Hangalapú fordítás a Google mobilos alkalmazásával. *Bárhol, bármikor*

GÉPI FORDÍTÁS | KORPUSZFÜGGÉS | MÁSODRENDŰ NYELVEK

Fordítottan arányos

Mennyivel lettek jobbak a fordítóprogramok a mesterséges intelligencia ugrásszerű fejlődése nyomán? A változások valóban forradalmiak, de csak néhány nagy nyelv profitál belőlük, sok ezer kicsit az eltűnés fenyeget. A magyar ebből a szempontból a nagyok közé tartozik.

„A medve lejött a hegyről. Ott volt egy út, és átment rajta." Vagy: „A medve lejött a hegyről. Ott volt egy úthenger, és átment rajta." Az efféle, egymástól alig különböző mondatokkal a régebbi gépi tanulású nyelvi modellekre épülő fordítóprogramok nemigen tudtak mit kezdeni (pontosan ki ment át kin vagy min?). A legújabb verzióknak viszont már nem okoz gondot, hogy érzékeljék, miként „öröklődik" az előző mondatból az egyes szám harmadik személy, és kire vonatkozik a névmás – hívta fel a figyelmet néhány hete a Magyar Tudományos Akadémia közgyűlésén tartott előadásában Prószéky Gábor matematikus-nyelvész, a Nyelvtudományi Kutatóközpont főigazgatója.

A nagy nyelvi modellekre (angol terminológiával: Large Language Model, LLM) épülő fordítóprogramoknak nem csak ez a képességük újdonság. Szakértők szerint az utóbbi 2-3 évben valóban forradalmi változások zajlottak le a gépi nyelvértés és fordítás területén. A különbség különösen a korábbi, nagyjából 70 éve tartó próbálkozásokkal összevetve válik nyilvánvalóvá.

Eleinte, évtizedeken át, a nyelvészek megpróbálták a számítógép számára értelmezhető formában betáplálni a nyelvtani szabályokat, de kiderült, hogy ez bizonyos szint felett áttekinthetetlenné válik, ráadásul kevéssé életszerű eredményekre vezet. Különösen a többértelműséggel nem tudtak mit kezdeni a szabályalapú fordítógépek. Az 1990-es évektől kezdtek statisztikai nyelvészeti módszerekre áttérni, és adtak nagy mennyiségű, emberek által fordított szöveghalmazt a kompütereknek, hogy azt használják fordításhoz, ami a gyakorisága alapján valószínűbb. Ezzel már könnyebb volt felismerni szókapcsolatokat vagy mondatrészeket, de a szövegkörnyezetet ez a rendszer sem tudta figyelembe venni, a nagyobb szövegadatbázisokból ellentmondásos eredmények jöttek ki, továbbá nem tudott mit kezdeni a nagyon ritka szószerkezetekkel és a nagyon hosszú mondatokkal – magyarázta Prószéky Gábor.

A mai áttörés alapját a mesterséges neurális háló alapú gépi tanulás teremtette meg, ami már nem sza-

Illényi Balázs
b.illenyi@hvg.hu

vakat, kifejezéseket és mondatokat akar egymásnak megfeleltetni, hanem hasonlóság alapján fordít. A nagy ötlet 2013-ban egy cseh PhD-hallgató programozó, Tomás Mikolov fejéből pattant ki. Ő találta ki, hogy a digitális térben az egyes szavakat az összes előfordulásukból (vagyis a szövegkörnyezetükből) kiszámított komplex vektorokkal jellemezzék, ezáltal a szemantikailag hasonló, azaz hasonló jelentésű nyelvi egységek fognak egymáshoz közel kerülni – foglalja össze kérdésünkre Váradi Tamás, a Nyelvtudományi Kutatóközpont főigazgató-helyettese.

A Mikolov-féle első modell minden egyes szót 300 dimenzió mentén ábrázolt, vagyis a vektorok az egyes szavaknak 300 jellemzőjét ragadták meg, ennek köszönhetően pedig az egymáshoz való viszonyuk matematikai pontossággal meghatározható lett. Így már nem fordulhat elő, ami egy hagyományos szótárban, hogy egymás mellé kerüljenek olyan, hasonló írásképű, ám jelentésükben távol álló szavak, mint például „tör” és „tőr”.

A nyelvfeldolgozásnak óriási lökést adott, hogy a mennyiség több szinten is úgymond átcsapott minőségbe. Egyrészt az említett vektortérben nagyságrendekkel pontosabban lehet meghatározni a szavak közti viszonyokat azáltal, hogy a szuperszámítógépek újabban már ésszel nemigen felfogható, közel 13 ezer dimenziót képesek kezelni, ráadásul emberi visszacsatolással ezeket még tudják finoman hangolni. Másrészt – ugyancsak a számítási kapacitás növekedése miatt – jóval nagyobbra lehet nyitni a „kontextusablakot”. Ez azt jelenti, hogy nem csupán szavak és mondatok, hanem bekezdések vagy akár teljes művek (több tízezer szó) belső viszonyrendszerét képes „fejben tartani” a gép, így a szavak közti összefüggéseket korábban elképzelhetetlen részletességgel tudja értelmezni – magyarázza Váradi Tamás.

Fordítási szempontból az igazi érdekesség, hogy a hatalmas szövegadatbázisokon (szaknyelven: korpuszokon) betanított nagy nyelvi modellek vektorterei a különböző nyelveken nagyon hasonlóak: az azonos jelentésű szavak, mondatok hasonló matematikai értékeket vesznek fel. Kiderült, hogy ez a nyelvfüggetlen vektortér valamiféle interlinguaként (közvetítő nyelvként) működik, amely képes absztrakt módon megragadni a jelentést, és ez már nem klasszikus értelemben vett fordítást tesz lehetővé, hanem – fogalmaz a nyelvtechnológus kutató – egyfajta „nyelvek közötti transzfert”.

A gépi fordításban is nagy áttörésnek számított, hogy az utóbbi években nemcsak a mesterséges neuronhálózatok méretét növelték több nagyságrenddel, hanem hasonló mértékben bővítették a nagy nyelvi modellek betanítására használt korpuszokat: akár több száz milliárd szóból álló, emberek által alkotott összefüggő szövegeket tápláltak be. Ennek, és a korábban említett fejlesztéseknek, valamint az emberi visszacsatolásos tanításnak köszönhetően hirtelen elkezdett sokkal jobban működni a nyelvértési és fordítási funkció is. Ráadásul a generatív modellek – ahogy azt a ChatGPT-vel és társaival próbálkozó felhasználók tapasztalják – egyszerre több feladatra is képesek: immár nem okoz nekik gondot kérdések megválaszolása, hosszabb szövegek alkotása, összefoglalása, átfogalmazása vagy értelmes párbeszédek folytatása sem, és mellettük nincs szükség külön gépi fordításra betanított rendszerre sem.

A látványos fejlődés árnyoldala, hogy a rendszer igen erőteljesen korpuszfüggő. Ez azt jelenti, hogy csak azzal a nyelvvel tud kezdeni valamit a fordítóprogram, amelyikből elegendő mennyiségű szöveget „látott”. A világon nagyjából 7 ezer nyelvet beszélnek ugyan, ám a nagy nyelvi modellek többségét angol vagy kínai korpuszokon tréningezik, és nem véletlen, hogy a generatív mesterséges intelligencia angolul működik a legjobban, hiszen a betanításhoz sokszor használt interneten a weboldalak több mint fele ezt a nyelvet használja – emlékeztet a The Atlantic áprilisi cikke.

A ChatGPT-t azért is könnyű még magyarul zavarba hozni, mert míg az alapmodellt 183 milliárd angol szóból álló korpuszon fejlesztették, magyarul csak 128 millió szavas szövegadatbázison gyakorolhatott. Igaz, a több nagyságrendi eltérés dacára is a magyar a 19. legnagyobb nyelvnek számít a csevegőmodellben.

Kutatók világszerte versenyt futnak az idővel, hogy a forrásszegény nyelvek ne tűnjenek el az angol dominanciája miatt. Egy nemzetközi csapat – hoz példát a The Atlantic – 517 afrikai nyelvnek épített hosszú évek munkájával egy 42 gigabájtos gyakorló adatállományt, ami a maga nemében páratlan ugyan, de alig látható töredéke mondjuk az OpenAI adatbázisának. Különösen a nem latin betűs írásmódú vagy az eleve gyenge írásbeliségű nyelvek vannak veszélyben,

A romani és a többiek

Több mint 614 millió legyet ütnek egy csapásra azzal, hogy bővítik a Google Fordító repertoárját. Ennyien – a világ népességének mintegy 8 százaléka – beszélik ugyanis azt a 110 új nyelvet, amelyeket a múlt csütörtökön bejelentett expanzió után használni lehet a világcég fordítóprogramjában. A mesterséges intelligencia, egészen pontosan a PaLM 2 nagy nyelvi modell segítségével ezekben a napokban élesedő változás a Google Fordító eddigi legnagyobb bővítési köre.

Nem meglepő, hogy megannyi kihalás szélén álló, kevesek által beszélt nyelvre csak most, a sok (száz)millió felhasználó által érdekesnek talált nyelvek után volt kapacitás. De akadnak a mostani újak között olyanok is, amelyek óriási táborral rendelkeznek, érthetetlen módon eddig mégis kimaradtak a világ legnagyobb fordítószolgáltatásából. Ilyen például a több mint 80 millió fő által használt kantoni. Ebben az esetben a Google azzal magyarázza a gépesítés elhúzódását, hogy a kantoni írott formája túlságosan hasonló az angol után a világ második legnagyobb nyelvének számító mandarinéhoz, ezért nagyon nehéz volt elegendő mennyiségű szöveget szerezni a mesterséges nyelvi modell betanításához.

Magyarországról nézve érdekes még a legszűkebb értelemben véve is a cigány nyelvet jelentő romani bekerülése, amelyet külön is kiemel tájékoztatójában a Google. Mint írják, a roma nyelvnek annyiféle dialektusa létezik, hogy lehetetlen küldetésnek tűnt igazságosan kiválasztani közülük egyetlent. A végül a Google Fordítóba került verzió a „vlax romani”, ami az oláhcigányok nyelvéhez áll a legközelebb, de más nyelvjárásokból is használ elemeket.

Fordítás kameraképről valós időben. *Látra szóló*

PROFIMEDIA

AFP / THOMAS SAMSON

Magyarországon például a cigány nyelv kerülhet jelentős hátrányba.

Ráadásul – panaszkodnak a kutatók – a kis nyelveket még az ág is húzza: a neten hozzáférhető szövegek igen rossz minőségűek, Afrikában például sokszor csak angolból igénytelenül fordított missziós honlapokat vagy kattintásvadász hirdetéseket találni egy-egy adott nyelven. Az ilyen korpuszon betanított fordítóprogramoktól nem érdemes sokat várni, hiszen aligha képesek figyelembe venni a kontextust, a többértelműséget vagy a kulturális jellegzetességeket. Mint Holy Lovenia szingapúri MI-kutató mondja, többet tudnak a hamburgerekről és a Big Benről, mint az adott ország gasztronómiájáról és helyi nevezetességeiről. Ezért is valószínű, hogy hamarosan minden eddiginél nagyobb szükség lesz fordítókra, hogy segítsenek a generatív MI által készített idegen nyelvi tartalmak értékelésében, felülvizsgálatában és minőség-ellenőrzésében – vélik többen is.

Szakértők attól tartanak, hogy a generatív nyelvi modellek egyenlőtlen fejlődése miatt világszerte sokan válnak másodrendű polgárokká, amennyiben nem fogják tudni az MI-t olyan sokrétűen – akár tanulásra, adóbevallás kitöltésére, munkahelyi feladatok végzésére vagy hivatalos ügyek intézésére – használni, mint a világnyelvekbe született emberek.

A magyarok ebből a szempontból szerencsések. Mint Prószéky Gábor az előadásában elmondta, bár a ChatGPT tréningezésére csak 128 millió szavas szövegadatbázist használtak, a Nyelvtudományi Kutatóközpont ennél sokkal nagyobb, immár 50 milliárd szónyi saját magyar korpusszal betanított nyelvmodelleket állított elő. A PULI nyelvmodellcsalád tagjaihoz hasonló alapmodellek az angolon kívül a világon legfeljebb tucatnyi nyelvre készültek eddig. Ennek köszönhetően bizonyos feladatok végzése során a hazai kutatók nem lesznek kiszolgáltatva világcégek webes szolgáltatásainak.

Ami pedig irodalmi szövegek fordítását illeti, jóllehet ez még a mai fejlettségi szinten nem igazán realitás, az elmúlt évek robbanásszerű fejlődését tekintve Váradi Tamás „nem merne mérget venni” arra, hogy a modellek pár év múlva sem lesznek erre képesek. ■