Dnevnik.si

Novice.Dnevnik.si Tiskane izdaje/Dnevnik

Slovenščina in računalniki

Trubarji digitalne dobe

Dnevnikov objektiv - sobota, 10.01.2009
Tekst: Sebastijan Kopušar

V Notranjem stoji vas, Vrh po imenu. V tej vasici je živel v starih časih Krpan, močan in silen človek." Prva slovenska umetna pripovedka je zvenela iz zvočnikov, z glasom, ki se je kdaj hipoma zataknil, včasih kovinsko zabrnel, a kljub vsemu presenetljivo jasno in razločno prebiral Levstikovo povest. Računalniki so spregovorili v zelo dostojno razumljivi slovenščini, ne več kot Američan, ki med branjem žveči žeblje. Zdaj jih razumemo. Pa oni nas?

Prof. dr. France Mihelič, vodja skupine za govorne tehnologije v Laboratoriju za umetno zaznavanje, sisteme in kibernetiko (LUKS) na ljubljanski fakulteti za elektrotehniko, je eden od digitalnih Trubarjev.

Računalniki slovenščino razumejo, če jim narekujete številke. Raven višje, a s precej več motnjami, je govorjenje o fiktivnih urnikih letalskih poletov. Še najbolje gre slovenskim radiologom, ki že lahko narekujejo svoje diagnoze, elektronika pa jih ubogljivo izpisuje. To je ta hip skrajni domet uporabnosti, saj sta prva dva programa le laboratorijska poskusna zajčka. Pa še to ni razumevanje, ampak le prepoznavanje govora in njegov zapis.

Človeštvo že več kot pol stoletja sanjari o kramljanju z elektronskimi stroji, a po vsem tem času še vedno bolj ali manj capljamo na izhodišču. Leta 1968, ko je režiser Stanley Kubrick v svojem znamenitem filmu Odiseja 2001 svetu predstavil razumni stroj HAL, so računalniški strokovnjaki verjeli, da bodo njihove naprave v roku generacije ali dveh dosegle raven inteligence, ki je ne bo mogoče ločiti od človeškega razuma. Pionir umetne inteligence Herbert Simon je celo napovedal, da bodo "stroji v dvajsetih letih zmožni opraviti katerokoli delo, ki ga zmorejo ljudje".

Praskanje po površju

Štirideset let kasneje je vprašanje, ali računalniki lahko postanejo inteligentni, oziroma v precej okrnjeni različici, ali lahko razumejo pomen besedila vsaj na ravni na primer petletnega otroka, še vedno eno osrednjih vprašanj na področju umetne inteligence. Besedila so eden od temeljnih načinov vsakodnevnega komuniciranja med ljudmi, današnja raven računalniške tehnologije pa ne zmore več od zgolj zelo površinskega razumevanja vsebine, ugotavlja **dr. Marko Grobelnik** z Instituta Jožef Stefan. Pravzaprav so v vsem tem času stroji zmogli en sam pomembnejši dosežek, v šahu so premagali svetovnega prvaka iz mesa in krvi.

Vdihovanje razuma v stroje se še vedno zdi podobno iskanju svetega grala, saj v tem času pravzaprav ni bilo resničnega preskoka, kljub obilici poskusov. Dr. Grobelnik pravi, da je v zadnjem času v svetu sicer opaziti napredek in povečano zanimanje za to področje, kar se pozna tudi po količini vloženega denarja.

"Odkrili smo tehnologije in postopke, ki uporabnikom prihranijo veliko časa, na primer detekcijo vsebin, strojno prevajanje, vizualizacijo vsebine, sumarizacijo, izločanje ključnih informacij iz besedila, še vedno pa smo skoraj enako oddaljeni od samega razumevanja vsebine," ugotavlja dr. Grobelnik.

Različne pristope do problema rad ponazori s prispodobo o slepcih, ki so odkrivali slona. Vsak ga je otipaval po svoje, tistemu pri repu se je zdel podoben vrvi, drugi je držal v roki rilec in rekel, da je podoben kopju, tretji je tipal njegovo telo in mislil, da je slon kot zid. "Podobno je z razumevanjem vsebine, lingvisti menijo, da gre za vprašanje jezika, raziskovalci sodobnega spleta govorijo o skupnosti, tu so semantiki..."

Praktičen primer vsakodnevnega preseka računalništva in razumevanja besedila so iskanja v spletnih brskalnikih. Če v enega od njih vnesete besedo jaguar, boste dobili 84 milijonov zadetkov oziroma povezav na spletne strani, kjer se beseda pojavlja. Iskalna tehnologija je že zelo dovršena, toda še vedno svetlobna leta daleč od tega, da bi bila pametna.

Prav iskanje pomena je eden od ciljev svetovnega spleta naslednje generacije 3.0, imenovanega semantični internet. Če trenutno vnesete v spletni iskalnik stavek "Želim si na počitnice v tople kraje, imam 2000 evrov in 11-letnega otroka...", bo Google poleg kupa reklamnih povezav na počitniške agencije našel le dva odgovora, Večerov TV vodič z Bernardo Žarn na naslovnici in star izvod Dolenjskega lista. Tudi ob bolj rafiniranih iskalnih pojmih ste še vedno obsojeni na dolgotrajno kopanje skozi desetine, če ne stotine strani s ponudbami hotelov in letalskih povezav. Splet 3.0 naj bi po idealni predstavi ob zgornjem stavku vrnil seznam pripravljenih počitniških paketov, kot bi jih posebej za vas pripravili v turistični agenciji.

Dr. Grobelnik poudarja, da je za razumevanje besedila treba preseči prepoznavanje črk in besed. Potreben je sistem znanja, nekakšen model sveta, skozi katerega prepoznavamo dejstva v besedilih, treba je razumeti kontekst posameznega besedila, pomembni so motivi in ozadja v besedilu. "V zadnjem desetletju je bilo nekaj poskusov, da bi vse to združili v sistem za računalniško razumevanje pomena, edini resen sistem, ki ga je mogoče najti na trgu, pa je Cyc, ki ga razvija ameriško-evropsko podjetje Cycorp," trdi.

Evropska izpostava Cycorpa ima sedež v tehnološkem parku Instituta Jožef Stefan in z njim zelo tesno sodeluje. "V Slovenijo smo jih pripeljali prav z namenom, da dopolnimo naše znanje na tem področju, Cycorp pa je iskal naša znanja, s katerimi sodimo med boljše v svetu. Zdaj zelo uspešno sodelujemo pri mednarodnih projektih, ki iščejo nove korake pri razumevanju vsebin," pojasnjuje dr. Grobelnik.

Po njegovih besedah je razumevanje besedil še vedno trd oreh za znanost. "Niti blizu nismo temu, kar zmore že nekajletni otrok. Še vedno praskamo po površju, vendar pa je napredek v zadnjih petih letih precejšen, zato lahko v prihodnjih petih letih pričakujemo obsežne premike na tem področju."

Prepoznati pomeni razumeti

Zatika se že na veliko bolj osnovni ravni, pri razpoznavanju govora oziroma njegovem pretvarjanju v zapisano obliko. Tudi razvijanje fonetičnega pisalnega stroja, ki bi prevzel naloge strojepisk in avtomatično zapisoval govor, ima namreč podobno dolgo brado kot iskanje računalniške pameti. **Prof. dr. France Mihelič**, vodja skupine za govorne tehnologije v Laboratoriju za umetno zaznavanje, sisteme in kibernetiko (LUKS) na ljubljanski fakulteti za elektrotehniko, pojasnjuje, da je to razumljivo, saj sta razumevanje in razpoznavanje jezika tesno povezana.

"Naloga je podobna, kot če bi poslušalcu narekovali besedilo v tujem jeziku, ki ga ne pozna, in ga prosili, naj zapiše glasove, ki jih sliši. Ugotovili bi, da bi bili rezultati razpoznave podobno slabi kot pri samodejnem računalniškem sistemu. Poskusite brez znanja in razumevanja francoščine napisati karkoli smiselnega po nareku Francoza," težavo slikovito opisuje dr. Mihelič. Po njegovih besedah govor uporabljamo zato, da se sporazumevamo, zato je razumevanje govorjenih sporočil bistveni dejavnik, ki vpliva na razpoznavo govora. Kar pomeni, da sta področji razumevanja in razpoznavanja povezani, zato razvoja uspešnega univerzalnega sistema za razpoznavanje govora, ki bi bil po svojih sposobnostih primerljiv s človekom, ne moremo pričakovati, dokler ne bo rešen problem razumevanja.

Kljub temu pa trenutni dosežki pri razpoznavanju govora močno presegajo začetne poskuse. Strokovnjaki so že

razvili samodejne postopke, ki s pomočjo statističnih metod ugotavljajo akustične lastnosti govora in sintakso posameznega jezika. Sestavljeni so iz ocenjevanja verjetnosti, da posamezne skupine prepoznanih glasov sestavljajo določene besede, nato pa še iz ocenjevanja, kako verjetno je, da se tako zaporedje besed sploh pojavlja v jeziku.

Tako je že mogoče kvalitetno razpoznavati tekoči govor z obsežnimi slovarji več kot 10.000 besed. To še posebej velja, če govorjenje poteka v zvočno kontroliranih pogojih brez motečih šumov, programi pa so prilagojeni posebnostim govorca. Z zmanjševanjem računalniškega besednega zaklada postaja sistem še bolj robusten, saj ga ni več treba prilagajati govornim lastnostim posameznika, hkrati pa se z manjšim besednjakom povečajo tudi možnosti za razmeroma uspešno samodejno analizo pomena govora.

V tujih jezikih dejansko že obstajajo programi, ki so sposobni zapisovati narek in imajo dokaj obsežno bazo besed, njihovi rezultati pa so zadovoljivi (kar pomeni 95 in več odstotkov prepoznanega govora). "Poznam prevajalca iz slovenščine v angleščino, ki si je na smučanju zlomil roko, zaradi česar se je bil primoran nasloniti na uporabo enega takšnih programov za zapisovanje besedila. In bil z njim zelo zadovoljen. Vendar to zahteva uporabo dobrega mikrofona, pravilnega govora, program pa potrebuje nekaj časa, da se prilagodi načinu govora posameznika. Učinkovitost takšne programske opreme se drastično zmanjša že pri prepoznavanju običajnega pogovora," dr. Mihelič opisuje trenutno raven tržno dosegljivega programja.

Slovanska zapletenost

Angleščina je v veliki prednosti pred drugimi jeziki, saj je bilo zanjo razvitih veliko dobro delujočih postopkov razpoznave govora. Žal se je pokazalo, da niso enako primerni za vse jezike, tako da jih ni mogoče preprosto prenašati na nove jezike. Učinkovitost obstoječih postopkov je v veliki meri odvisna od sintakse jezika, pri čemer jo slabše odnesejo jeziki z velikim številom pregibnih oblik (sklanjatve, spregatve, število, spol) besed. Med slednjimi je vsa slovanska skupina, slovenščina pa je ta že tako slabši izhodiščni položaj začinila še z našo ljubo dvojino.

"Pri angleščini mora računalnik prepoznati toliko besed, kolikor jih je v slovarju, pri slovenščini je treba vsaki posamezni besedi dodati še vse njene različne oblike zaradi sklanjatev, spregatve, spola in števila, računalnik jih mora vse ločiti med sabo. Naslednja težava je slovnična struktura besed v stavku. Pri slovenščini in podobnih jezikih je besedni vrstni red v stavku zelo svoboden, medtem ko je pri angleščini veliko bolj strogo določeno, kako si smejo slediti posamezne besede. Kar je za računalniško prepoznavanje zelo koristno," **dr. Simon Dobrišek**, asistent dr. Miheliča, našteva težave z našim maternim jezikom.

Jezikovni modeli temeljijo na prepoznavanju njim že znanih besed, kar dr. Dobrišek ponazori s programom, s katerim so poskusno simulirali prepoznavanje govora pri komunikaciji za rezervacijo letalskih kart.

"Jutri popoldne bi rad letel iz Londona za Ljubljano" je računalnik prepoznal in pridno ponovil, stavek "Ali lahko pri stari mami naročim tri pice" pa se je sfižil v digitalno jecljanje "A lahko ... u ... dora guli ... tako ... cirih cirih ...".

Velika večina svetovnega znanja in razvoja se tako osredotoča na jezik, ki ni blizu našemu, zaradi česar so novi dosežki le deloma uporabni. Da je kupica še bolj grenka, poskrbita naša slovnica in pravopis, saj ima vsako pravilo najmanj eno izjemo. "Več je izjem kot pravil," ugotavlja tudi dr. Mihelič. "Ob začetku dela smo naivno mislili, da bomo vzeli pravopis in prenesli pravila v računalniški sistem, nazadnje pa ugotovili, da je še naglasno mesto v besedi zelo težko določiti." Zato ves čas poudarja povezovanje in sodelovanje s klasičnimi jezikoslovci ter pripravo "smiselne slovnične interpretacije", ki je nujna za slovenski preboj na tehničnem področju materinščine.

Nezanimivi za velike

Dr. Mihelič ocenjuje, da smo v Sloveniji kmalu po začetku raziskav na področju prepoznavanja govora dokaj hitro dosegli stopnjo razvoja v Evropski uniji. Hkrati pa meni, da "v zadnjem času znova zaostajamo za trendi razvoja, Čehi na primer so nas že začeli prehitevati". Za raziskave govornih tehnologij so potrebne obsežne zbirke ustrezno označenega govora in njegovega zapisa. Priprava takšne zbirke zahteva ogromno dela, od desetkrat pa tudi do štiridesetkrat toliko, kot znaša dolžina posnetka, za enourni posnetek na primer tudi do štirideset ur. Seveda se takoj pojavi vprašanje stroškov, pri čemer se domači strokovnjaki srečujejo še s težavo sistema financiranja v znanosti. Denar za raziskave in raziskovalne nazive je mogoče dobiti namreč le na podlagi objav v mednarodnih revijah, "revije Nature pa ne zanimajo pretirano raziskave slovenskega jezika, kaj šele oblikovanje slovenskih podatkovnih zbirk", pravi dr. Mihelič.

Zato se dogaja, da se raziskovalci na področju govornih tehnologij raje usmerjajo v raziskave od jezika neodvisnih postopkov ali že razvite postopke preizkušajo na tujejezičnih govornih zbirkah. Kar pomeni še manjše število raziskovalcev, ki hkrati grizejo skozi enako zahtevne raziskave kot kolegi v tujini, saj te niso odvisne od števila ljudi, ki neki jezik uporabljajo. Zaradi prej naštetih posebnosti pa je njihovo delo še težje.

Dr. Mihelič poudarja, da je trenutno financiranje raziskovalnih skupin iz javnih sredstev preskromno. "Raziskav in razvoja na področju tehnične obdelave slovenskega govora ni mogoče financirati le s stališča gole tržne logike ali raziskovalne vrednosti. Gre za družbeni nacionalni interes, povezan z ohranjanjem in razvojem slovenskega jezika. Jezikoslovju v tem primeru še namenjamo določene materialne stroške, medtem ko so vlaganja v tehnične rešitve marginalna."

Pri tem navaja primer že omenjene Češke, ki nas je po vloženih sredstvih, številu raziskovalnih skupin in dosežkov že prehitela, večina raziskav pa poteka na javnih raziskovalnih ustanovah in univerzah, ki se financirajo iz državnega proračuna. Tudi precej večja Nizozemska je v zadnjih dveh desetletjih v glavnem z javnimi sredstvi izpeljala več projektov za pripravo obsežnih nacionalnih govornih in tekstovnih zbirk, Japonska pa se je z 200 milijoni dolarjev večinoma državnega denarja v zadnjem desetletju lotila doslej verjetno najbolj temeljitega in obsežnega projekta zbiranja in dokumentiranja, s posebnim poudarkom na spontanem govoru. "Pri nas še ni prodrlo spoznanje, da je avtomatična obdelava slovenskega jezika prav tako pomembna kot pisanje slovarjev in slovnic," meni dr. Mihelič.

Jezik je eden najpomembnejših nosilcev nacionalne identitete, zahtevnih sistemov za razpoznavanje in sintezo slovenščine pa tujci za nas ne bodo razvili - tako zaradi nepoznavanja slovenskega jezika kot premajhnega tržišča. Dober primer je virtualna davčna svetovalka Vida, saj podpora, ki jo je tuji partner razvil za govorno komunikacijo na samodejnem davčnem telefonskem odzivniku, ni delovala. Dr. Dobrišek navaja primer Microsofta, s katerim se je poskušal dogovoriti za razvoj slovenskega govornega pogona v operacijskem sistemu Windows. "Najprej jih je zanimalo, potem so verjetno ugotovili, kako velika je Slovenija, in raven pogovora se je hitro prenesla na manj pomembne razvijalce v podjetju, nazadnje pa so dejali, da bi lahko sodelovali le pri razvoju angleških govornih modulov, saj je slovenščina zelo nizko na prioritetni listi," opisuje dr. Dobrišek svoje izkušnje.

Če bomo želeli v prihodnosti komunicirati z umetnimi sistemi v svojem jeziku, bo treba tudi pri nas vlagati v njihov razvoj. Drugače bomo najverjetneje prisiljeni uporabljati angleščino. "Mogoče bo za mlajše generacije to celo 'fensi', po mojem pa je to veliko bolj nevarno, kot so recimo tujejezična imena podjetij ali naslovi v časopisih," opozarja dr. Mihelič.

Slovenska babilonska ribica

*S trdim orehom prepoznavanja govora se spopada tudi slovensko podjetje Alpineon. V idiličnem kotu Letenic pri Golniku se skupina petih razvojnikov pod vodstvom **dr. Jerneje Žganec Gros** ukvarja z razvojem jezikovnih tehnologij. Izhajajo iz "gnezda" prof. dr. Franceta Miheliča in Laboratorija za umetno zaznavanje, sisteme in kibernetiko, moči pa vlagajo v raziskave in razvoj prepoznavanja govora, strojnega prevajanja in sintezo govora. "Uspešno že tržimo projekt sintetizatorja slovenskega govora. Naš modul govori razumljivo in že kolikor toliko naravno, ga pa še dodatno izboljšujemo. Prvenstveno je namenjen slepim in slabovidnim, kupujejo pa ga tudi na primer telekomunikacijska podjetja," prve uspehe podjetja razkriva dr. Žganec-Grosova.*

Čeprav nameravajo vse tri sisteme - za prepoznavanje, prevajanje in tvorjenje govora - tržiti tudi posamično, pa je njihov osrednji cilj glasovni komunikator VoiceTRAN, ki naj bi omogočal avtomatsko prevajanje iz enega v drug jezik. Radi bi naredili žepno napravo, v katero bo mogoče govoriti v enem jeziku, nakar bo naredila prevod v drugega, uporabna pa naj bi bila v obe smeri. "Ja, resnično nekakšna babilonska ribica," se direktorica razvoja nasmeji ob primerjavi z vsevedno prevajalsko živalco iz kultne knjige Štoparski vodnik po galaksiji.

Njegova uporabnost bo zagotovo zelo široka, precej zanimanja zanj je pokazala Slovenska vojska, ki je eden od glavnih investitorjev v razvoj. "Vojska naj bi ga uporabljala za humanitarne namene, denimo za komunikacijo s civilnim prebivalstvom na mirovnih misijah," poudarja prva dama Alpineona.

Med njihovo "premoženje" sodi tudi digitalni slovar izgovarjav, ki so ga zgradili v sodelovanju z Inštitutom za slovenski jezik pri ZRC SAZU. "Slovenščina ima prosto naglasno mesto, kar pomeni, da je računalniku za vsako besedo treba povedati, kje je naglas. Vsaka beseda v slovarju izgovarjav ima poleg fonetičnega prepisa določeno

tudi naglasno mesto," poudarja dr. Žganec-Grosova.

Poleg "običajnih" težav s prepoznavanjem govora in napakami, ki se ob tem pojavljajo, ledino orjejo tudi pri strojnem prevajanju. Napake pri prepoznavanju se med strojnim prevajanjem samo še multiplicirajo. Čeprav to velja za tehnologijo prihodnosti, se "tudi v tujih strokovnih krogih še vedno sliši kot neke vrste znanstvena fantastika", pravi dr. Žganec Grosova. Prevajalniki temeljijo na statističnem modelu, pri katerem najprej poiščejo veliko količino že prevedenih tekstov, prevajalnik pa se nato sam uči prevajalnih pravil. Na srečo je ob našem vstopu v EU nastalo ogromno takšnega učnega gradiva. Model še vedno izpopolnjujejo, trenutno dosegajo okoli 60-odstotno natančnost na omejeni aplikacijski domeni.

"Pri VoiceTRANU je ta natančnost precej višja, saj je nabor besed omejen le na nekaj tisoč. Imamo že dvosmerni prevajalnik v angleščino, za vojsko razvijajo še modula za arabščino in albanščino, sistem pa je zasnovan tako, da je mogoče relativno preprosto dodajati nove jezikovne pare," pojasnjuje sodelavec Alpineona ***dr. Aleš Mihelič*** *(ki ni v sorodu s profesorjem Miheličem).*

Pri delu sodelujejo s tremi fakultetami Univerze v Ljubljani in Institutom Jožef Stefan, dr. Jerneja Žganec Gros pa meni, da je bilo v zadnjem času na voljo precej razvojnih pomoči. "Brez vložka države se verjetno ne bi upali lotiti tako zahtevnega projekta, kot je VoiceTRAN," pravi in dodaja, da jih recesije ni strah, saj so si sredstva za dokončanje projekta, ki naj bi ga zaključili v dveh letih, že zagotovili iz tržnih projektov. Hkrati je prepričana, da je prav recesija čas za razvoj novih tehnologij za zalogo. Po njej bo prišel čas za njihovo trženje.

Komentarji