Značenje hrvatskog jezika otkriva se digitalno

Zahvaljujući Fakultetu elektrotehnike i računarstva Sveučilišta u Zagrebu tajne koje sadrži riječ mogle bi biti odgonetnute, a prije nedokučive korelacije dokumentiranih događaja postati očite i usmjeriti nas na ispravni put u kriznim trenucima

TEKST: LUKA FIŠIĆ
FOTO: RATKO MAVAR

Napredak tehnologije neće izbrisati i osiromašiti hrvatski jezik. Naprotiv, zahvaljujući Fakultetu elektrotehnike i računarstva u Zagrebu tajne koje sadrži riječ mogle bi biti odgonetnute, a prije nedokučive korelacije dokumentiranih događaja postati očite i usmjeriti nas na ispravan put u kriznim trenucima.

Na zagrebačkom FER-u već deset godina postoji istraživačka grupa Laboratorij za analizu teksta i inženjerstvo znanja (TakeLab). U njemu se bave obradom informacija iz tekstnih izvora, a njihova istraživanja obuhvaćaju obradu prirodnog jezika, pretraživanja informacija i strojnog učenja, odnosno primjene tehnika umjetne inteligencije, s ciljem razvoja naprednih jezičnih tehnologija za semantičku analizu digitalnog sadržaja. Voditeljica laboratorija je prof. dr. sc. Bojana Dalbelo Bašić, a njezin zamjenik izv. prof. dr. sc. Jan Šnajder.

– Naš je primarni fokus obrada prirodnog jezika i primjena metoda strojnog učenja za rješavanja tog problema u svrhu ekstrakcije znanja iz nestrukturiranih, tekstnih podataka i omogućavanja komunikacije između čovjeka i stroja prirodnim jezikom. Drugim riječima fokus je jezik, konkretno njegova manifestacija u tekstu. Naša je misija razviti modele i alate koji omogućavaju semantičku analizu velike količine tekstnih podataka – elaborira Šnajder.

Jan Šnajder, zamjenik voditeljice TakeLaba, objašnjava koliko je važno što u tom FER-ovom laboratoriju razvijaju jezične tehnologije za hrvatski jezik. Osjećaju, kaže, veliku odgovornost jer ako ne razvijemo tu tehnologiju hrvatski jezik bit će isključen iz digitalnog svijeta. Mnogo toga rade i na engleskom, a nešto manje i na njemačkom

Primjerice, ako vas zanima neki događaj ili pojam koji se nalazi u velikoj zbirci dokumenata, morali biste dobro pogoditi ključne riječi da pronađete željeno. No kod semantičkog pretraživanja postoji model pomoću kojeg se može prebroditi jaz između onog što trebate i načina na koji je to prirodnim jezikom zapisano. Nadalje, ako imate veliku količinu tekstnih podataka i zanima vas koji su glavni protagonisti u dokumentima (kako su povezani ili u kojoj su tvrtki zaposleni) isto tako će vam trebati pomoć modela na kakvim Laboratorij radi. U pitanju je klasičan primjeru ekstrakcije informacije gdje se iz teksta pokušava dobiti podatke iz kojih se zatim mogu donijeti zaključci i izvoditi hipoteze.

TakeLab iza sebe ima mnogo uspješnih završenih projekata, a među njima su i oni nastali u suradnji s industrijom. Riječ je o praktično orijentiranim projektima koji rješavaju konkretnu potrebu industrije za analitikom teksta (tipična situacija bila bi analiza sentimenta: preko komentara na društvenim mrežama saznati što ljudi misle o nekom proizvodu ili usluzi). Na jednom takvom projektu u njemu upravo rade. Projekt CATACX (Kognitivno-afektivna tekstna analitika društvenih medija za analizu korisničkog iskustva) financira HAMAG-BICRO, a cilj mu je analiza iskustva i mišljenja korisnika telekomunikacijskih usluga u komentarima na Facebooku. Saznati što je pošlo po zlu iz tisuće komentara nije lak zadatak. Riječ je o kompleksnom, skupom i dugom procesu. No CATACX bi taj proces trebao učiniti bezbolnim i lakim. Razvijen je model koji u konverzacijama korisnika s korisničkoj službom na Facebooku analizira niz afektivnih i kognitivnih pokazatelja, kao što su sentiment, komunikacijske namjere, teme razgovora te čak deset različitih, ponekad vrlo suptilnih emocija. Za pola godine prototip projekta trebao bi biti završen te zatim dalje usmjeren prema komercijalnoj uporabi.

– Imali smo niz vrlo uspješnih suradnji s tvrtkama u Hrvatskoj i inozemstvu te aktivno surađujemo s tvrtkom Photomath, koja ulaže u naše doktorande. Velik trud ulažemo i u to da dovedemo u našu grupu izvrsne znanstvenike iz inozemstva. Na primjer, imamo poslijedoktorsku znanstvenicu iz Italije – kaže Šnajder.

Uz financiranje Vlade Flandrije i Ministarstva znanosti i sporta izradili su i tražilicu CADIAL za pravne dokumente Republike Hrvatske, koja je javno dostupna svim građanima naše zemlje. Pomoću nje korisnici mogu

Prebujna hollywoodska mašta

Još desetak godina do naprednije umjetne inteligencije

Kada govori o umjetnoj inteligenciji, Šnajder je oprezan pri prognozama. Vjeruje da u sljedećih desetak godina neće promijeniti svijet na način kako to Hollywood često predviđa.

– U bliskoj budućnosti možemo očekivati razvoj alata za pomoć pri teškom fizičkim poslovima, modele koji će se moći baviti boljim strojnim prevođenjem, osobne asistente. No za ostalo morat ćemo pričekati više od 10 sljedećih godina. Problemi umjetne inteligencije sežu duboko u područje kognitivne znanosti i filozofije – vjeruje profesor.

semantički pretraživati hrvatske zakone i dobiti bolje informacije o pravnom poretku ili dobiti uvid u željene pravne dokumente. Razvili su i sustav za Hrvatsku informativnu novinsku agenciju (Hina) pomoću kojeg se dokumenti automatski razvrstavaju pretplatnicima prema temama, što je rezultiralo znatnom uštede vremena za korisnika.

– Jednako intenzivno radimo i na engleskom i na hrvatskom jeziku. Smatramo da je bitno da razvijamo jezične tehnologije za hrvatski jezik. To je iznimno važno za zemlje s malim brojem govornika i malim proračunom. Neki problemi zanimljivi su za rješavanje bez obzira o kojem je jeziku riječ – ističe Šnajder. Laboratorij u ovome trenutku radi na još dva projekta. SenseHive financira Hrvatska zaklada za znanost, a bavi se izgradnjom semantičkih mreža riječi na temelju prikupljanja odgovora velikog broja izvornih govornika hrvatskog jezika.

– Razvijamo platformu pomoću koje bismo pitali izvorne govornike hrvatskoga jezika što misle o značenju pojedinih riječi i kako ih upotrebljavaju u svakodnevnom govoru. Skupljajući te informacije i koristeći se metodama strojnog učenja s vremenom bismo modelirali značenje svake pojedinačne riječi u hrvatskom jeziku. Krećemo s bazom od 10.000 riječi i to bi trebao biti važan resurs za razvoj jezičnih alata za hrvatski jezik. Na projektu surađujemo s Filozofskim fakultetom u Zagrebu, Institutom za hrvatski jezik i jezikoslovlje te partnerima iz Slovenije i Njemačke – objašnjava profesor.

EVERBEST financira fond Jedinstvo uz pomoć znanja, a obuhvaća semantičku analizu događaja ekstrahiranih iz novinskih tekstova, s ciljem boljeg pretraživanja i preporučivanja događaja krajnjim korisnicima, a provodi se u suradnji s Institutom Jožefa Stefana u Ljubljani te sveučilištima u Stuttgartu i Nottinghamu.

– Dovođenje događaja u vezu jednog s drugim te njihovo razumijevanje bitno je da bismo mogli donositi bolje odluke u sadašnjosti za budućnost. Nadamo se uskoro predložiti model koji bi bio znatno napredniji od postojećih – dodaje Šnajder.

FER-ov Laboratorij za analizu teksta i inženjerstvo znanja (TakeLab) bavi se obradom informacija iz tekstnih izvora. Upravo provodi projekt CATACX, u kojem se analizira mišljenje korisnika telekomunikacijskih usluga u komentarima na Facebooku, razvija platformu pomoću koje istražuje što izvorni govornici hrvatskoga jezika misle o značenju pojedinih riječi i kako ih upotrebljavaju, a provodi i semantičku analizu događaja ekstrahiranih iz novinskih tekstova. Izradio je i tražilicu CADIAL, koja je javno dostupna svim građanima naše zemlje i pomoću koje korisnici mogu semantički pretraživati hrvatske zakone

Na Laboratoriju za analizu teksta i inženjerstva iznimno su predani radu sa studentima te trenutačno surađuju s tridesetak studenata preddiplomskog i diplomskog studija. Studenti su ove godine sudjelovali i na međunarodnom natjecanju u semantičkoj analizi teksta (SemEval). Jedan zadatak na kojemu su sudjelovali je detekcija humora na Twitteru, na kojemu su osvojili drugo mjesto. Do sada je u grupi na temama obrade prirodnog jezika i strojnog učenja diplomiralo više od 70 studenata.

– Većina ljudskog znanja kodirana je u tekstu, u kojem se nalazi bogati izvor informacija koje se mogu pretvoriti u znanje i pomoći u odlučivanju u raznim domenama ljudske djelatnosti. Radimo vrlo zanimljiva istraživanja koja su u srži umjetne inteligencije, a s druge strane proizlaze iz ljudske potrebe da se nosimo s velikom količinom informacija i da u moru podataka pronađemo smisao – zaključuje Šnajder.

VEĆINA LJUDSKOG ZNANJA KODIRANA JE U TEKSTU, U KOJEM SE NALAZI BOGAT IZVOR INFORMACIJA KOJE SE MOGU PRETVORITI U ZNANJE I POMOĆI U ODLUČIVANJU U RAZNIM DOMENAMA LJUDSKE DJELATNOSTI. ZATO NA FER-U RADE VRLO ZANIMLJIVA ISTRAŽIVANJA KOJA SU U SRŽI UMJETNE INTELIGENCIJE