Govorne tehnologije

VIRTUALNE ASISTENTICE UČE HRVATSKI, ALI TEŠKO IM IDE

Obrada prirodnog jezika ili, kraće, NLP jedna je od užarenih tehnologija bez kojih je nezamisliva upotreba svih onih silnih pametnih zvučnika i telefona koje je najjednostavnije aktivirati glasom. Anglizirani svijet lakonski se predao glasovnim naredbama, ali Hrvatima to teže pada. Umjetni sugovornici njihov jezik jedva natucaju

piše **ALEKSANDAR TEŠIĆ**
aleksandar.tesic@lider.media

Laičkom umu naučenom na usvojene koncepte i isprobane obrasce nije baš lako shvatljiv proces u kojem umjetna inteligencija (AI) ovladava razumijevanjem ljudskoga jezika toliko da je osobna asistentica u stanju razumjeti što joj govorimo i dok joj se obraćamo uobičajenim govornim jezikom ili da je redakcijski asistent sposoban sam napisati vijest. Ne treba ni biti jer priča nije nimalo jednostavna. Bavi se njome čitava jedna grana računalne znanosti koju u tehnološki dominantnom engleskom govornom području već prepoznaju po skraćenici NLP (engl. natural language processing), a kod nas nazivaju obradom prirodnog jezika.

S razvojem NLP-ja mnogi umjetnointeligentni sustavi uče sve bolje razumjeti jezične finese. Mnogo im bolje ide učenje engleskoga i drugih velikih jezika, a s malim hrvatskim, unatoč rastućoj tržišnoj potražnji, još uvijek muku muče. Obradom prirodnog hrvatskog jezika domaće znanstvenoistraživačke institucije bave se desetljećima, ali masovniji razvoj i tržišni plasman umjetnointeligentnih sustava koji podržavaju hrvatski jezik za sada je izostao, uglavnom zbog malog tržišta i nedovoljnog ulaganja u istraživanja. Pa ipak, postoje i domaće tvrtke koje su u razvoju prirodnojezičnih alata i usluga pronašle svoju poslovnu nišu.

Teško i algoritmima

Lingvistica **Nevena Erak Camaj**, koja je zahvaljujući NLP-ju sa svojim životnim i poslovnim par-

Nevena Erak Camaj i Eduard Camaj, osnivači tvrtke Erato, razvili su prvu hrvatsku platformu za izradu 'chatbota' koji razumiju pitanja korisnika i daju smislene odgovore. Platforma je specijalizirana ponajprije za hrvatski jezik

Desetljećima se domaće znanstvenoistraživačke institucije bave obradom prirodnog jezika, ali masovniji razvoj i tržišni plasman jetnointeligentnih sustava koji podržavaju hrvatski jezik zasad je izostao, uglavnom zbog malog tržišta

ša Martinić, ektorica racija ke Newton nologies Adria, glavni izazov radi prirodnih venskih jezika odi količinu ataka koje a analizirati. Za leski ili druge ke jezike postoji ogo besplatnih a podataka, a atski, kaže, u smislu nije oljno bogat

tnerom, programerom **Eduardom Camajom** razvila prvu hrvatsku platformu za izradu 'chatbota' specijaliziranu ponajprije za hrvatski jezik, u objavi na blogu njihove tvrtke Erato približava nam donekle složenu problematiku obrade prirodnog jezika. O famoznom NLP-ju, grani računalne znanosti koja umjetnoj inteligenciji pomaže razumjeti ljudski jezik i dok joj se ne obraćamo jasnom formom, uputama i naredbama, kreativno piše kao o magiji koja računalima omogućuje da analiziraju, razumiju i izdvoje značenje iz ljudskog jezika. Navodi da zahvaljujući NLP-ju i različitim algoritmima i mehanizmima strojnog učenja rečenice ljudskog jezika možemo strukturirati i preoblikovati na kojekakve načine – označavanjem, razlaganjem na manje smislene jedinice, određivanjem njihovih međusobnih odnosa, utvrđivanjem entiteta, sažimanjem ili prepoznavanjem izvanjezičnih obilježja.

Sada kada smo koliko-toliko definirali i shvatili terminologiju, lakše ćemo razumjeti potrebu da s umjetnointeligentnim sustavom komuniciramo na prirodnom hrvatskom jeziku i naslutimo izazove koji proizlaze iz mnogih specifičnosti našeg malog južnoslavenskog jezika.

– Imajući na umu da je hrvatsko tržište iznimno malo i donekle tehnološki zaostaje za najnovijim trendovima, može se dovesti u pitanje isplativost razvoja tehnologija specifičnih za hrvatski jezik. To i jest glavni razlog zašto se velike kompanije nisu pretjerano bavile hrvatskim jezikom u dijelu razumijevanja teksta jer je hrvatski u tom pogledu specifičan i ne može se algoritamski samo 'preslikati' s engleskoga. S druge strane, potreba na tržištu postoji i svakim je danom sve veća, pogotovo ako znamo da sazrijevanje hrvatskog tržišta u određenim tehnološkim segmentima kasni od tri do pet godina, a na svjetskom su tržištu proizvodi i usluge utemeljeni na ovim tehnologijama u punom jeku već neko vrijeme – govori nam Erak Camaj, čija tvrtka Erato tehnologiju obrade prirodnog jezika primjenjuje u izradi 'chatbotova' – virtualnih asistenata koji razumiju pitanja korisnika i daju smislene odgovore.

Nastojanja EU

Da prirodnojezične tehnologije za hrvatski jezik još znatno zaostaju za velikim jezicima, ali i da njihov razvoj napreduje velikom brzinom, upućuje nas, pak, **Nives Mikelić Preradović**, koja studente Odsjeka za informacijske i komunikacijske znanosti zagrebačkog Filozofskog fakulteta podučava primjeni pojmova, trendovima i sustavima za obradu prirodnog jezika.

– Tvrtke na hrvatskom tržištu mogu zadovoljiti rastuću potražnju domaćega gospodarstva za tehnologijama otvorenoga koda, na primjer strojno prevođenje, prepoznavanje govora, diktiranje te inteligentne sustave za višejezično pretraživanje informacija, sažimanje i generiranje teksta, samo razvojem naprednih visokojezičnih alata u obliku visokotehnoloških rješenja za koje je potrebno pozamašno financiranje, ali i računalno razumijevanje prirodnog jezika, što je još uvijek iznimno teško zbog višeznačnosti prirodnog jezika. Trenutačna nastojanja na razini EU koja obuhvaćaju i hrvatski jezik je da se europskim malim i srednjim poduzećima omogući pristup prirodnojezičnim tehnologijama, primjerice, strojnom prevođenju, kako bi im se omogućilo pristupanjem novim tržištima i razvojnim mogućnostima te razvoj višejezičnih javnih e-usluga. Tržišna potreba postoji, ne samo u Hrvatskoj nego u cijeloj EU, budući da su prirodnojezične tehnologije u ekonomskom pogledu nužne za razvoj jedinstvenog digitalnog tržišta, kojim trenutačno dominiraju rascjepkanost i neeuropski akteri koji ne uzimaju u obzir potrebe višejezične Europe – kaže Nives Mikelić Preradović i navodi primjer Styria Grupe, čija tvrtka Styria Data Science sudjeluje u tri mi-

Nives Mikelić Preradović, redovita profesorica zagrebačkog Filozofskog fakulteta, navodi primjer tvrtke Styria Data Science koja sudjeluje u tri milijuna eura teškom istraživačkom EU projektu 'Embeddia' koji bi 2021. trebao rezultirati razvojem redakcijskoga robotskog asistenta koji će, među ostalim, biti u stanju sâm napisati vijest

Gordan Gledec, znanstvenik iz Zavoda za primijenjeno računarstvo zagrebačkog Fakulteta elektrotehnike i računarstva, navodi da potražnju hrvatskoga gospodarstva za prirodnojezičnim AI alatima i uslugama uglavnom zadovoljavaju strane tvrtke jer domaćih koje bi im konkurirale nema

lijuna eura teškom istraživačkom EU projektu 'Embeddia', koji bi 2021. trebao rezultirati razvojem redakcijskog robotskog asistenta koji će iz strukturiranih podataka biti u stanju sam napisati vijest, kao i sustava koji će se razumijevanjem prirodnog hrvatskoga koristiti za bolje povezivanje sličnih članaka i filtriranje neprimjerenih komentara.

Eksperimentalni sustavi

Obradom hrvatskog prirodnog jezika bavi se i tvrtka Newton Tehnologies Adria u vlasništvu Newton Media Grupe iz Praga. Specijalizirana je za razvoj sustava za pretvaranje govora u tekst s fokusom na slavenske jezike, a jedan od njezinih proizvoda je NEWTON Dictate – sustav za pretvaranje govora u tekst, trenutačno dostupan na hrvatskom, srpskom, slovenskom, slovačkom, poljskom i češkom jeziku. Direktorica operacija **Maša Martinić** kao glavni izazov u obradi prirodnih slavenskih jezika navodi količinu podataka koju je potrebno analizirati. Kaže da obrada podataka troši vrijeme i resurse i da za engleski ili druge velike jezike postoji mnogo besplatnih baza podataka, a da je hrvatski jezik u tom pogledu nedovoljno bogat.

Imajući u vidu da su troškovi istraživanja i razvoja visoki, **Sandu Martinčić-Ipšić**, profesoricu Odjela za informatiku Sveučilišta u Rijeci, ne čudi da je stupanj razvijenosti jezičnih i govornih tehnologija viši za jezike koji imaju veći broj govornika i posljedično veće tržište, a istodobno i ulažu više novca u znanstvena istraživanja. Svjedoči, međutim, da se i na našim sveučilištima itekako radi i navodi da je u sklopu istraživanja u području govornih tehnologija razvijen eksperimentalni sustav za govorni dijalog između korisnika i računala za tematski ograničeno područje komunikacije na hrvatskome jeziku. Govori nam da su razvijeni eksperimentalni sustavi za raspoznavanje i sintezu hrvatskoga govora te sustav za upravljanje dijalogom i da se rezultati tih istraživanja mogu implementirati u razne aplikacije koje podržavaju govornu komunikaciju na hrvatskom i razvijaju se za mobilne uređaje, ali da je njihova veća komercijalizacija izostala.

Moguća rješenja

Izdvojila je ipak riječku tvrtku e-glas, koja razvija asistivne tehnologije za hrvatski jezik i spomenula mogućnosti upotrebe Googleovih rješenja u aplikacijama, čija šira primjena opet izostaje ponajprije zbog premalog tržišta.

– Na Odjelu za informatiku postoji manji tim koji se već dulje bavi istraživanjima računalnih postupaka za jezične i govorne aplikacije za hrvatski jezik. Trenutačno razvijamo postupke sažimanja (sumarizacije) tekstova te postupke mjerenja semantičke sličnosti, koji uključuju reprezentaciju teksta u mrežama te se koriste dubokim učenjem. Ti postupci primjenjivi su u aplikacijama koje određuju razinu parafraziranja, odnosno plagijarizma, automatskog razvrstavanja tekstova u kategorije, automatskog određivanja ključnih pojmova ili tagova u tekstu, procjenjivanju kvalitete (koherentnosti) i čitljivosti teksta itd. Nadalje nastavljamo istraživanja vezano uz postupke slogovanja i za druge jezike, kao važan dio jezičnih i govornih aplikacija. Na području govornih tehnologija trenutačni projekti bave se primjenom postupaka dubokog učenja za analizu i sintezu hrvatskoga govora, pri čemu je cilj vjerno simulirati ljudske percepcijske sposobnosti za razumijevanje govora – otkriva Martinčić-Ipšić kojim su izazovima u obradi prirodnog hrvatskog zaokupljeni u Rijeci.

Prilika za poduzetne

Hrvatski problem nije u nedostatku znanja, nego u nedostatnoj potpori da se to znanje opredmeti do razine tržišnog proizvoda, zaključuje ovu priču **Gordan Gledec**, znanstvenik sa Zavoda za primijenjeno računarstvo zagrebačkog Fakulteta elektrotehnike i računarstva, jedne od domaćih znanstvenoistraživačkih institucija, koja se obradom hrvatskog jezika sustavno bavi već desetljećima.

– Prije desetak godina FER je, na poticaj jedne domaće tvrtke s kojom je tada surađivao, razvio prototip sustava za govorno upravljanje prema dobivenim specifikacijama. Riječ je bila o vrlo učinkovitom i robusnom sustavu, što potvrđuje činjenica da je besprijekorno reagirao i na izgovorene naredbe našega nesavršenoga Text-to-Speech sustava, tj. na artificijelni govor niske kvalitete. Nažalost, sve je ostalo na prototipu, jer je FER znanstvenoistraživačka institucija, a ne tvrtka koja bi se bavila održavanjem, 'upgradeanjem' i sličnim poslovima u vezi s isporučevinama namijenjenima širokoj uporabi. U Hrvatskoj, zbog nedovoljnih ulaganja u prirodnojezične tehnologije, nema poduzeća osposobljenih da ovakve proizvode preuzmu i dalje se o njima brinu – ispričao nam je Gordan Gledec i napomenuo na kraju da potražnju hrvatskoga gospodarstva za prirodnojezičnim AI alatima i uslugama uglavnom zadovoljavaju strane tvrtke, jer domaćih koje bi im konkurirale nema, da je ograničena ponuda daleko od potreba, a nuđenom kvalitetom i traženom cijenom često i prilično upitna te zaključio da bi hrvatski poduzetnici u suradnji s domaćim R&D-jem tu trebali tražiti svoje poslovne niše.

Profesorica Odjela za informatiku Sveučilišta u Rijeci Sanda Martinčić-Ipšić svjedoči da se i na našim sveučilištima itekako radi. Govori da su razvijeni eksperimentalni sustavi za raspoznavanje i sintezu hrvatskog govora te sustav za upravljanje dijalogom i da se rezultati tih istraživanja mogu ugraditi u razne aplikacije koje podržavaju govornu komunikaciju na hrvatskom i razvijaju se za mobilne uređaje, ali da nisu previše komercijalizirani

S razvojem NLP-ja (engl. natural language processing) mnogi umjetnointeligentni sustavi, poput virtualnih asistentica Siri i Alexe, uče sve bolje razumjeti jezične finese. No mnogo im bolje ide učenje engleskog i drugih velikih jezika – unatoč sve većoj tržišnoj potražnji s malim hrvatskim još muku muče