Izvorni članak UDK 004.838.3(045)
doi: 10.21464/fi38101
Primljeno 14. 9. 2017.

Sandro Skansi,[1] Davor Lauc[2]

[1] Sveučilište u Zagrebu, Hrvatski studiji, Kampus Borongaj, Borongajska cesta 83d, HR–10000 Zagreb
[2] Sveučilište u Zagrebu, Filozofski fakultet, Ivana Lučića 3, HR–10000 Zagreb
[1] sskansi@hrstud.hr, [2] dlauc@ffzg.hr

Analogijsko zaključivanje i značenja riječi u višedimenzionalnom prostoru

Sažetak

Rad istražuje temeljnu misao i pretpostavku simboličke logike oko pojmova kao atomarnih komponenti (uvedenih prilikom definiranja sustava i koji se ne mogu dalje razlagati), i uvodi drugačiji formalizam, baziran na umjetnim neuralnim mrežama za formalizaciju logičkog zaključivanja kao kognitivnog procesa, što definira pristup koji nazivamo subsimboličkom logikom primijenjenoj na analogijsko zaključivanje kao punopravnom obliku zaključivanja. Istražujemo i kognitivne aspekte takvog pristupa, posebice u kontekstu izolacije i reprodukcije spontanih, ali neispravnih formi zaključivanja (logičkih pogreški) svojstvenih logičkom zaključivanju kao kognitivnom procesu. Ovo je danas dominantna tehnika u umjetnoj inteligenciji, no filozofijske su posljedice ovog pristupa u potpunosti neistražene. Prema našim spoznajama, ovo je prvi pokušaj da se uz pomoć umjetnih neuralnih mreža analizira fenomen analogijskog zaključivanja.

Ključne riječi

analogijsko zaključivanje, značenje riječi, umjetne neuralne mreže, neuralni jezični modeli, kognitivni konekcionizam, subsimbolička logika

Uvod

Premda je »logika« izrazito polisemična riječ, bremenita brojnim konotacijama, u suvremenoj filozofiji, matematici, jezikoslovlju, računarstvu te drugim disciplinama koje istražuju logiku, logika se poglavito razumijeva kao formalna, deduktivna disciplina, nepovezana s problemom ljudskog zaključivanja. Ono što nazivamo problemom ljudskog zaključivanja prvenstveno je problem formalne replikacije ljudskog zaključivanja, odnosno njegovih komponenti. Ova potraga za replikacijom zahtijeva jedan deskriptivan pristup. Ovo ističemo zato što je dominantan pristup u logici u dvadesetom stoljeću bio normativan, a takvo razumijevanje logike prvenstveno dugujemo glavnom utemeljitelju suvremene logike Gottlobu Fregeu i njegovoj kritici psihologizma. Takav istraživački program logike duhovito je sažeo Bertrand Russell u definiciji logike koja mu se pripisuje, u kojoj se navodi da je ona »predmet u kojem nitko ne zna o čemu razgovara, niti je li to što je izrečeno istinito« (Boden 1988), bez problematiziranja problema toga kako znamo da su početne tvrdnje istinite, ali ni kakve to ima veze sa sadržajem tvrdnji ili načinom kako ljudi zaključuju.

Usprkos razvoju brojnih alternativnih logika koje se djelomično udaljuju od ovog prihvaćenog istraživačkog programa i istražuju ne-deduktivno zaključivanje (induktivne, neizrazite, probabilističke, abduktivne i druge logike),

te zaključivanje koje uzima u obzir sadržaj iskaza ili ljudska ograničenja (relevantne, linearne i druge substrukturalne logike), smatramo kako glavni recentni izazov prihvaćenom razumijevanju logike dolazi iz područja umjetne inteligencije i kognitivne znanosti.

Uz pretpostavku da logika proučava i formalizira kako ispravno zaključivanje tako i ljudsko zaključivanje, filozofijska logika, poznata u matematičkim krugovima kao neklasična logika, formalizira upravo raznolike aspekte ljudskog zaključivanja. Možda je ovdje najistaknutija *fuzzy logika* koja pokušava formalizirati ideju »računanja s riječima« (*computing with words*), odnosno logika prirodnog jezika (usp. Zadeh 1996). Iako usko povezano s našim pristupom, nećemo ulaziti dublje u sličnosti s fuzzy logikom.

U kontekstu umjetne inteligencije, jednu od uloga logike možemo razumjeti kao nadilaženje jaza između čovjeka u kojem se zaključivanje nalazi kao empirijski fenomen i stroja koji bi trebao imitirati ljudske sposobnosti. Dakako, ovo nisu jedine primjene logike u umjetnoj inteligenciji. Štoviše, spadaju u računalno manje važne, ali iz kognitivnog aspekta su među najzanimljivijima. Radi li stroj imitacijom isto što i čovjek kada zaključuje, veliko je i otvoreno pitanje na koje ne pokušavamo ni djelomično odgovoriti u ovom radu. Ovdje nas prvenstveno zanimaju opće strukture zaključivanja, posebno one najjednostavnije i najrudimentarnije poput analogijskog zaključivanja. Premda je analogijsko zaključivanje, promatrano kao kognitivni i pragmatički proces, vrlo složeno, promatrano kao oblik simboličkog zaključivanja, analogijsko je zaključivanje najjednostavniji oblik zaključivanja.

Unutar područja umjetne inteligencije danas prevladava mišljenje da se ovo zaključivanje može naučiti iz iskustva, a mi u ovom radu istražujemo uvjete mogućnosti takova procesuiranja kod strojeva. Ti su uvjeti dijelom zadani algoritmima, a dijelom naučeni iz podataka. Znakovito je to kako te empirijski najuspješnije istraživačke metodologije umjetne inteligencije karakterizira upravo napuštanje izravne primjene logike i zaključivanja putem pravila, ali i odmak od pokušaja imitiranja načina kako ljudi zaključuju. Jedno od zanimljivih pitanja jest: predstavljaju li novi pristupi u kognitivnoj znanosti i umjetnoj inteligenciji, poput dubokih umjetnih neuralnih mreža, bolji pristup za rješavanje tradicionalnih filozofskih problema zaključivanja i time definiraju pristup koji bi se mogao nazvati *subsimboličkom logikom*?

Logika kao empirijska i deskriptivna

Pitanje o tome je li logika empirijska postavili su Hilary Putnam (1969) i Michael Dummett (1978). Obje su kritike imale nešto zajedničko: kritizirale su nužnost aksioma klasične logike i potrebu njihove revizije zbog otkrića kvantne mehanike. Sličnu ideju imao je i Benno Erdmann (1892). Naša kritika nije kritika pojedinih aksioma, nego preispitivanje je li ljudsko zaključivanje sustav koji je moguće adekvatno formalizirati pomoću aksioma i pravila ili su potrebno radikalno drugačiji strukturni modeli. Mi želimo istraživati zaključivanje kakvo ono jest, što znači da treba kritički prosuditi koji su formalizmi u stanju reproducirati ljudsko zaključivanje i tada se njima prikloniti. Vrlo je zanimljivo pitanje jesu li kognitivne pristranosti u zaključivanju usko vezane uz druga ljudska kognitivna ili afektivna stanja. Premda vjerujemo da je odgovor potvrdan, ovim se smjerom istraživanja ne bavimo u ovom radu. Ovome se može dati znatno jači naglasak unutar našeg pristupa koji značajno približava logiku filozofiji uma i epistemologiji i predstavlja teorijsku i kognitivnu podlogu subsimboličkoj logici. Prvi je psihologizam zastupao John

Stuart Mill u svojem *System of Logic* (1843), pri čemu je opisao logiku kao »znanost o zaključivanju«, ali i kao »umijeće zaključivanja«.

Psihologizam je uglavnom bio derivirana teza, ali i teza koja je pokušala logiku podvesti pod psihologiju. Povijesno je to imalo tri smjera argumentacije. Prvi, koji dugujemo Theodoru Lippsu (1893) i Gerardusu Heymansu (1894, 1905) polazi od pretpostavke da je psihologija disciplina koja proučava *sve* zakone mišljenja, no smatramo da je ovo vrlo nespretno jer u to doba nije postojala psihologija kao znanost odvojena od filozofije, ali je već postojala znanost i disciplina u filozofiji koja je proučavala zakone mišljenja – logika. Zbog toga se ovdje javlja jedna specifična cirkularnost koja nam (uz poprilično benevolentno tumačenje) ukazuje na neprihvatljivost ove teze. Druga teorija, koju su postavili Wilhelm Jerusalem (1905) i Christoph von Sigwart (1904), puno je bliža našoj tezi, a zasniva se na tome da je zaključivanje mentalni proces. Mi se, međutim, razlikujemo od Jerusalema i Sigwarta u tome što smatramo da je zaključivanje, osim ljudskog mentalnog procesa, i računalni proces koji se može realizirati u računalu u obliku simboličkog sustava kakvi su poznati već pedeset godina, ali i u obliku subsimboličkog pristupa učenog nad podatcima, poput na primjer umjetne neuralne mreže. Preciznije rečeno, smatramo da je zaključivanje i kod ljudi i u strojevima isti simbolički proces, koji se samo realizira na drugačijem supstratu. Psihologija se bavi isključivo mentalnim fenomenima nad biološkim supstratom, dok mi razmatramo fenomen zaključivanja u općoj formi, kao konekcionistički opisiv kognitivni proces, realiziran uz pomoć umjetnih neuralnih mreža, koje su u isto vrijeme pojednostavljeni formalni model biološkog neurona, ali i ostvarive kao računalni proces.

Može se ustvrditi kako podvođenje logike pod psihologiju, kako su to predlagali rani psihologisti u logici, u potpunosti negira njenu prirodu. Danas sa sigurnošću možemo ustvrditi da je logiku, kao formalni okvir za zaključivanje, u bilo kojoj njenoj formi moguće replicirati na računalu bez gubitka koji nazivamo *subjektivnim doživljajem*. Računalni model zaključivanja građen je prema ljudskom zaključivanju (čak se u nekim zadatcima i uspješnost rješavanja mjeri s obzirom na sličnost s ljudima, pri čemu je Turingov test najistaknutiji primjer), i time se dobiva osnova za usporedbu. Simulacija drugih kognitivnih procesa, poput emocija, može biti kritizirana iz perspektive gubljenja subjektivnog doživljaja. Želimo naglasiti da je zaključivanje tradicionalno shvaćeno kao odvojeno od konkretnog sadržaja, što znači da *subjektivni doživljaji* ne igraju ulogu, odnosno nije potrebno moći replicirati *te doživljaje* da bismo mogli ustvrditi da se u stroju odvija zaključivanje kao kognitivni proces. Ono što je ovdje zanimljivo naglasiti jest da je, prema shvaćanju simboličke logike, jedino *ispravno* ono zaključivanje koje je lišeno *subjektivnih doživljaja*, dok je pogrešno zaključivanje ono koje je karakterizirano vrlo specifičnim logičkim pogreškama, prožeto subjektivnostima, ali u jednom bitno interpersonalnom smislu jer se logičke pogreške, poput na primjer afirmacije konzekventa, ponavljaju kod različitih ljudi koji nisu ni u kakvom kontaktu. To govori u prilog tezi da su logičke pogreške bliske ljudskom razmišljanju i sastavni su dio zaključivanja kao kognitivnog procesa.

Subsimboličkim pristupom ovaj se fenomen može očuvati na način da se repliciraju logičke pogreške. Premda ovo djeluje nepoželjno, logičke pogreške sastavni su dio zaključivanja *kao* kognitivnog procesa i bilo koji formalan sustav koji želi modelirati zaključivanje *kao* kognitivni proces mora moći modelirati i karakteristične pogreške u tom procesu jer su one sastavni dio procesa zaključivanja *kao* kognitivnog procesa. Ovime se postiže bolji for-

malni opis zaključivanja jer se naglasak pomiče s ispravnog zaključivanja na ljudsko zaključivanje. Logičke pogreške nisu pogreške formalnog sustava već vrlo precizne devijacije koje su karakteristične za ljudsko zaključivanje i zbog toga su sastavni dio zaključivanja kao kognitivnog procesa koji želimo opisati toliko precizno da se može replicirati na drugim supstratima koji nisu ljudski um. Ono što predlažemo u ovom radu jest jedna nova vrsta psihologizma, u bitnome različita od onoga protiv kojeg su se Frege i Husserl borili, a taj tip psihologizma utemeljen je na pokušaju modeliranja novim, neuralnim, formalnim alatima sličnima onima kakvima ljudska bića razmišljaju.

Više dimenzija i vektori

Ključna ideja za formalnu realizaciju subsimboličke logike, odnosno formalnog sustava zaključivanja temeljenog na umjetnim neuralnim mrežama, distribuirana je reprezentacija pojmova, koju ćemo kasnije pojasniti. Distribuirane reprezentacije pojmova reprezentacije su pojma ne u obliku specifičnog simbola (koji može biti prikazan kao jedan član liste odnosno vektora s bulovskim vrijednostima), nego u obliku vektora s realnim brojevima (bez nule). Osnovni fizikalni model stvarnosti koji se danas koristi temeljen je na specijalnoj teoriji relativnosti (Einstein 1905) koja pretpostavlja četvoro-dimenzionalan Minkovskijev prostor. Ovime želimo naglasiti da, premda nisu bliski fizikalnoj intuiciji, modeli s više dimenzija koriste se već više od sto godina u drugim disciplinama, no u filozofiji nikada nisu bili pretjerano popularni, prvenstveno radi želje da se pojam vektora fizikalno interpretira, pri čemu se četvrta dimenzija poistovjećivala s vremenom. No kako su vektori matematički pojmovi, potrebno ih je na taj način razumjeti da bi smo se odmakli od privida intuitivnosti nastalog zbog njihove fizikalne interpretacije.

Iz računalne perspektive, vektori se mogu shvatiti kao uređena n-torka ili lista koja ima dodatna svojstva koje obične n-torke ili liste nemaju. U ovom nam radu ta dodatna svojstva nisu važna, a sve što je ovdje izneseno *a fortiori* vrijedit će i za potpuno rigorozno definirane vektore. Ako su x1,x2,...,xn proizvoljni realni brojevi, tada strukturu (x1,x2,...,xn) nazivamo n-dimenzionalnim vektorom. Ako govorimo o dvodimenzionalnim vektorima oblika (x1,x2), tada ih intuitivno možemo zamisliti kao točke u ravnini. Trodimenzionalne vektore oblika (x1,x2,x3) možemo predočiti kao točke u prostoru, a četvoro i više-dimenzionalne vektore zamisliti kao točke u hiperprostoru.

No četvoro-dimenzionalnom prostoru je i dalje moguće dati limitiranu, ali intuitivno bližu, vizualnu reprezentaciju. Klasičan je školski primjer da zamislimo tri dimenzije i dva vektora u njima (x1,x2,z) i (x1,x2,y). Znamo da će to biti ista točka ako i samo ako z=y. Istu intuiciju možemo proširiti na neke točke (x1,x2,x3,z) i (x1,x2,x3,y): ovi će vektori biti isti ako i samo ako z=y. Primjer takvog svojstva može biti boja ili bilo što slično tome. Tako možemo zamisliti 4-dimenzionalni prostor kao 3D prostor u kojem dvije točke mogu biti na istom mjestu, ali ako su različite boje, onda nisu ista točka.

Predmeti reprezentirani kao searleovski grozd opisa sačinjenih od njihovih svojstava mogu se shvatiti kao n-dimenzionalni vektori, odnosno liste gdje je svaka komponenta numerička vrijednost svojstva koje je u tom stupcu. Premda ovo nije slučaj u čisto matematičkom shvaćanju vektora i matrica, proširenje do računalnog shvaćanja lista i matrica očito je. Čovjek koji ima visinu, težinu, širinu i boju očiju je entitet s četiri svojstva i taj entitet može se reprezentirati jedino u četvoro-dimenzionalnom prostoru. Čovjek ima puno

više od ovih četiriju mjera, pa će shodno tome trebati i višedimenzionalan prostor za pohranjivanje ovih svojstava. Neka svojstva reprezentirana su numeričkim vrijednostima, neka ordinalnim vrijednostima, a treće su pak da/ne bulovske vrijednosti. Ono što je vrlo zanimljivo jest da se u sličnom višedimenzionalnom prostoru mogu trivijalno reprezentirati riječi. Ovaj se pristup, koji je već pedesetak godina klasična metoda u umjetnoj inteligenciji, naziva *Bag of Words*. Zamislimo sličnu tablicu svojstava, gdje su sada svojstva riječi, a redci fragmenti duljeg teksta, npr. rečenice. Tada se ispod svake riječi upisuje koliko se puta ona javlja u danoj rečenici. Može se pratiti frekventnost, u kojem se slučaju ne radi o bulovskoj tablici ili jednostavno pojavnost, pri čemu svi redci imaju isključivo bulovske vrijednosti.

Ovo je reprezentacija riječi, ali ova reprezentacija ima mnoge probleme. Prije svega, ona ne vodi računa o tome pojavljuje li se neki redoslijed riječi ili ne. Moguće je proširiti ovaj jednostavan jezični model tako da se uvaži inherentna sekvencijalnost jezika da bi se uzimali parovi ili trijade riječi umjesto individualnih riječi. Premda bi ovaj pristup uključio određenu sekvencijalnost, on nam ne rješava drugi problem, a to je da slične riječi ili infleksije riječi imaju slično značenje. Treći, srodan problem, sastoji se u tome da nema načina za značenjsko povezivanje sintaktički različitih riječi, poput npr. »pravnik« i »odvjetnik«. Pravo je vrijeme da se postavi ključno pitanje za analogijsko zaključivanje: bismo li neljudskog agenta koji drži da su ovo slični pojmovi smatrali sposobnim za zaključivanje? Ili možda preciznije: ako bismo neljudskog agenta, bez ikakvog pozadinskog znanja, mogli naučiti dajući mu isključivo tekstualne podatke iz kojih je moguće tek implicitno iščitati da su pravnik i odvjetnik slični, bismo li priznali da posjeduje sposobnost zaključivanja? Tradicionalna simbolička logika ovo je u stanju napraviti isključivo kroz zaključivanje silogističkog tipa:

Svi odvjetnici su pravnici.
Većina pravnika su odvjetnici.

Dakle

Većinom su odvjetnici i pravnici iste osobe.

Premda ovo nije jedna od klasičnih figura silogizma, nego modifikacija silogističkog zaključivanja da bi se prihvatio jezični kvantifikator »većina«, neosporno je da se radi o punopravnom zaključivanju, što sugerira da je analogijsko zaključivanje također punopravno zaključivanje.

Analogijsko zaključivanje kao punopravno zaključivanje

Ideja ekstenzionalnosti kao kriterij zamjenjivosti u određenom broju konteksta je u filozofijski diskurs uveo W. V. Quine (1953) koji je govorio o koekstenzionalnosti pojmova *salva veritate*. Kako naučene analogije nisu podložne savršenom preklapanju, tako možemo govoriti o stupnjevima koekstenzionalnosti resolvenata analogije, čime uvodimo esencijalnu neizrazitost, odnosno neizrazitosti koja nije podložna probabilističkoj reinterpretaciji. Analogijsko zaključivanje može se promatrati kao punopravno zaključivanje jer ono djeluje ne samo kao prijedlog instacijacije nego i definira klasu ekvivalencije koja nudi kandidate koji su zamjenjivi *salva veritate*, odnosno klasu pojmova koja definira konkluziju. Ako se inzistira na simboličkoj definiciji zaključka, analogijsko se zaključivanje može prikazati kao:

t=s
t><t’

Dakle

t’=s

Pri čemu su t, t’ i s pojmovi, x=y je jednakost (»x je y«), a x><y je neka relacija analogije (koja je relacija ekvivalencije). Kada bi konkluzija bila t’><s, tada bismo mogli vrlo elegantno reći da se radi o klasično valjanom zaključivanju, no kod analogijskog zaključivanja riječ je o jednom (silogistički gledano) neprirodnom osnaženju na punu jednakost u konkluziji. Ovo je specifikum analogijskog zaključivanja koje ga čini različitim od pukog primjera klasičnog silogizma. Mogli bismo reći da je tome tako zato što analogijsko zaključivanje nije isključivo deduktivno, nego ima elemente generalizacije i osnaživanja tipičnih za indukciju. Ovime se već nazire primjerenost promatranja analogijskog zaključivanja s metodama strojnog učenja, koje generalno formaliziraju induktivno zaključivanje. Preostaje nam pokazati zbog čega smatramo da je potrebno od svih mogućih metoda koristiti umjetne neuralne mreže.

Ono što možemo zaključiti kao međurezultat sljedeće je: analogijsko zaključivanje (a) ide ispod razine pojma (i propozicije), i (b) predstavlja punopravno zaključivanje. Zanimljivo je usporediti tezu Douglasa Hofsdatera (Hofstadter i Sander 2013) o tome kako je analogija jezgra svakog mišljenja i zaključivanja jer bez pojmova nema mišljenja, a bez analogija nema pojmova.

Analogijsko zaključivanje je uz pomoć umjetnih neuralnih mreža moguće naučiti iz čistog teksta bez dodatnog znanja (Mikolov i dr. 2013), na način da se s naučenim reprezentacijama može zaključivati uz pomoć operacija inherentnih u vektorskom prostoru u kojemu one postoje. Naučene reprezentacije aktivacije su srednjeg sloja, u obliku matrice realnih brojeva. Svaki redak te matrice (redak matrice je vektor) odgovara jednoj riječi u tekstu, a ti se vektori mogu zbrajati (i oduzimati). Poznat je primjer iz (Mikolov i dr. 2013) gdje je iz utrenirane matrice izdvojen vektor za riječ »kralj«, oduzet mu je vektor za »muškarac« i dodan mu je vektor za »žena« i (gledajući euklidsku udaljenost) taj rezultat bio je jako blizu vektoru za »kraljica«.

U tradiciji simboličke umjetne inteligencije postoje na prvi pogled slični sustavi, poput intuicionističkih substrukturnih računa termova (Benton i dr. 1993), gdje termovi nisu više pravi atomi, nego imaju određenu strukturu, često opisanu pravilima njihove transformacije. No problem se javlja zato što su ovo sve ručno prilagođeni sustavi bazirani na znanju, koji nisu u stanju procesirati stvarne tekstove bez da ih se opskrbi dodatnim definicijama. U određenom smislu ovi sustavi postali su sami sebi svrha, bez obzira na specifičnost primjena zbog kojih su izmišljeni.

Slična tema u formalnoj epistemologiji bila je takozvana AGM teorija revizije vjerovanja (Alchourrón, Gärdenfors, Makinson 1985) koja definira mehanizme ekspanzije i kontrakcije vjerovanja u općim crtama, ali ne govori koje točno vjerovanje agent revidira u kojem slučaju. Čak i kasnije ekstenzije (Darwiche, Pearl 1997) nailaze na slične probleme. Osnovni problem ovakvih simboličkih pristupa u modeliranju kognitivnih procesa sastoji se u tome što nisu dinamični u smislu učenja i potrebno je ručno kodirati proširenja ili smanjenja vjerovanja da bismo kasnije imali jednostavni sustav operacija nad njima, odnosno da bismo mogli u tom prostoru jednostavnim klasičnim vektorskim operacijama zbrajanja i oduzimanja dobiti v(kralj)-v(muškarac)+v(žena)=v(kraljica), kako je opisano u (Mikolov i dr. 2013). U određenom smislu, ovime se postiže slično kao i kada se P&Q, T(P)=1, T(Q)=0 poistovjeti s 1*0.

Potrebno je naći prikladno preslikavanje pojmova na vektore (koje smo priručno označili s v(x)), a u (Mikolov i dr. 2013) se za to koristi obična umjetna neuralna mreža, što daje, s jedne strane, iznimne rezultate, ali s druge strane, ostavlja velike mogućnosti za proširenja.

Glavni problem dosadašnjih logičkih sustava zaključivanja temeljenih na pojmovima sastoji se u tome da takvi sustavi imaju lokalne reprezentacije pojmova, a za implementirati i eventualno nadograditi pristup (Mikolov i dr. 2013) ili neki dugi subsimbolički model do punog subsimboličkog zaključivanja, potrebno je imati distribuirane reprezentacije pojmova, odnosno vektore realnih brojeva bez komponenata jednakih nuli.

Distribuirane reprezentacije

Drugačiji tip reprezentacije distribuirana je reprezentacija kakva je prisutna u umjetnim neuralnim mrežama, kao i u biološkim neuronima. Prema današnjem razumijevanju u kognitivnoj znanosti, pojam »crvena lopta« ne postoji u jednom neuronu (bilo biološkom bilo umjetnom), nego u puno njih istovremeno gdje postoji kao istovremena pobuđenost tih neurona. Nadalje, pojam »zelena lopta« ne mora postojati u istim neuronima na »zelen, a ne crven način«, nego može postojati u drugim neuronima kao binarna vrijednost. To znači da neuroni ne moraju biti nosioci svojstava, što ih oslobađa od problema da imaju različite interne reprezentacije za »crveno« i »zeleno«.

Zamislimo da nije tako, i da u neuronima *svi* pojmovi imaju lokalnu reprezentaciju. To bi značilo da crvena lopta dobije reprezentaciju u neuronima X, Y, Z. Imamo dvije mogućnosti. Ti neuroni mogu biti ili svaki zadužen za jednu komponentu propozicije (X=crveno, Y=je, Z=lopta), pa je onda propozicija »cinober lopta« ista kao crvena lopta, ali s nekim X2 za cinober. Uzmimo bez smanjenja općenitosti da je neimenovana »crvena« svjetlija od »cinober« nijanse. To znači da je pojam »svjetlije« u stvari distribuiran između X, X2, X3 itd., što znači da nemaju svi pojmovi lokalnu reprezentaciju jer »svjetlije« očito ima distribuiranu reprezentaciju.

Ako uzmemo da neuron može imati samo dva stanja (poput električne sklopke), onda distribuirane reprezentacije tvore skup svih mogućih podskupova dostupnih neurona. Zanimljivo je razmotriti mogućnosti takova uređaja kada bismo mogli imati prebrojivo beskonačno neurona (bilo bioloških, bilo umjetnih), no ovo nećemo istraživati u ovom radu.

Distribuirane reprezentacije u kognitivnim znanostima

Kao metodološki temelj neuralnih jezičnih modela uglavnom se navodi poznata krilatica britanskog jezikoslovca Johna R. Firtha:

»Znat ćeš riječ prema društvu s kojim je.«[1] (Firth 1957)

Kao i drugim brojnim tezama u lingvistici i ovoj možemo pronaći preteče u logici i filozofiji jezika. Važnost konteksta kao preduvjeta razumijevanja značenja najpoznatije je formulirao Frege u svom metodološkom načelu konteksta:

1
»You shall know a word by the company it keeps.«

»... nikada (...) ne pitaj za izolirano značenje riječi, nego samo u kontekstu propozicija.« (Frege 1884)

Cjelokupnu paradigmu značenja kao uporabe kasnog Wittgensteina možemo smatrati stvarnom metodološkom prethodnicom ovom pristupu, posebno uzevši u obzir utjecaj Wittgensteinovih ideja na britansku lingvistiku.

Naš argument ne pretpostavlja da pojmova ima beskonačno. Premda može djelovati očito da ne može postojati bijekcija između pojmova i neurona, zbog toga što prijašnjih ima konačno, a potonjih beskonačno. Postojanje beskonačno mogućih pojmova dvojbena je pretpostavka. Dva su oblika na koji beskonačnost može ući u jezično-pojmovni svijet. Prvo, jezik se može proširivati novim izrazima, poput »Labradoodle« (za križanca pudle i labradora). Svako takvo proširenje događa se u vremenu, i čak ako se uzme mogućnost simultane definicije N novih pojmova, da bi skup svih pojmova bio beskonačan, nužno je da u nekom trenutku T simultano uvedemo beskonačno mnogo pojmova ili da imamo beskonačno trenutaka u kojima uvodimo novi pojam. Drugo, većina jezika posjeduje (prefiksalne) generatore, poput »pra« u »pradjed«, čime je moguće tvoriti izraze poput »prapraprapradjed«. Imaju dva razloga zašto beskonačni nizovi ovog tipa nisu dio jezika. Prvo, nakon određenog broja predaka, predaka više nema. Koliko god se ovo proširivalo, načelno svi imaju konačan broj predaka, pa izraz s više »pra« od broja predaka jednostavno ne referira. Drugo, i možda suptilnije, je da ljudi kognitivno reduciraju broj pojmova koje koriste. Tako, na primjer, kada netko kaže »pradjed«, »prapradjed«, »prapraprapradjed«, »praprapraprapradjed« itd. sugovornik može to interpretirati kao »Ivan«, »Stjepan«, »Daljnji predak« itd. Bez obzira koliko daleko predci dobivaju imena, u jednom trenutku kognitivno im se pridaje jednostavno pojam »Daljnji predak«.

Suvremenom operacionalizacijom ovog Wittgensteinova pristupa problemu značenja možemo smatrati distribucijsku semantiku unutar jezikoslovlja i kognitivne znanosti (McDonald i dr. 2001). Osnovna je hipoteza ovoga pristupa da riječi sličnog značenja imaju sličnu distribuciju. Praktična posljedica toga je da značenje riječi možemo naučiti iz konteksta tih riječi u uporabi, odnosno u tekstu; riječi koje se dovoljno puta pojavljuju u istim, odnosno sličnim kontekstima, semantički su bliske.

Recentna empirijska verifikacija ove ideje kroz algoritam Word2vec (Bengio i dr. 2003, Mikolov i dr. 2013), koji je vrsta umjetne neuralne mreže, metodologija subsimboličkog višedimenzionalnog prikaza značenja riječi, pokazala se izuzetno uspješnom te potiče na brojna pitanja poput koji je značaj ovih otkrića za tradicionalne filozofske i logičke probleme poput problema značenja, odnosa denotacije i konotacije, neizrazito odnosno fuzzy zaključivanje i slične. Smatramo da njihov rezultat nije pretjerano usporediti s drugim znanstvenim otkrićima koja su redefinirala stare filozofske probleme i otvorila nove.

Osnovna ideja suvremenog višedimenzionalnog prikaza značenja riječi iznenađujuće je jednostavna – prikazujemo ih višedimenzionalnim vektorima koji su dobiveni učenjem koje daje veću vjerojatnost da se neka riječ pojavi uz neku drugu ako se one u podatcima često pojavljuju zajedno. Ovime se može također vidjeti koje se riječi često pojavljuju uz iste riječi odnosno u sličnim rečenicama (npr. »flaša« i »boca«). Konkretno, ako riječ kojoj želimo otkriti značenje označimo s r_t, a npr. prethodnu riječ s r_{t-1}, želimo takav višedimenzionalni prikaz gdje će P(rt-1|rt), kao i druge riječi iz okoline, biti najveća. Taj postupak naziva se *Skipgram*, dok se analogni postupak koji maksimalizira vjerojatnost riječi na temelju okoline P(rt|rt-1,rt+1, ...) naziva *CBOW*. *Skipgram* i *CBOW* su implementacijske varijante Word2vec i predstavljaju

naučene neuralne jezične modele. Primjenom ova dva jednostavna postupka na velikoj količini uporabe jezika, ostvareni su rezultati koji često nadmašuju prosudbu ljudskog zaključivanja.

Budući da neuralni jezični modeli uče distribuirane reprezentacije pojmova iz čistog teksta, nemaju dodatnog ljudskog znanja. To se učenje provodi operacijama koje su dobro definirane nad određenim vektorskim prostorom i time predstavljaju interne mehanizme tog prostora. Premda o stvarnoj realizaciji komputacijskog prostora ljudskog agenta znamo vrlo malo, znamo da su ljudi u stanju naučiti razmišljati koristeći analogije jer je vrlo teško zamislivo da je analogijsko zaključivanje u bilo kojem vidu apriorno zaključivanje. Jednostavnost neuralnih jezičnih modela moguće je koristiti za ekstrapolaciju jednostavnosti ljudskih kognitivnih procesa potrebnih za analogijsko zaključivanje. Ovo možda zvuči novo, no zasniva se na činjenici da su prethodne teorije koje su poistovjećivale kognitivne procese sa simboličkim, a ne vektorskim operacijama, morale prihvatiti komputacijska ograničenja tih tehnika i stipulirati veću moć ljudskog uma s obzirom na računalo. Mi ovdje uzimamo kao prešutnu pretpostavku suprotan smjer: ako je na računalu ideju moguće jednostavno realizirati, vjerojatno se u ljudskom umu, gdje se neosporno taj proces i realizira, on realizira na podjednako jednostavan način.

Silogizam, klasična simbolička logika i analogijsko zaključivanje

Subsimbolička logika analogijskog zaključivanja, kao naučenog zaključivanje, otvara i pojam proksimiteta. Što su pojmovi u višedimenzionalnom prostoru bliži, to im je proksimitet veći, ali time je informativnost zaključka niža. Ovo je srodno Blackburnovu i Bosovu (2005) poimanju informativnosti u simboličkim sustavima, ali tamo je informativnost binarno svojstvo, ako je sustav valjan, onda nije informativan i obrnuto. Subsimbolička logika tretira informativnost kao proksimitet, što znači da je razina informativnosti opisana realnim intervalom između 0 i 1. Kao primjer uzmimo dva klasična silogistička zaključka:

Svi ljudi su sisavci.
Svi sisavci su živa bića.

Dakle

Svi ljudi su živa bića.

Zaključak ovog tipa je poprilično neinformativan zaključak gledano iz perspektive opće populacije jer bi teško bilo izbjeći naučiti ove dvije premise, a i konkluzija, premda slijedi iz premisa, ne donosi ništa novog u epistemičkom smislu. Vjerojatno je agent već znao ovu konkluziju kao nezavisnu činjenicu. S druge strane, zaključak

Svi dvonožni dinosauri mesojedi iz Krede su pernati.
Sve pernate životinje su ptice.

Dakle

Svi dvonozni dinosauri mesojedi iz Krede su ptice.

je u mnogočemu vrlo informativan zaključak za opću populaciju (i sama konkluzija, čak uz uvjet da znamo premise može nam biti informativna). Općeni-

to, cijela simbolička logika i matematika kao sustavi teorema primjer su deduktivnog znanja koje je iznimno informativno. Na primjer, dokaz centralnog graničnog teorema nije trivijalan za opću akademsku populaciju, premda svi znaju i intuitivno prihvaćaju osnovne aksiome vjerojatnosti, kao i posljedice navedenog teorema. Replikacija ovog zaključivanja u simboličko-logičkom sustavu je vrlo jednostavna:

Za svaki x, ako Sx onda Mx.
Za svaki x, ako Mx onda Px.

Dakle

Za svaki x, ako Sx onda Px.

Problem u simboličkom pristupu je da simboli ne nose sa sobom ništa o informativnosti zaključka (definicijom se poistovjećuju s konkretnim pojmovima), a bez dodatne informativnosti nije moguće prenijeti karakteristike zaključivanja kao kognitivnog procesa, ali ni provesti bilo kakvo zaključivanje na razini pojmova poput analogijskog zaključivanja.

Zaključak

Dva se pitanja sama po sebi nameću. Ako je analogijsko zaključivanje zaista zaključivanje, i ako je njega moguće naučiti iz čistog teksta, je li tada moguće naučiti zaključivati iz teksta? Ako da, logičko je zaključivanje posljedica, a ne preduvjet razumijevanja teksta, što je interesantno i s hermeneutičke strane, a procesiranje teksta s metodama poput neuralnih jezičnih modela može se smatrati prototipom ovog novog pristupa. Ovdje nije potrebno tvrditi da se radi o punom razumijevanju jer čak i na ovoj razini naš pristup implicira reviziju danas općeprihvaćene teze da je zaključivanje preduvjet za razumijevanje, a time radikalno redefinira odnos tih dvaju pojmova.

Druga zanimljiva posljedica je da je analogijsko zaključivanje, kakvo je prisutno u neuralnim jezičnim modelima, u stvari naučeno zaključivanje. To znači da je analogijsko zaključivanje u stvari izvlačenje obrazaca koji su prisutni u napisanom tekstu. Što bi bilo ako bi se isti pristup mogao proširiti na opće zaključivanje? Ono što bi bilo sigurno jest da bi taj proces preslikao *input* u *output* preko distribuiranih mentalnih reprezentacija. Možemo zamisliti da se neki ispravan zaključak sastoji od 100 aktiviranih neurona. To otvara put fenomenima nalik *sorites* paradoksu, ali kod mentalnih procesa. Pitanje koje se može postaviti bilo bi sljedeće: bi li *sorites* proces doveo od ispravnog do neispravnog zaključka ili bi jednostavno zapriječio put prema bilo kakvoj konkluziji? Ovo je vrlo zanimljivo otvoreno pitanje jer klasična simbolička logika nije u stanju analizirati gradualni prijelaz valjanog zaključka u nevaljani ili u izostanak istog, barem ne na dinamičan način na koji to mogu neuralne mreže i predloženi program subsimboličke logike.

U ovom smo radu pokazali da se rudimentarni oblici zaključivanja mogu naučiti, prikazujući novije algoritme iz umjetne inteligencije koje su u stanju naučiti sličnost pojmova, što je osnova za analogijsko zaključivanje kao rudimentarno zaključivanje slično jednostavnom silogizmu. U algoritmu Word2vec, temeljenom na umjetnoj neuralnoj mreži na koji smo se referencirali, analogijsko je zaključivanje dobiveno srednjim slojem neuralne mreže i tako dobiveni vektori za pojedine riječi takvi su da su vektori riječi sličnih značenja bliži, nego onih različitih. Ovo je veliki pomak jer se možemo oslo-

niti na značenjsku sličnost koja je nositelj analogije. Naš doprinos u ovom radu sastoji se u istraživanju mogućnosti primjene Word2vec reprezentacija kao osnove za subsimboličko zaključivanje, u nadi da ćemo u budućnosti moći adaptirati umjetne neuralne mreže za složenije zaključivanje. Ovo je jaki argument za empiričnost i desktiptivnost logike koja vodi do otvaranja pitanja jesu li logika i zaključivanje naučene kognitivne vještine, a ne (kako se često pretpostavlja) čovjeku urođene kao uvjet mogućnosti spoznaje i razumijevanja.

Literatura

Alchourrón, C. E.; Gärdenfors, P.; Makinson, D. (1985). »On the logic of theory change: Partial meet contraction and revision functions«. *Journal of Symbolic Logic* 50 (1985) 2, str. 510–530. doi: https://doi.org/10.2307/2274239.

Bengio, Y. i dr. (2003). »A Neural Probabilistic Language Model«. *Journal of Machine Learning Research* 3 (2003), str. 1137–1155.

Benton, N. i dr. (1993). »A term calculus for Intuitionistic Linear Logic«. U: Bezem, M.; Groote, J. F. (ur.). *Typed Lambda Calculi and Applications: International Conference on Typed Lambda Calculi and Applications TLCA '93*, str. 75–90. Berlin: Springer. doi: https://doi.org/10.1007/bfb0037099.

Boole, G. ([1854] 2010). *An Investigation of the Laws of Thought*. London: Watchmaker Publishing.

Blackburn, P.; Bos, J. (2005). *Representation and Inference for Natural Language*. Amsterdam: CSLI.

Boden, M. A. (1988). *Computer Models of Mind: Computational Approaches in Theoretical Psychology*. Cambridge: Cambridge University Press.

Darwiche, A.; Pearl, J. (1997). »On the Logic of Iterated Belief Revision«. *Artificial Intelligence* 89 (1997) 1–2, str. 1–29. doi: https://doi.org/10.1016/s0004-3702(96)00038-0.

Dummett, M. (1978). »Is Logic Empirical?«. U: *Truth and Other Enigmas*. Cambridge: Harvard University Press.

Einstein, A. (1905). »Zur Elektrodynamik bewegter Körper«. *Annalen der Physik* 17 (1095) 10, str. 891–921. doi: https://doi.org/10.1002/andp.19053221004.

Erdmann, B. (1892). *Logische Elementarlehre*. Halle: Niemeyer.

Firth, J. R. (1957). »A Synopsis of Linguistic Theory 1930–1955«. U: *Studies in Linguistic Analysis*, str. 1–32. London: Philological Society.

Heymans, G. (1911). *Das künftige Jahrhundert der Psychologie*. Leipzig: Barth.

Hofstadter, D.; Sander, E. (2013). *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*. New York: Basic Books.

Leibniz, G. W. ([1686] 1992). *Discourse on Metaphysics and the Monadology.* London: Prometheus Books.

Lipps, T. (1893). *Grundzüge der Logik*. Leipzig: Verlag von Leopold Voss.

McDonald, S.; Ramscar, M. (2001). »Testing the Distributional Hypothesis: The Influence of Context on Judgements of Semantic Similarity«. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, str. 611–616.

Mikolov, T. i dr. (2013). »Efficient Estimation of Word Representations in Vector Space«. *ICLR Workshop*, arXiv:1301.3781.

Mill, J. S. ([1843] 2002). *A System of Logic: Ratiocinative and Inductive*. Honolulu: University Press of the Pacific.

Putnam, H. (1969). »Is Logic Empirical?«. U: Cohen, R. S.; Wartofsky, M. W. (ur.): *Boston Studies in the Philosophy of Science: Proceedings of the Boston Colloquium for the Philosophy of Science 1966/1968*, str. 216–241. Dordrecht: D. Reidel Publishing Company.

Quine, W. V. (1953). »On What There Is«. U: *From a Logical Point of View*. Cambridge: Harvard University Press.

Sigwart, C. (1904). *Logik*. Tübingen: Mohr.

Wittgenstein, L. (1953). *Philosophical Investigations*. London: MacMillan Publishing Company.

Zadeh, L. (1996). »Fuzzy logic = computing with words«. *IEEE Transactions on Fuzzy Systems* 4 (1996) 2, str. 103–111.

Sandro Skansi, Davor Lauc

Analogical Reasoning and Word-Meanings in a Multidimensional Space

Abstract

The present work explores the underlying thought behind symbolic logic which accepts concepts as atomic components, and we introduce a different formalism based on artificial neural networks for the formalization of logical reasoning as a cognitive process, which defines an approach we call subsymbolic logic. We apply this approach to analogical reasoning, which we argue is the proper reasoning. We also explore the cognitive aspects of this approach, especially in isolating and reproducing spontaneous but erroneous forms of reasoning (cognitive biases) which are a part of logical reasoning viewed as a cognitive process. Today, it is the dominant technique in artificial intelligence, but the philosophical aspects of such an approach remain mostly unexplored. To the best of our knowledge, this is the first such attempt at using artificial neural networks to analyse analogical reasoning.

Key words

analogical reasoning, meaning of words, artificial neural networks, neural language models, cognitive connectionism, subsymbolic logic