ALTERNATOR
Misliti znanost.

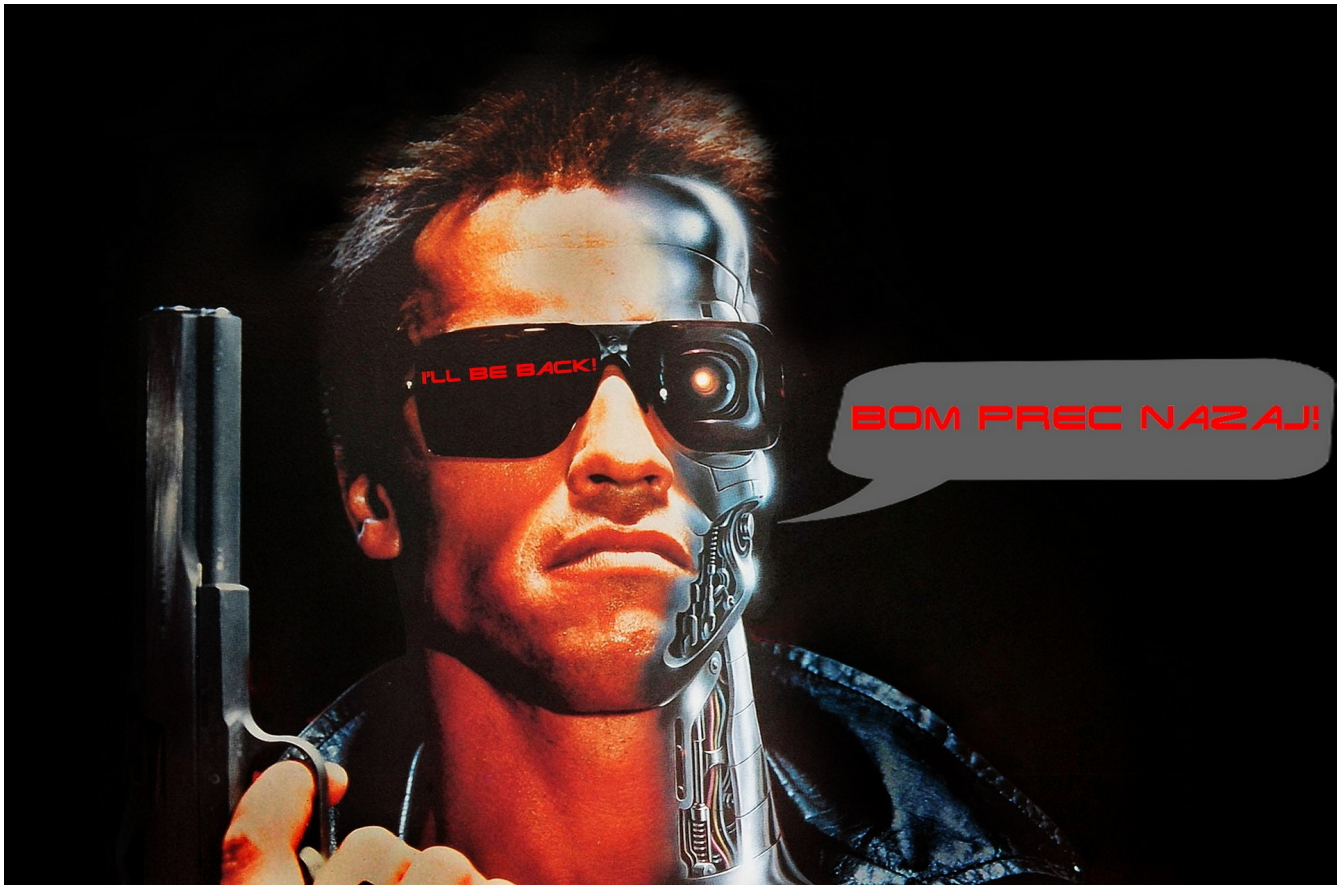
Z nevronščino v prihodnost

9. 12. 2021

Number: 49/2021

Author:

- Špela Vintar



Računalniška obdelava: Katja Bidovec

V zadnjih letih je marsikdo opazil, da prevodi Googlovega prevajalnika niso več neusahljiv vir zabave in smeha, ampak se tudi daljše in kompleksnejše povedi pretvorijo v presenetljivo tekoče in razumljivo besedilo. Kaj se je zgodilo na področju strojnega prevajanja in drugih jezikovnih tehnologij in kaj lahko pričakujemo v prihodnosti?

Če najprej zavrtimo čas tri četrt stoletja nazaj ali si v spomin prikličemo odličen film *Igra imitacije*, se srečamo z Alanom Turingom, matematikom in kriptografom, ki ni bil le avtor Ultre, ki je dešifrirala Enigmo, stroj za dešifriranje, s katerim se je jeziček na tehtnici moči med drugo svetovno vojno prevesil na stran zaveznikov, ampak danes velja za očeta sodobnega računalništva in vizionarja umetne inteligence. Predvidel je, da bo proti koncu stoletja računalniška inteligenca postala del vsakdanjosti, da bodo »stroji komunicirali med seboj in si tako ostrili um«. Vprašanje o zmožnosti strojev, da mislijo, pa je preoblikoval v »igro imitacije«: če izolirani spraševalec v komunikaciji z osebo A in osebo B, od katerih je ena stroj, ni zmožen prepoznati stroja, je stroj uspešno prestal preskus mišljenja. Številni strojni prevodi bi danes gladko prestali Turingov test, saj jih po razumljivosti in berljivosti ni mogoče razlikovati od človeških. V nadaljevanju pojasnjujemo glavne razvojne mejnike, ki so do tega pripeljali, in se ukvarjamo z vprašanjem, ali strojni prevajalniki resnično že dosegajo človeško spretnost pri prenašanju sporočil iz enega jezika v drugega in kaj to pomeni za človeške prevajalce ter jezikovne poklice nasploh.

Od pravil do podatkov

Prvih nekaj desetletij razvoja strojnih prevajalnikov je minilo v prizadevanjih, da bi vso leksikalno, slovnično in semantično kompleksnost naravnih jezikov računalniško modelirali s pravili, ki bi na posameznih nivojih analize poskrbela za ustrezno

pretvorbo iz izhodiščnega v ciljni jezik. Čeprav so bili tedanji prevajalniki z današnjega vidika še daleč od uporabnosti, je najbolj znani prevajalnik stare generacije Systran leta 1997 zaživel tudi na rosno mladem svetovnem spletu v okviru iskalnika AltaVista. Za slovenščino pa je prvi tak prevajalnik razvilo jezikovnotehnološko podjetje Amebis iz Kamnika.

Konec devetdesetih let prejšnjega stoletja je v raziskavah prišlo do prvega pomembnega mejnika s prehodom na statistične strojne prevajalnike, ki niso več potrebovali dvojezičnih slovarjev in slovničnih pravil. Delovali so na podlagi strojnega učenja in se urili na obstoječih zbirkah že prevedenih besedil in velikih enojezičnih korpusih. Ti prevajalniki so temeljili na verjetnostnem modelu prevodne ustreznosti, ki je bil naučen na vzporednih in stavčno poravnanih dvojezičnih besedilih in je za vsako besedo ali besedno zvezo predvidel najverjetnejši prevod. Za slovnično pravilnejše ciljno besedilo pa je poskrbel verjetnostni jezikovni model, izdelan iz velikih količin podatkov za ciljni jezik. Tako je leta 2006 Google objavil svojo prvo različico spletnega prevajalnika, ki je deloval po statistični metodi, med jeziki, ki jih je vključeval, pa se je leta 2008 znašla tudi slovenščina. Frazni model statističnega prevajanja, ki je kmalu prerasel v najbolj priljubljeno metodo in se je uveljavil ne le pri Googlu, ampak tudi pri številnih nekomercialnih sistemih, ki temeljijo na ogrodju Moses (https://www.statmt.org/moses/) , je zaradi jezikovnega modela presenetljivo dobro reševal oblikoslovne zakonitosti posameznih besednih zvez in fraz. Težko bi se denimo primerila napaka, da bi »black cat« v slovenščino prispela kot »črni mačka«, saj je takšna besedna zveza za slovenščino netipična. A po drugi strani ti sistemi niso imeli nikakršnih mehanizmov za upoštevanje širšega konteksta, saj je bilo frazno modeliranje omejeno na kombinacije do pet ali šest besed.

Ali govorite nevronsko?

Prvi nevronski prevajalniki so se pojavili okrog leta 2014 in naglo nadomestili prejšnjo tehnologijo, saj so bili njihovi prevodi precej bolj berljivi. Nevronska mreža je način procesiranja pri globokem učenju, tj. vrsti strojnega učenja, ki se uporablja za izdelavo najkompleksnejših napovednih modelov. Po zgradbi spominja na ustroj človeških možganov in biološko delovanje nevronov, saj gre za nelinearno obdelavo podatkov na več medsebojno povezanih nivojih omreženih vozlišč. Nevronske mreže so precej zmogljive za verjetnostno modeliranje, pri katerem gre za napovedovanje verjetnosti določenega ciljnega niza ob danem vhodnem nizu. Zato se jih uspešno uporablja za najrazličnejše jezikovne naloge, od razpoznavanja govora, generiranja odgovorov na vprašanja, pa vse do prevajanja in – v zadnjem času – tolmačenja. Nevronski strojni prevajalnik je sestavljen iz kodirnika, ki vhodni niz pretvori v številsko predstavitev oziroma vektorsko vložitev (angl. *word embedding*), in dekodirnika, ki generira prevod besedo za besedo, pri vsaki besedi pa spremlja podobnost ciljnega vektorja z izhodiščnim. Ker se tako v fazi učenja kot pri samem prevajanju obdelujejo cele povedi, se ne dogaja več, da prvi del povedi ne bi bil skladen s preostalim delom povedi. Prav tako se pri prevajanju vsake posamezne besede vedno upošteva kontekst celotne povedi. Morda je še najbolj fascinantna lastnost nevronskih jezikovnih tehnologij, da se pomen modelira s pomočjo vektorskih vložitev. To je abstraktna številska predstavitev besede, ki se izračuna iz vseh pojavitev v različnih kontekstih in besedo umesti v mnogorazsežnostni semantični prostor. V tem prostoru se besede s podobnim pomenom – tudi v različnih jezikih – znajdejo druga blizu druge, pomenska razmerja med sorodnimi besedami pa si je mogoče predstavljati kot matematične preslikave med vektorji. Novejše metode za računanje vektorskih vložitev znajo upoštevati tudi vsak posamezen kontekst in so zato boljše pri prevajanju dvoumnih besed.

Nevronski prevajalniki pa imajo vgrajeno še eno novost, in sicer posebno metodo za obdelavo neznanih besed. Naravni jezik je namreč živ in se neprenehno spreminja, tako da tudi ob ogromnih količinah učnih podatkov ne moremo biti prepričani, da se prevajalnik pri novih besedilih ne bo srečal s kako besedo, ki je v učnih podatkih ni bilo. V fazi učenja tako nevronski prevajalnik zgradi vektorske vložitve ne le za cele besede, ampak tudi njihove krajše sestavne dele. Ko se v novem besedilu pojavi neznana beseda, se njena vektorska reprezentacija zgradi iz posameznih delov, prav tako se po delih generira tudi prevod.

Tudi Google se uči

Izvirnik

A new soft bot mimics octopuses and inchworms to climb walls

This kind of free-climbing machine, described April 3 at the Materials Research Society spring meeting, could one day help conduct surveillance or inspect buildings and bridges.

Google Translate 2016

A nove mehke bot posnema hobotnice in inchworms plezati stene

Ta vrsta prosto plezanje napravo, opisano 3. april na spomladanskem srečanju Materials Research Society, bi lahko nekega dne pomoč ravnanje nadzora ali pregleda objektov in mostov.

Google Translate 2018

Novi mehki bot posnema hobotnice in pentlje, da se plezajo po stena

Ta vrsta prostega plezalnega stroja, opisanega 3. aprila na spomladanskem sestanku družbe Material Science Society, bi lahko nekega dne pomagal opraviti nadzor ali pregled zgradb in mostov.

Google Translate 2021

Nov mehki bot posnema hobotnice in palčaste črve za plezanje po stenah

Takšen stroj za prosto plezanje, opisan 3. aprila na spomladanskem zasedanju Društva za raziskave materialov, bi lahko nekoč pomagal pri nadzoru ali pregledovanju stavb in mostov.

Primeri Googlovih prevodov iz let 2016, 2018 in 2021 ponazarjajo skokovit napredek, ki so ga prinesle nevronske tehnologije. Če so bili statistični prevodi še okorni in zaradi številnih slovničnih napak in manjkajočih stavčnih členov le stežka berljivi, je že prva nevronska različica Googlovega prevajalnika za slovenščino opazno boljša, vsaj kar zadeva berljivost. A tudi nevronščina ni brez težav, saj si prevod »pentlje« za »inchworms« lahko pojasnimo le s prevajanjem po delnih vektorjih, skoraj nevarno neopazna pa je sprememba, ki si jo prevajalnik privošči pri nazivu organizacije Materials Research Society. Naziv v prvi nevronski različici ostane nepreveden, kar bi bilo povsem ustrezno, če ne bi prevajalnik »Research« skrivoma spremenil v »Science«.

Če skušamo kakovost strojnih prevajalnikov izraziti na merljiv in eksakten način, uporabimo eno od številnih metod za evalvacijo, med katerimi je še najbolj znana metrika BLEU (https://aclanthology.org/P02-1040/) . Ta strojni prevod ocenjuje na podlagi precej primitivne primerjave z referenčnim (človeškim) prevodom in mu podeli oceno od 1 do 100. V praksi velja, da so prevodi z ocenami, nižjimi od 10 precej neuporabni, med 20 in 35 uporabni, a s precej napakami, od 40 dalje pa imamo opravka z zelo berljivimi, slovnično in pomensko ustreznimi prevodi. Na zadnji svetovni konferenci o strojnem prevajanju WMT21 (https://www.statmt.org/wmt21/) so predstavljeni sistemi za številne jezikovne pare dosegali oceno BLEU okrog 50. Toda pri tem je treba upoštevati, da ta metrika pravzaprav meri podobnost z referenčnim prevodom, skoraj vsako avtentično besedilo pa ima več možnih prevodov, zato neujemanje ne pomeni nujno neustreznega prevoda.

Tudi za slovenščino sodobni prevajalniki, kot so Google Translate, DeepL in eTranslation, dosegajo oceno okrog 40, v okviru projekta Razvoj slovenščine v digitalnem okolju (https://slovenscina.eu/) , ki ga financira Ministrstvo za kulturo prek Evropskega sklada za regionalni razvoj, pa konzorcij slovenskih raziskovalnih institucij razvija lasten angleško-slovenski prevajalnik.

Uhan, ki prevaja

Glede na nagel razvoj tehnologij se kar samo postavlja vprašanje, ali bomo v prihodnosti sploh še potrebovali človeške prevajalce, pa tudi, ali se bodo ljudje še vedno učili tujih jezikov. Precej vizionarski pogled na prihodnost komunikacije v dobi umetne inteligence podaja dokument Obzorje človeško-strojnega veka: napoved novih in prihodnjih jezikovnih tehnologij (https://lithme.eu/wp-content/uploads/2021/05/The-dawn-of-the-human-machine-era-a-forecast-report-2021-final.pdf) , ki je nastal v okviru programa COST s sodelovanjem številnih univerz in raziskovalnih inštitutov. Avtorji uvodoma opominjajo na dejstvo, da se nam ni treba spraševati, kdaj bodo jezikovne in govorne tehnologije vstopile v naš vsakdanjik, saj so v njem že nekaj časa prisotne. Razpoznavanje in tvorjenje govora je tako vključeno v vsak pametni telefon in malo boljši avto, našo digitalno komunikacijo pa preko neštetih spletnih storitev neprestano analizirajo in uporabljajo za različne trženjske namene. V prihodnosti lahko pričakujemo, da bodo pripomočki za brezšivno večjezično pisno in govorno komunikacijo postajali vse boljši, pa tudi vse manjši in manj opazni, da jih bo mogoče integrirati v obleko, modne dodatke ali samo telo. Pričakujemo lahko tudi velike korake pri obdelavi in modeliranju gestike, mimike in čustev, na drugi strani pa razvoj jezikovnih pripomočkov za osebe s posebnimi potrebami.

Strojno prevajanje torej že danes vpliva na naše jezikovno vedenje, še posebej na spletu. A kljub naglemu razvoju za zdaj ni videti, da bi lahko nalogo prevajanja tudi v profesionalnih okoljih v celoti zaupali računalnikom. Pogosti argumenti, ki jih prevajalci uperijo proti svojim strojnim tekmecem, so nezmožnost upoštevanja širšega konteksta (tj. daljšega od posamezne povedi), nedoslednost pri prevajanju terminologije in drugih ustaljenih izrazov, »zahrbtnost« nevronskih napak, ki so pogosto skrite in na prvi pogled neopazne, ter slogovna in kulturna neustreznost. K temu lahko dodamo še vseprisotni očitek umetni inteligenci, da deluje pristransko (https://arxiv.org/pdf/1906.00591.pdf) in v družbi še poglablja že obstoječe neenakosti in predsodke. Pozivi k večji regulaciji in nadzoru nad etičnostjo razvijajočih se tehnologij prihajajo od vsepovsod, na področju prevajanja pa se podobno kot na drugih področjih dogaja, da se ljudje počutijo le še kot podatkovna hrana za požrešne algoritme, saj se prevajalniki učijo iz človeških prevodov.

Čeprav na mnoga vprašanja ni enoznačnih odgovorov, lahko vsaj glede prihodnosti jezikovnih poklicev ostanemo optimisti, saj so potrebe po takšnih in drugačnih jezikovnih profilih vse večje, res pa bodo mnogi od prihodnjih poklicev izrazito

tehnološko obarvani.

https://www.alternator.science/en/long/z-nevronscino-v-prihodnost/