sustav može biti klasificiran s obzirom na primjenu metode: MT

temeljen na pravilima (engl. *Rule-based* ili RBMT), statistički MT ili SMT, temeljen na primjerima MT-a (engl. *Example based* ili EBMT), hibridni ili HMT te MT temeljen na dubokom učenju (engl. *Deep learning*), i to najčešće metodama neuralnih mreža (engl. *Neural networks* ili NMT) (Poibeau, 2017). U nastavku slijedi opis metoda.

3. 2 Strojno prevođenje temeljeno na pravilima

Strojno prevođenje temeljeno na pravilima (engl. *Rule based machine translation*, RBMT) poznato je i kao strojno prevođenje temeljeno na znanju ili klasični pristup MT-a. Općenito označava sustave strojnog prevođenja koji se temelje na lingvističkim informacijama o izvornim i ciljnim jezicima u osnovi preuzetim iz (dvojezičnih) rječnika i gramatika koji pokrivaju glavne semantičke, morfološke i sintaktičke pravilnosti svakog jezika. RBMT sustav na temelju ulaznih rečenica (na nekom izvornom jeziku) generira izlazne rečenice (na nekom ciljnom jeziku) na temelju morfološke, sintaktičke i semantičke analize obje vrste izvornog i ciljnog jezika uključenih u dani prevoditeljski proces.

RBMT pristup primjenjuje skup jezičnih pravila u fazama analize, prijenosa i generiranja. Sustav temeljen na pravilima zahtijeva sintaktičku i semantičku analizu te sintaktičko i semantičko generiranje.

Nedostatci povezani s pristupom RBMT jesu nedovoljna količina dobrih rječnika (dok je izgradnja novih skupa), činjenica da se neke jezične informacije moraju postaviti ručno, složenost s interakcijama pravila u velikim sustavima, dvosmislenosti i idiomatskim izrazima te neuspješna prilagodba na nove domene (Okpor, 2014).

3. 3 Statistički pristup kod strojnog prevođenja

Statistički pristup temeljen je na podatcima i oslanja se na kvantitativne metode za otkrivanje odnosa. Ovakvi pristupi zahtijevaju tekstualne korpuse kako bi osigurali pouzdanost i razvili modele za automatizaciju različitih jezičnih konteksta i modele koji opisuju jezične pojave računanjem različitih parametara. Općenito, NLP temeljen na statistici ne ograničava se na osnovne statističke konstrukte iz teorije vjerojatnosti već obuhvaća više kvantitativnih pristupa za automatsku obradu jezika kao što su modeliranje vjerojatnosti, teorija informacija i linearna algebra (Hogenboom, 2010).

Različitosti sustava SMT uglavnom proizlaze iz definiranja minimalne jedinice za prijevod (riječ, fraza ili konstituent – sastavni dio). Prijevod koji se temelji na frazama – PBSMT (Koehn i sur., 2003) predstavlja evoluciju prijevoda od riječi do riječi. Ako se u obzir uzima fraza umjesto riječi, uzima se u obzir kontekst i lokalni razmještaj svake riječi (Koehn, 2010). Fraza ne predstavlja nužno jezični element izvornog jezika (engl. *Source language*, SL) koji mora imati odgovarajuću vrijednost u jeziku na koji se prevodi (engl. *Target language*, TL). Kako je broj fraza koje se mogu izvući iz korpusa veći od broja sintaktičkih elemenata, omogućena je veća sloboda u prijevodu, ali također i veći broj gramatički neispravnih prijevoda (España-Bonet, Costa-jussà, 2016).

Sustavi temeljeni na sintaksi (SSMT) primjenjuju suprotan pristup i prevode pojedine sintaktičke elemente (Wu, 1997; Quirk i sur., 2005). Osnovna je ideja pristupa upotreba sinkronih gramatika koje mogu stvoriti vrijednost iz izvora i cilja prevođenja istovremeno. Sinkrone gramatike kreirane su iz paralelnog korpusa što ovaj pristup čini vrlo sporim u odnosu na sustave PBSMT.

Hijerarhijski sustavi temeljeni na frazama (HPBSMT) svojevrsna su ravnoteža između pristupa PBSMT i SSMT (Chiang, 2005). Hijerarhijska se fraza sastoji od hijerarhije riječi i podfraza koja je namijenjena uspostavljanju pravilna redoslijeda među frazama (España-Bonet, Costa-jussà, 2016).

Statistički pristup omogućava precizno modeliranje različitih značenja riječi u skladu s kontekstom, a statistički su podatci efikasni za pronalaženje ekvivalenta prijevoda na razini riječi ili fraze. Statističkim pristupom danas je relativno dobro formalizirana leksička semantika dok su razumijevanje semantike rečenica i odnosa među njima još uvijek izazov koji se danas pokušava rješavati pristupom poznatim kao dubinsko učenje (engl. *Deep learning*) (Poibeau, 2017).

3. 4 Strojno prevođenje temeljeno na primjerima

Prevođenje temeljeno na primjerima ili analogiji (engl. *Example based machine translation*, EBMT) uvedeno je s ciljem prevladavanja problema prevođenja temeljenog na pravilima. EBMT obično djeluje u tri faze kako bi se prevela rečenica. U prvoj fazi sustav pokušava pronaći fragmente rečenice koja će se prevesti u korpusima izvornog jezika. U drugoj fazi, nakon što su svi relevantni fragmenti prikupljeni i pohranjeni, sustav traži translacijske ekvivalente na ciljnom jeziku, zahvaljujući tekstovima iz dvojezičnog korpusa koji se koriste za prijevod. Treća i zadnja faza kombinira fragmente prijevoda kako bi se dobila točna rečenica na ciljnom jeziku (Poibeau, 2017).

3. 5 Hibridni pristup kod strojnog prevođenja

Hibridni je pristup razvijen kombinacijom SMT-a i RBMT-a te je kompenzirao slabe točke oba pristupa s njihovim prednostima (Hogenboom, 2010). Postoje hibridni sustavi kojima je osnovni mehanizam statistički i oni kojima je osnovni mehanizam temeljen na pravilima. *Hibridni sustavi kojima je osnovni mehanizam zasnovan na SMT-u*: statistički se sustavi mogu graditi kad god postoje paralelni korpusi, a na kvalitetu prijevoda utječe i količina podataka. Dobar leksički izbor predstavlja prednost sustava SMT kada su dostupni podatci o određenoj domeni, budući da se leksički odabir modelira iz tih podataka (Costa-jussà i sur., 2012). Statistički sustavi pokušavaju poboljšati gramatičku točnost integrirajući značajke iz sustava RBMT, na primjer u fazama prededitiranja i posteditiranja.

Hibridnim sustavima strojnog prevođenja pristupa se iz perspektive temeljene na pravilima ili korpusu, dakle statistici.

Hibridni sustavi kojima je glavni mehanizam zasnovan na pravilima: sustavi RBMT inače zahtijevaju puno vremena i napora budući da su rječnici i pravila ručno kreirani i definirani. No uz ispravnu analizu izvornog teksta i dobro definiranih pravila, a kako bi poboljšali točnost, čisti RBMT sustavi upotrebljavaju sve više komponenata temeljenih na korpusima i statistici i time se automatiziraju neki dijelovi procesa prevođenja, koji su se provodili ručno.

Kad su hibridni sustavi vođeni SMT sustavima, integracija pravila izvršena je od faze prededitiranja/posteditiranja do središnjega dijela sustava. Kad su hibridni sustavi vođeni RBMT sustavima, integracija statistike upotrebljava se kako bi se obogatili resursi, moduli temeljeni na podatcima ili okosnica sustava temeljenog na pravilima (España-Bonet, Costa-jussà, 2016).

4. METODE DUBOKOG UČENJA

Razvojem procesorske moći računala otvaraju se nove mogućnosti u obradi ogromnog broja različitih podataka, prvenstveno nestrukturirana tipa, kao što je tekst, govor, slika ili video. Tijekom posljednjih se godina razvojem neuronskih mreža pojavila nova vrsta statističkog učenja nazvana „duboko učenje" (engl. *Deep learning*, DL u daljnjem tekstu) ili „hijerarhijsko učenje". Prema autorima Goodfellow i sur. (2016), „suvremeni pojam" DL-a nadilazi neuroznanstvenu perspektivu trenutne generacije modela strojnog učenja te obuhvaća općenitiji princip učenja više razina kompozicije, koji se može primijeniti u okvirima strojnog učenja koji nisu nužno neuronske mreže. Umjesto da koristi skupinu unaprijed definiranih karakteristika, DL djeluje iz jako velikog niza primjera. Ovo je učenje hijerarhijsko jer započinje osnovnim elementima (znakovima ili riječima u slučaju jezika), nastavlja identifikacijom složenijih struktura (nizovi riječi ili izraza u slučaju jezika) dok ne dobije cjelovitu analizu predmeta koji se analizira (rečenica) (Poibeau, 2017; Deng, Liu, 2018). Metode DL-a utemeljene na neuronskim mrežama pokazuju obećavajući trend u obradi prirodnog jezika, posebno u području strojnog prevođenja (Zheng i sur., 2019).

Primjena DL-a kod strojnog prevođenja omogućava stvaranje sustava koji sam izvodi najbolju reprezentaciju iz podataka uz vrlo malo ručnog podešavanja. Sustav za prevođenje koji je temeljen isključivo na dubokom učenju (engl. *Deep learning machine translation* ili *Neural machine translation*) sastoji se od *encodera* i *decodera*, koji se temelje na neuronskoj mreži. *Encoder* je dio sustava koji analizira trening-podatke, odnosno podatke za učenje, a *decoder* dio sustava koji automatski generira prijevod iz određene rečenice, na temelju podataka koje je analizirao *encoder*.

Pristupom dubokog učenja kod strojnog prijevoda odjednom se razmatra cijela rečenica bez njezine dekompozicije na manje segmente, istovremeno razmatrajući sve vrste odnosa u kontekstu. DL pristup može otkriti strukturu rečenica (odnose riječi ili skupina riječi) na temelju pravilnosti koje su identificirane u velikom broju primjera danih sustavu tijekom treninga. Spomenuti odnosi u rečenici mogu biti vertikalni (skupine sličnih riječi koje mogu popuniti poziciju u rečenici) ili horizontalni (sintaktički povezane skupine riječi u rečenici), što čini ovakav pristup fleksibilnim i kognitivno zanimljivim, kao i računalno vrlo zahtjevnim.

DL je značajan korak naprijed i omogućio je velika poboljšanja u području obrade prirodnih jezika, strojnog prevođenja, prepoznavanja slika i obrade govora (Poibeau, 2017).

Iako je tehnologija automatskog prevođenja već dugo predmetom istraživanja, izvorni cilj strojnog prevođenja (MT) – zamijeniti ljudske prevoditelje – još nije postignut (Skadiņa i sur., 2019).

Iako je primjena metoda DL-a u kratkom roku doprinijela povećanoj kvaliteti prevođenja, sustavi za strojno prevođenje još uvijek nisu u mogućnosti proizvesti izlaz iste kvalitete kao ljudski prevoditelj.

Prostora za napredak ima pogotovo na gramatičkim razinama složenih klauza (Wu i sur., 2018; Ge i sur., 2019).

5. ZAKLJUČAK

Obrada prirodnih jezika kao i konkretnije, strojno prevođenje, jako su napredovali u zadnja dva desetljeća, pogotovo primjenom dubokog učenja zasnovanog na neuronskim mrežama. Izvođenje zadataka kao što su razumijevanje teksta (uz mogućnost odgovaranja na pitanja), razumijevanje govora, vođenje dijaloga između računala i čovjeka već se naveliko koriste u praksi s različitim stupnjem kvalitete. Upravo sve šira uporaba metoda obrade prirodnih jezika pretpostavlja i dobro temeljno znanje o ovom području što je i predstavljalo motivaciju pisanja ovog rada.

Kako bi se koncizno i sistematski dao pregled i povezanost područja, u radu su prikazani odnosi između niza povezanih područja i njihovih sastavnica: umjetna inteligencija, obrada prirodnih jezika, prepoznavač i generator prirodnog jezika, strojno prevođenje, strojno učenje i duboko učenje. Osim osnovnih značajki opisana je temeljna problematika područja obrade prirodnih jezika i konkretnije metoda strojnog prevođenja. Za opisane metode diskutirala se i praktičnost i smjernice za primjenu pa je tako klasični pristup, temeljen na pravilima, prikladan za uže i specijalizirane domene NLP-a budući da pravilima može formalizirati ekspertno znanje dok je statistički pristup prikladniji kada postoje veliki paralelni jezični korpusi. Hibridnim se pristupom obuhvaćaju prednosti oba pristupa. U radu je opisan i najnoviji pristup dubokog učenja zasnovan na neuronskim mrežama kojem se povećala kvaliteta prevođenja prvenstveno za široko korištene jezike za koje postoje veliki paralelni korpusi. Iako još ne može generirati prevođenje kvalitete ljudskog prevoditelja, strojno je prevođenje, kao i ostala potpodručja NLP-a, već naširoko korišteno u praksi: sumarizacije dokumenata, prilagođeni rezultati analitika, ekstrakcija informacija i znanja, jeftino ili besplatno prevođenje i sl. Svaka od navedenih metoda strojnog prevođenja prikladna je za određene domene primjene te se očekuje i njihov daljnji razvoj, pogotovo hibridnih i metoda dubokog učenja. Također, kako bi se smanjila potreba za prededitiranjem i posteditiranjem te pogreške strojnog prevođenja prvenstveno na semantičkoj, govornoj i pragmatičkoj razini, bitno je uključiti i velik broj ljudskih eksperata koji procjenjuju i ispravljaju rezultate strojnog prevođenja.

U ovom je području još mnogo postojećih izazova, prostora za napredak i poboljšanje kvalitete automatskih obrada/prijevoda/generiranja jezika kako bi one bile na razini ljudske komunikacije ili prevođenja koje izvode ljudski eksperti. Daljnji smjer istraživanja je usmjeren na konkretne primjere obrade prirodnih jezika, izradu specijaliziranih korpusa za primjenu različitih algoritama strojnog prevođenja te istraživanje mogućnosti i primjene prevođenja prirodnog jezika u formalni i obrnuto.

REFERENCE

Aikawa, T., Schwartz, L., King, R., Corston-Oliver, M., & Lozano, C. (2007) „Impact of controlled language on translation quality and post-editing in a statistical machine translation environment", Proc. of MT Summit, 1-7

Arendse, B., Gdaniec, C. (2001) „MTranslatability", Machine Translation, 16(3), 175-218

Brachman, R. J., Levesque, H. J. (2004) Knowledge representation and reasoning, San Francisco: Elsevier

Chiang, D. (2005) „A hierarchical phrase-based model for statistical machine translation", u: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), 263-270

Choi, K.-S., Isahara, H. (2014) „Toward ISO Standard for Controlled Natural Language", u :LREC Workshop W2: Controlled Natural Language Simplifying Language Use, Reykjavik

Costa-jussà, M. R., Farrús, M., Marino, J. B., Fonollosa, J. A. R. (2012) „Study and comparison of rule-based and statistical catalan-spanish machine translation systems", Computing and Informatics, 31(2), 245-270

Deng, L., Liu, Y. (2018) „A Joint Introduction to Natural Language Processing and to Deep Learning", u: Y. Deng, Li, Liu (ur.), Deep Learning in Natural Language Processing: Springer Nature

Dovedan, Z. (1993) „Postupci sintaktičke analize prirodnih jezika", u: Radovi Zavoda za informacijske studije. Zagreb: Filozofski fakultet

Eisenstein, J. (2019) Introduction to natural language processing: MIT Press.

España-Bonet, C., Costa-jussà, M. R. (2016) „Hybrid Machine Translation Overview", u: M. R. Costa-jussà, R. Rapp, P. Lambert, R. E. Banchs, i B. Babych (ur.), Hybrid Approaches to Machine Translation: Springer International Publishing Switzerland

Feldman, S. (1999) „NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval", Information Today

Ge, S., Wu, S., Chen, X., i Song, R. (2019) „A Grammatical Analysis on Machine Translation Errors", u: Z. J. Chen J. (ur.), Machine Translation. CWMT 2018. Communications in Computer and Information Science, vol 954. Springer, Singapore. Springer Singapore. http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-981-13-3083-4_1

Pre-editing and post-editing, University of Geneve, dostupno na https://www.unige.ch/fti/en/faculte/departements/dtim/recherches/ta/ (pristupljeno 20. 10. 2020.)

Gerlach, J., Porro, V., Bouillon, P., Lehmann, S. (2013) „Combining pre-editing and post-editing to improve SMT of user-generated content" u: M. S. and L. S. Sharon O'Brien (ur.), Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice, Nice (pp. 45-53)

Goodfellow, I., Bengio, Y., Courville, A. (2016) Deep Learning, Cambridge: MIT Press

Gottschalk-Mazouz, N. (2007) „Internet and the flow of knowledge: Which ethical and political challenges will we face?", u: Proceedings of the 30. International Ludwig Wittgenstein Symposium. Volume 2. Kirchberg Am Wechsel, 215-232

Hogenboom, F. (2010) „An Overview of Approaches to Extract Information from Natural Language Corpora", u: 10th Dutch-Belgian Information Retrieval Workshop, DIR 2010 January 25. Nijmegen, Netherlands

Johnson, M. (2012) Natural Language Processing and Computational Linguistics: from Theory to Application, dostupno na: http://web.science.mq.edu.au/~mjohnson/papers/CLandTopicModels.pdf (pristupljeno 10. 10. 2020.)

Jordan, M. I., Russell, S. (2001) „Computational Intelligence", u: R. A. Wilson i F. C. Keil (ur.), The MIT Encyclopedia of the Cognitive Sciences (LXXIII – XC), A Bradford book, Massachusetts

Jurafsky, D., Martin, J. H. (2008) Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition: Prentice Hall

Koehn, P. (2010) Statistical machine translation: Cambridge University Press

Koehn, P., Och, F. J., Marcu, D. (2003) „Statistical phrase-based translation", u: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, NAACL '03, 48-54

Koehn, P. (2020) Neural Machine Translation: Cambridge University Press, https://doi.org/10.1017/9781108608480

Kuhn, T. (2014) „A Survey and Classification of Controlled Natural Languages", Computational Linguistics, 40(1), 121-170

Lane, H., Howard, C., Hapke, H. (2019) Natural Language Processing in Action – Understanding, analyzing, and generating text with Python, New York: Manning Publications Co.

Liddy, E. D. (2003) „Natural Language Processing", u: Encyclopedia of Library and Information Science, 2nd edition (pp. 2126-2136), Marcel Decker, Inc.

Majcunić, S., Matetić, M., Brkić Bakarić,M.(2019) „Translation error analysis in treat: a windows app using the MQM framework", Zbornik Veleučilišta u Rijeci, Vol. 7 No. 1, 140-162.

Manning, C. D., Schutze, H. (1999) Foundations of Statistical Natural Language Processing: MIT Press

Miyata, R., Fujita, A. (2017) „Dissecting Human Pre-Editing toward Better Use of Off-the-Shelf Machine Translation Systems", u: Proceedings of the 20th Annual Conference of the European Association for Machine Translation (EAMT), Prague

Miyata, R., Hartley, A., Paris, C., Tatsumi, M., Kageura, K. (2015) „Japanese controlled language rules to improve machine translatability of municipal documents", u: Proc. of MT Summit, 90-103

Nitzke, J. (2019) Problem solving activities in post-editing and translation from scratch. A multi-method study, Berlin: Language Science Press

O'Brien, S., Roturier, J. (2007) „How portable are controlled language rules? A comparison of two empirical MT studies", u: Proc. of MT Summit, 345-352

Okpor, M. D. (2014) „Machine Translation Approaches: Issues and Challenges", International Journal of Computer Science Issues, Vol. 11 (Issue 5, No 2)

Poibeau, T. (2017) Machine Translation, Cambridge MA: MIT Press Essential Knowledge series

Quirk, C., Menezes, A., Cherry, C. (2005) „Dependency treelet translation: Syntactically informed phrasal SMT", u: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Association for Computational Linguistics, 271-279

Rajesh, K. S., Reddy Lokanatha C. (2009) „Natural Language processing – an intelligent way to understand context sensitive languages", International Journal of Intelligent Information Processing, 3(2), 421-428

Sciforce. (2019) A Comprehensive Guide to Natural Language Generation, dostupno na https://medium.com/sciforce/a-comprehensive-guide-to-natural-language-generation-dd63a4b6e548 (pristupljeno 1. 9. 2020.)

Silva, C. S. R., Fonseca, J. M. (2018) „Artificial intelligence and algorithms in intelligent systems: Advanced analytics: Moving forward artificial intelligence (AI), Algorithm intelligent systems (AIS) and General impressions from the field", u Artificial Intelligence and Algorithms in Intelligent Systems - Proceedings of 7th Computer Science On-line Conference, 2018 (pp. 308–317), R. Silhavy (ur.), Springer Verlag. https://doi.org/10.1007/978-3-319-91189-2_30

Skadiņa, I., Gaizauskas, R., Vasiļjevs, A., Paramita, M. L. (2019) „Introduction", u: V. A. Skadiņa I., Gaizauskas R., Babych B., Ljubešić N., Tufiş D. (ur.), Using Comparable Corpora for Under-Resourced Areas of Machine Translation. Theory and Applications of Natural Language Processing. Springer, Cham. https://doi.org/10.1007/978-3-319-99004-0_1

Su, S.-Y., Huang, C.-W., Chen, Y.-N. (2019) „Dual Supervised Learning for Natural Language Understanding and Generation", u: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 5472-5477, https://doi.org/10.18653/v1/P19-1545

Su, S.-Y., Yuan, P.-C., Chen, Y.-N. (2018) „How time matters: Learning time-decay attention for contextual spoken language understanding in dialogues", u: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, 2133-2142

Šuman,S. (2019) „Sustav za prevođenje poslovnih opisa u model podataka entiteta i veza", doktorska disertacija, Odjel Informatike, Rijeka

The Machine translation Odyssey. (2018), dostupno na http://translation-blog.trustedtranslations.com/the-machine-translation-odyssey-2018-05-16.html (pristupljeno 9. 9. 2020.)

Tsujii, J. (2011) „Computational Linguistics and Natural Language Processing", u: Computational Linguistics and Intelligent Text Processing. CICLing 2011. Lecture Notes in Computer Science, Gelbukh A.F. (ur.), Vol. 6608. Springer Berlin, Heidelberg

Wu, D. (1997) „Stochastic inversion transduction grammars and bilingual parsing of parallel", Computational Linguistics, 23(3), 377-403

Wu, H., Zhang, H., Li, J., Zhu, J., Yang, M., Li, S. (2018) „Training machine translation quality estimation model based on pseudo dana", Acta Sci. Nat. Univ. Pekin., 54(2), 279-285, https://doi.org/10.13209/j.0479-8023.2017.158

Zheng, Z., Huang, S., Dai, X., Chen, J. (2019) „Controlling the Transition of Hidden States for Neural Machine Translation", u Machine Translation. CWMT 2018. Communications in Computer and Information Science, Z. J. Chen J. (ur.), vol 954. Springer, Singapore. Springer Singapore. https://doi.org/10.1007/978-981-13-3083-4_8

Review article
https://doi.org/10.31784/zvr.9.1.23
Received: 17. 11. 2020.
Accepted: 20. 1. 2021.

OVERVIEW OF NATURAL LANGUAGE PROCESSING AND MACHINE TRANSLATION METHODS

Sabrina Šuman
PhD, Senior Lecturer, Polytechnic of Rijeka, Vukovarska 58, 51000 Rijeka, Croatia; e-mail: ssuman@veleri.hr

ABSTRACT

The paper provides an overview of areas related to the processing of natural languages and their interrelationships, starting from a broader domain such as artificial intelligence, through machine learning, computational linguistics, machine translation methods and especially those based on deep learning. The characteristics, applications, phases and main problems of natural language processing from the lexical, syntactic, semantic, speech and pragmatic perspective are described. The phases of natural language recognition and analysis as well as the natural language generation phase are described. Pre-editing and post-editing procedures using controlled natural languages are given as examples of practices that increase the accuracy and quality of automatic translation and text processing in general. Special focus is given to machine translation and machine translation methods. Approaches to machine translation as statistical, rule-based, example-based, hybrid and deep learning-based approach are described and discussed with regard to their advantages and disadvantages including appropriate application in practice. In the end, still unresolved challenges are given as a direction of future research related to natural language processing and the importance of further development of a deep learning-based approach.

Key words: *natural language processing, machine translation, computational linguistics, deep learning, artificial intelligence*

TEHNIČKE ZNANOSTI