

A CHATGPT TÖRTÉNETE¹

THE HISTORY OF CHATGPT

Héja Enikő

PhD, tudományos munkatárs

HUN-REN Nyelvtudományi Kutatóközpont Nyelvtechnológiai kutatócsoport, Budapest

heja.eniko@nytud.hun-ren.hu

ÖSSZEFOGLALÁS

A természetesnyelv-feldolgozás (NLP) célja a természetes nyelvek számítógépes eszközökkel történő elemzése és megértése. Bár a tudományterület egészen az 1940-es évekre nyúlik vissza, a 2010-es évekig nem született olyan egységes paradigma, amely a legfontosabb problémákra robusztus megoldást tudott volna nyújtani. Ezen hiányosság legfontosabb okai közé tartozik a természetes nyelvek változatossága, illetve a feladatok széles köre. Az utóbbi tíz évben azonban a mély neurális hálók és a viszonylag könnyen elérhető, soha nem látott mennyiségű digitális szöveg együttesen olyan áttöréseket hoztak a tudományterületen belül, amelyek sok esetben még a kutatókat is meglepték. Jelen cikk az utóbbi évek egyik legnagyobb hatású fejlesztésével, a ChatGPT-vel foglalkozik. Célunk, hogy bemutassuk a ChatGPT létrejöttéhez vezető legfontosabb lépéseket, amelyek mindegyike önmagában is paradigmaváltásnak tekinthető az NLP területén belül.

ABSTRACT

Natural language processing (NLP) is the analysis and processing of natural languages by computational means. Although the field dates back to the 1940s, until the 2010s no single paradigm had emerged that could provide a robust solution to the most important problems. The main reasons for this lack are the diversity of natural languages and the wide range of tasks. In the last decade, however, the combination of deep neural networks and the relatively easy access to unprecedented amounts of digital text has led to breakthroughs in the field, many of which have surprised even researchers. This article focuses on one of the most influential developments in recent years, ChatGPT. Our goal is to describe the key steps that led to the creation of ChatGPT, each of which can be considered a paradigm shift within the field of NLP.

¹ Jelen cikk az *Érintő* online folyóirat 2023. szeptemberi számában megjelent közlemény módosított változata.

Kulcsszavak: disztribúciós szemantika, szavak vektoros ábrázolása, előtanított nyelvmodellek, generatív nyelvmodellek, finomhangolás, prompt programozás, megerősítéses tanulás emberi visszajelzés alapján (RLFH)

Keywords: distributional semantics, vector representation of words, pretrained language models, generative language models, fine-tuning, prompt programming, reinforcement learning from human feedback (RLFH)

1. BEVEZETÉS

Az utóbbi időszak legnagyobb port kavarázó nyelvközpontú mesterséges intelligencia (MI) fejlesztése a ChatGPT, amellyel több nyelven is elvégezhetünk különböző feladatokat, legyen az forgatókönyvírás vagy akár bibliográfiakészítés tudományos írásokhoz. Fontosságának megfelelően a ChatGPT-ről rengeteget írtak, néhol lekicsinylő, méltatlankodó hangnemben. Kritizálták többek között az érettségin elért eredményeit, de olvashattunk arról is, hogy „pofátlanul hazudik és blöfföl” (URL1).

A jelen cikkben azt kívánjuk bemutatni, hogy bár a ChatGPT működése korántsem hibátlan, miért számít mégis áttörésnek a *természetesnyelv-feldolgozás* (továbbiakban NLP – *natural language processing*) területén. Ehhez először röviden tekintsük át, hogy mivel foglalkozik az NLP. Manapság az NLP az élet minden területét áthatja: a leggyakrabban használt eszközök talán az internetes keresők, illetve a fordítóprogramok. Számos olyan NLP-alkalmazás is létezik, amelyet közvetlenül nem használunk, de különféle eszközökbe beépítve már találkozhattunk velük. Ilyenek a *névelem-felismerő rendszerek*, amelyek nemcsak a tulajdonneveket, de ezek kategóriáit is képesek beazonosítani a szövegekben, vagy a *szentimentelemzők*, amelyek internetes véleményekről döntenek el automatikusan, hogy azok pozitívak vagy negatívak, esetleg semlegesek-e.

Összességében elmondhatjuk, hogy a legutóbbi időkig a hagyományos NLP eszközök *nyelv- és feladatspecifikusak* voltak, vagyis ezeket mindig egy meghatározott nyelvre és egy meghatározott feladatra fejlesztették ki. Így például, készült névelem-felismerő az angolra vagy szentimentelemző a magyarra. A nagy mennyiségű szövegen tanított mély neurális hálók, vagyis a *nagy nyelvi modellek* (továbbiakban LLM – *large language model*) megjelenésével ez a helyzet azonban drasztikusan megváltozott. Egy hasonlattal élve, az LLM-ek olyanok, mint a svájci bicska: a természetes nyelvek feldolgozása során többé már nincs szükségünk külön késre, villára vagy ollóra, ehelyett egy sokkal általánosabb eszköz áll a kutatók rendelkezésére, amely kis erőfeszítéssel tovább alakítható a konkrét eszközök bármelyikévé. Azonban bármennyire is áttörésnek számított, a ChatGPT színrelépésével a „svájcbicska-paradigma” rövid idő után szintén meg-

haladottá vált: egy előre definiált eszköztár helyett egy nagyon mély tudással és mostanáig feltáratlan képességekkel rendelkező mesterséges intelligencia jött létre. A hasonlatunk keretei között maradva, a ChatGPT tekinthető egy univerzális konyhai robotgépnek, amely a hagyományos feladatok mellett számtalan meglepő új dologra is képes, olyannyira, hogy új kutatási terület vizsgálja, hogy pontosan milyen feladatok elvégzésére is alkalmas. A ChatGPT további érdekes tulajdonsága, hogy a feladatokat számos nyelven képes elvégezni, így végeredményben a ChatGPT-vel egy univerzális, poliglott MI született.

A cikk hátralevő részében áttekintjük, hogy milyen főbb lépések vezettek a ChatGPT megjelenéséhez. Bár meglepő módon – és kissé leszűkítve – a ChatGPT története mindössze tíz évre tekint vissza, fontos szem előtt tartani, hogy a maguk idejében (mindössze néhány éve) az ide vezető lépések külön-külön is áttörésnek számítottak a tudományterületen belül.

2. AZ ELSŐ ELŐTANÍTOTT NYELVMODELLEK

Az első jelentősebb fejlemény az *előtanított nyelvmodellek*, vagyis nagy mennyiségű szövegen tanított neurális hálók megjelenése volt 2013-ban, illetve 2014-ben (Mikolov et al., 2013a, 2013b; Pennington et al., 2014). Mindkét nyelvmodell – némileg eltérő módon – a szójelentések egy geometriai reprezentációját tanulta meg, melynek eredményeként az egyes szavak jelentését egy sokdimenziós vektortérbe képezték le. A *szavak vektorábrázolása* a lexikális jelentés egy akkor még szokatlanul mély reprezentációját nyújtotta. Például a jelentéshasonlóság a vektorok között bezárt szögként lett értelmezhető, így például a 'kutya' szónak megfeleltethető vektor kisebb szöget zár be a 'macska' szó vektorával, mint például a 'szekrény' szóhoz tartozó vektorral.

Továbbá, bizonyos esetekben a szójelentés egyes komponensei leképezhetővé váltak vektorműveletekre (összeadás, kivonás). A leghíresebb kapcsolódó példa, hogy a 'woman' (nő) és a 'man' (férfi) különbsége megegyezik a 'queen' (királynő) és 'king' (király) különbségével. Vagyis: *queen* – *king* = *woman* – *man*. Az első előtanított nyelvmodellek *sekély*, egyetlenegy rejtett réteggel rendelkező *neurális hálók* voltak, amelyeket nagy mennyiségű szövegen tanítottak: a neurális hálókat olyan feladatokra tanítják, amelyek egy szövegbeli célszó és annak lehetséges kontextusainak viszonyát vizsgálják. Ez lehet akár a célszó predikciója a kontextus alapján vagy a célszó reprezentációja a kontextusok gyakorisága alapján.

Fontos kiemelni, hogy – bár a konkrét tanítási feladat nyelvmodellenként elteríthet – ezek mindig a *szóelőfordulások kontextusait* ragadják meg valamilyen módon, és ez alapján készítik el a szavak, szóelőfordulások vektorreprezentációit. A szóelőfordulás kontextusainak vizsgálata az ún. *disztribúciós vizsgálat*.

A nyelvészetben a *disztribúció* fogalma alatt azon környezetek összességét értjük, amelyekben egy nyelvi elem megjelenhet (Harris, 1954).

Az alábbiakban egy példával illusztráljuk, hogy hogyan lehet a szavakat vektorokkal ábrázolni. Ehhez egy egyszerű nyelvészeti jelenség elemi reprezentációját választottuk: a mellékneveket reprezentáljuk a főnévi kontextusok gyakorisága alapján. Az *1. táblázatban* egy mátrixot látunk, amelynek sorait melléknevek alkotják, a melléknév-vektorok koordinátái pedig a mellékneveket közvetlenül követő főnevek korpuszbeli gyakoriságai.

1. táblázat. Egy egyszerű melléknév-főnév gyakorisági mátrix: a mátrix soraiban a melléknevek szerepelnek, amelyeket a mellettük közvetlenül előforduló főnevek korpuszbeli gyakoriságával jellemzünk

	Vélemény	Megítélés	Száj	Ember	Gyerek	Ülés	Sajtótájékoztató
szubjektív	598	246	0	5	0	0	0
éhes	0	0	274	237	185	0	0
keddi	0	0	0	0	0	2348	1558
szerdai	0	0	0	2	3	2726	1896

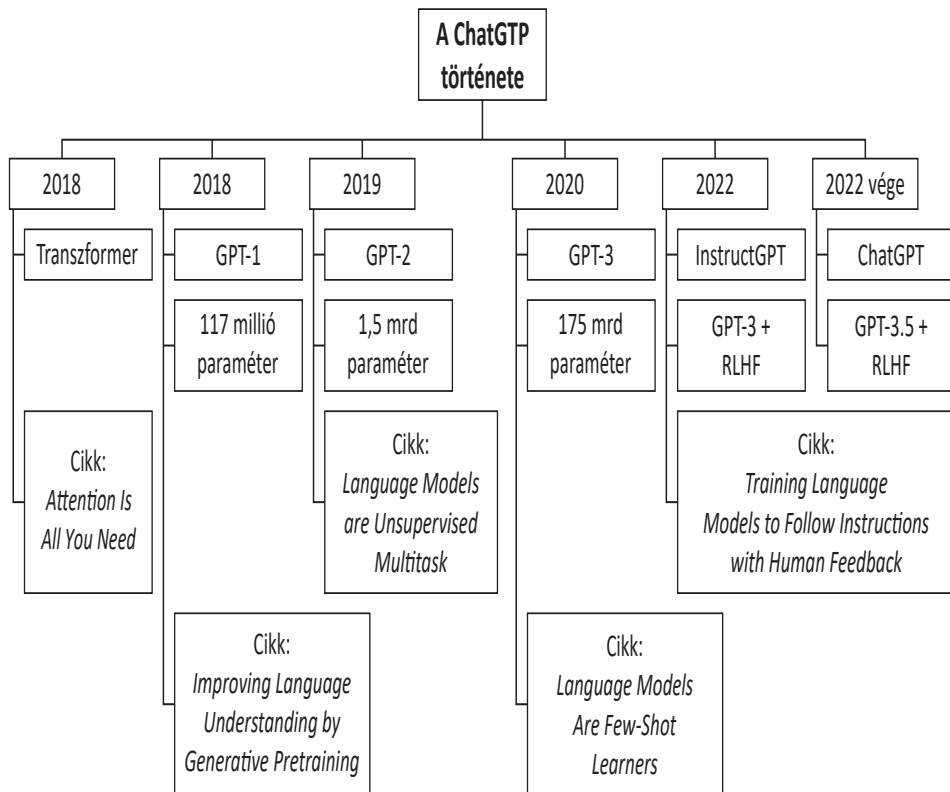
Már ezen az egyszerű reprezentáción is láthatjuk, hogy a *szubjektív* és az *éhes* melléknevek teljesen más főnevek előtt fordulnak elő, míg a *keddi* és a *szerdai* melléknevek hasonló gyakorisággal fordulnak elő ugyanazon főnevek előtt (*ülés, sajtótájékoztató*). Így a kontextusok korpuszbeli gyakorisága alapján következtethetünk a célszavak jelentésbeli hasonlóságára.

Bár a fenti megközelítés nagy előnye, hogy a modellezés során *objektív*, megfigyelhető adatokból indulunk ki, két alapvető problémát ki kell emelnünk. Egyfelől, a vizsgált jelenség jellemzéséhez használt jegykészlet nyelvészeti tudást igényel, így az elméletfüggő. Másfelől – mivel a melléknevek többsége a legtöbb főnévvel egyáltalán nem fordul elő – az így létrejövő vektorok sok 0-t tartalmazó ún. *ritka vektorok*, amelyek dimenziószáma is igen magas: megegyezik az összes olyan korpuszban előforduló főnév számával, amely közvetlenül követ egy melléknevet. A legtöbb gépi tanuló algoritmus azonban nem tudja hatékonyan kezelni a ritka vektorokat. Ehelyett alacsonyabb dimenziószámú, sűrű (kevesebb 0-t tartalmazó) vektorokra, ún. *szóbeágyazásokra* van szükség. A sekély neurális hálók tanításával létrejövő nyelvmodellek mindkét nehézséget kiküszöbölik: egyfelől, a kontextust kizárólag ablakméret (a figyelembe veendő bal oldali, illetve jobb oldali kontextusok mérete) és nem nyelvi jegyek alapján határozzák meg. Másfelől, az előtanított nyelvmodell a tanítás után súlyokkal ellátott, általában 300 dimenziós rejtett réteg lesz.

3. A CHATGPT TÖRTÉNETE

Bár a disztribúció fogalma régi, a ChatGPT evolúciója mindössze öt év alatt ment végbe, amit elsősorban az interneten fellelhető nagy mennyiségű digitális szöveg, valamint a nagy neurális hálók megjelenése tett lehetővé.

Az alábbiakban ennek a folyamatnak a főbb lépéseit mutatjuk be. A ChatGPT létrejöttéhez vezető út főbb állomásait az 1. ábra foglalja össze. Ezek közül a fontosabbakat a következőkben részletesebben is bemutatjuk.



1. ábra. A ChatGPT története (saját szerkesztés)

3.1. A nagy méretű nyelvmódel (LLM)

A ChatGPT-hez vezető első lépést a *transformer architektúrájú* mély neurális hálók jelentették 2018-ban. Mostani perspektívából nehéz elképzelni, hogy már ezek a nyelvmodellek is mennyire felforgatták az NLP-t. A legfontosabb változás az volt, hogy egyértelművé vált, nagy mennyiségű szövegen valamilyen egyszerű, disztribúciós feladatra tanított mély neurális hálók elképesztő mélységű és fedésű

„nyelvi tudást” – feltehetően mintafelismerési készséget – képesek felhalmozni. A nagy méretű nyelvmodellek is a már említett, *előtanított nyelvmodellek* csoportjába tartoznak, de míg a szójelentéseket reprezentáló első generációs neurális hálók durva becsléssel legfeljebb 10 millió paraméterből álltak, addig a ChatGPT őse, a GPT-1 már 117 millió paramétert tartalmazott. Tehát nemcsak a neurális háló architektúrája lett jóval bonyolultabb, de a paraméterek száma is legalább egy nagyságrenddel nagyobb lett. Visszatérve kedvenc hasonlatunkhoz, az LLM-ek olyanok, mint a svájci bicska, hiszen ezek általános tudást tartalmazó eszközök, amelyeket viszonylag kevés munkával sokfajta különböző specifikus feladatot elvégző eszközzé alakíthatunk tovább. Ezt a lépést *finomhangolásnak* (*fine-tuning*) nevezzük, amelynek lényege, hogy a pusztán kontextuális információon tanított LLM-et a kívánt feladatnak megfelelő adatokon tanítjuk tovább. A névelemfelismerés kontextusában maradvá vegyük a következő mondatot: „*Arany János reggel Marival kávézott a szegedi Tisza Hotelben.*” Ha meg szeretnénk tudni, hogy ez a mondat milyen tulajdonneveket tartalmaz, és ezek milyen típusú entitásokra referálnak, akkor építenünk kell egy *tanítókorpuszt*, amely a feladatnak megfelelően címkézett adatokat tartalmazza, majd ezen a címkézett korpuszon tovább kell tanítani az előtanított nyelvmodellünket. Pusztán a szemléltetés kedvéért, a fenti példamondat a megfelelő címkékkel nagyjából ilyen lenne:

<<”Arany János”=személy>reggel<<”Marival”=személy>kávézott a<<”szegedi Tisza Hotelben”=hely>.

Nagyon fontos hangsúlyozni, hogy a finomhangolás során az előtanított nyelvmodellt ténylegesen továbbtanítjuk, tehát az eredeti nyelvmodell súlyai módosulnak a felcímkézett tanítóadat alapján. A svájci bicska hasonlat pedig azt a fontos tényt szemlélteti, hogy a finomhangolás feleslegessé tette az addig szokásos *feladatspecifikus* neurálisháló-architektúrákat vagy az egyéb specifikus eszközök fejlesztését.

3.2. Már megint egy új paradigma? A promptok

A finomhangolás hátrányai

Alig két év telt el a transzformer alapú LLM-ek megjelenése után, amikor Tom B. Brown és kollégái, az OpenAI² munkatársai *Language Models Are Few-Shot Learners* című cikkükben (2020) azt vizsgálták, hogy milyen hátulütői vannak a finomhangolásnak.

² A 2015-ben alapított OpenAI mesterséges intelligenciával kapcsolatos kutatásokat folytat. Célul tűzte ki az általános mesterséges intelligencia (AGI – artificial general intelligence) kifejlesztését. Ez az elnevezés olyan „nagymértékben autonóm rendszerekre utal, amelyek képesek felülmúlni az embereket a legtöbb gazdaságilag értékes munkában” (OpenAI Charter).

Először is, nem áll és nem is állhat rendelkezésre minden lehetséges feladat vonatkozásában címkézett adat, hiszen ezek létrehozása rendkívül erőforrás-igényes. Némileg előreszaladva, ha összehasonlítjuk a finomhangolt nyelvmodelleket és a ChatGPT-t, észrevehetjük, hogy az utóbbi esetében már fordítva gondolkozunk. Az eddig megszokott módszerrel szemben nem az éppen felmerülő feladatokhoz fejlesztünk egy alkalmazást, hanem pont fordítva: rendelkezésünkre áll egy univerzális eszköz, és azt próbáljuk feltérképezni, hogy milyen feladatokra lehet ezt egyáltalán használni.

Másodszor, a finomhangolás során megszerzett feladatspecifikus tudás sok esetben *nehezen általánosítható új adatokra*, így a kiértékelés során mért eredmények gyakran jobbak a nyelvmodell tényleges eredményeinél.

Harmadrészt, pszichológiaiilag nem hiteles a finomhangolás paradigmája: az emberi idegrendszer nem címkézett adatokon tanul meg konkrét nyelvi feladatokat, hanem konkrét utasításokat követünk, amelyeket sok esetben példakkal is szemléltetnek (például: „*Mondd meg, hogy az alábbi vélemény pozitív-e!*”). Így a transzformeralapú neurális háló megjelenése után nem sokkal a kutatók már azt vizsgálták, hogy vajon kiváltható-e a finomhangolás.

Promptok és példák

A GPT-3 2020-as megjelenésével ennek a kutatásnak az eredménye rövid időn belül a sokadik nagy meglepetést okozta szakmai körökben. Engedve a pszichológiai analógia csábításának, az OpenAI kutatói arra a kérdésre keresték a választ, hogy milyen hatással van az előtanított nyelvmodell feladatmegoldó készségére, ha az LLM-et rövid természetes nyelvi utasításokkal próbálják rávenni arra, hogy azt csinálja, amit elvárnak tőle. Ezek a rövid természetes nyelvi utasítások a *promptok*. Sok esetben a promptok mellett a feladatnak megfelelő példákat is mutatnak a rendszernek. Ezt hívják *one-shot*, illetve *few-shot* tanulásnak. Az előbbi esetben a prompt egy példával, az utóbbi esetben néhány példával van kiegészítve, ahol a példák száma általában 10 és 100 közötti. Az így megadott példák száma nagyságrendekkel kisebb, mint amennyi tanítóadatra szükség van finomhangolás esetén.

A GPT-3 egy ún. *generatív nyelvmodell (Generative Pre-trained Transformer)*, amely szabad szöveg generálására lett kifejlesztve. Különleges képessége, hogy képes bármilyen promptot – akár példakkal is kiegészítve – számtalan módon folytatni, a jazz-zenészek improvizációs készségét idézi, akik egyetlen témából végtelen számú szót képesek létrehozni. Miután kapnak egy promptot, ezek a modellek számtalan különböző választ képesek abból generálni, amelyek stílusukban, hangnembükben mind különbözhetnek.

A kutatók tehát arra a kérdésre keresték a választ, hogy milyen promptokkal lehet ezt a nyelvmodellt rábírni arra, hogy az a finomhangoláshoz hasonló eredményeket produkáljon. Azt találták, hogy elég egy néhány példát is tartalmazó

prompt (*few-shot* tanulás), hogy a GPT-3 a hagyományos, finomhangolt modellekkel összehasonlítható eredményeket érjen el a tesztfeladatokon. Például, bár a GPT-3 tanítóanyagának mindössze 7%-a nem angol nyelvű szöveg, számos nyelvről (például: francia, német, román) jobban fordított angolra, mint a direkt gépi fordítás céljára kifejlesztett legfrissebb neurális gépi fordító rendszerek. Ez az eredmény szinte hihetetlen: a gépi fordítás nyelvpáronként párhuzamos adatokat tételez – vagyis olyan szövegeket, amelyek egymás fordításai – ezzel szemben a GPT-3 csekély mennyiségű nem angol tanítóadaton tanult, amelyben a különböző nyelvű szövegek random fordulnak elő.

A kísérletek során egyértelműen kiderült, hogy a nyelvmodell mérete, vagyis a beállításra kerülő súlyok száma nagyban meghatározza, hogy mennyire eredményes a *few-shot* tanulás. Különösen meglepő, hogy a GPT-3 konzisztensen jó teljesítménye *few-shot* tanulás esetén nem a szó szoros értelmében vett tanulás eredménye, hiszen a nyelvmodell súlyai már nem módosulnak tovább, mint a finomhangolás során. Nem értjük pontosan, hogy valóságos tanulás nélkül a promptok hogyan képesek befolyásolni a nyelvmodell működését. Ennek megfejtése az NLP egyik fontos új kutatási iránya.

Sajnos azonban a nagy nyelvmodellek promptolása számos nem kívánt viselkedést is eredményez: a válaszok gyakran nem tényszerűek (ilyenkor azt mondjuk, hogy a nyelvmodell hallucinál), ugyanarra a promptra sok esetben nem megjósolható módon különböző válaszok születnek; a válaszok gyakran társadalmi előítéleteket jelenítenek meg, vagy egész egyszerűen nem tükrözik azt, amit a felhasználó szeretne.

3.3. ChatGPT: hogyan regulázzuk meg a nyelvmodellt?

A GPT-3 esetében tulajdonképpen egyáltalán nem meglepő az eltérés az elvárt és a valós viselkedés között, hiszen az eredeti disztribúciós tanítási feladat: „*Jó-sold meg, hogy mi lesz a következő szó egy szövegben!*” igencsak eltér attól, amit elvárunk: „*Kövessd a felhasználó utasításait anélkül, hogy udvariatlan vagy előítéletes lennél!*”. Így a kutatók arra a kérdésre kezdték el keresni a választ, hogy hogyan lehet elérni, hogy az LLM-ek a felhasználók szándékának megfelelően viselkedjenek az alapvető emberi értékek tiszteletben tartásával. Ez a kutatási irány az ún. *alignment research*, amelynek eredményeképpen 2022 novemberében megjelent a ChatGPT³ (Ouyang et al., 2022).

³ A ChatGPT-nek nincsen hivatalos cikke, azaz az OpenAI nem írt olyan cikket, amely specifikusan a ChatGPT-re fókuszált volna, ehelyett általánosságban írtak az utasításkövető *instruct* modellekről. A chat mód az utasításkövetés egy alesetének tekinthető. Ezért itt az InstructGPT-t mutatjuk be, amely elsősorban a mögöttes GPT-modellben tér el a ChatGPT-től.

A feladatból viszonylag természetesen adódik a következő lépés, az eddigi promptok és promptokra adott válaszok figyelembevétele elsősorban abból a szempontból, hogy mennyiben tükrözik a felhasználó szándékát. A ChatGPT tanítása során éppen ez történt: gépi tanulási eljárással elérték, hogy a GPT-3 figyelembe vegye azt, hogy a múltban mennyire adott jó válaszokat a promptokra. Ez a módszer a *mege erősítés es tanulás emberi visszajelzés alapján (RLHF – reinforcement learning from human feedback)*. Az RLHF-t általában olyan feladatokra használják, amikor nehéz jó veszteségfüggvényt találni a megoldásra, ám az ember számára mégis könnyű eldönteni, hogy mennyire jó a megoldás. Ilyen például annak eldöntése, hogy egy nyelvmó dell kimenete mennyire vicces, vagy egy promptra adott válasz mennyire jó.

Az RLHF leegyszerűsítve két fő lépésből áll: az első lépésben egy jutalmazási mó dellt tanítunk, amely egy tetszőleges <prompt, kimenet> párhoz egy értéket rendel, például 1-től 5-ig. Ez az érték azt tükrözi, hogy az adott prompthoz mennyire jó az adott kimenet. A jutalmazási mó dell célja, hogy kiváltsa az emberi címkézést a ChatGPT tanítása során, de a jutalmazási mó dell tanításához szükség van emberek által címkézett tanító adatokra is. Így a tanítóadat létrehozása úgy történik, hogy a <prompt, kimenet> párokhoz emberek egy értéket rendelnek aszerint, hogy mennyire tartják jónak a kimenetet. A következő lépésben mege erősítés es tanulással finomhangoljuk az eredeti nyelvmó dellt, jelen esetben a GPT-3-at optimalizálták a jutalmazási mó dell alapján.

Úgy tűnik, hogy az RLHF segítségével sikerült kiküszöbölni a GPT-3 gyermekbetegségeit: a ChatGPT már sokkal inkább a felhasználó szándékainak megfelelően válaszol, és válaszai sokkal következetesebbek, mint a GPT-3 válaszai. Példánknál maradván, nézzük meg, hogy hogyan oldja meg a ChatGPT a névelim-felismerési, illetve a szentimentelemzési feladatunkat.

Első példa

Felhasználó:

Kérlek, sorold fel, hogy milyen tulajdonneveket tartalmaz a következő mondat, és azt is, hogy ezek a tulajdonnevek milyen típusú entitásra referálnak! „Arany János reggel Marival kávézott a szegedi Tisza Hotelben.”

ChatGPT:

A következő mondatban található tulajdonnevek és az általuk referált típusú entitások:

[..]

„Arany János” egy személyre utal.

„Marival” egy másik személyre utal.

„Tisza Hotel” egy konkrét helyet jelöl, valószínűleg egy szállodát.

„Szeged” egy másik konkrét helyre utal, valószínűleg városra.

Itt érdemes kiemelni, hogy a ChatGPT mennyire pontos elemzést adott: a névelim-felismerés során általában a <HELY>-et egy atomi egységnek kezeljük, a nyelvmodell ezzel szemben azt is megmondta, hogy a <HELY> kategórián belül egy szállodáról van szó és egy városról. Vegyük észre, hogy a ChatGPT már nem volt ilyen éles szemű „Arany János” esetében, hiszen elvárás lehetne, hogy tudja, hogy ez a tulajdonnév egy híres költőre utal. Ugyanakkor ez a viselkedés összhangban van a jelenlegi névelim-felismerő rendszerek működésével, amelyek általában felismerik a személyneveket, de nem osztályozzák őket tovább. Most nézzük meg, hogy ugyancsak külön finomhangolás nélkül mire jutott a példamondattal a ChatGPT a szentimentelemzés kontextusában!

Második példa

Felhasználó:

Kérlek, jelöld meg, hogy milyen típusú érzelmet fejez ki a következő mondat: „Arany János reggel Marival kávézott a szegedi Tisza Hotelben.”

ChatGPT:

A következő mondat semleges érzelemkifejezést tartalmaz. A mondat pusztán egy információt közöl, egy esemény leírását adja, de az érzelem nem jelenik meg benne. Nem érzelmileg színezett vagy hangsúlyozott az állítás, csak tényeket tartalmaz.

Ez a válasz is meglepően jó. A szentimentelemzés során általában két (negatív, pozitív) vagy három (negatív, semleges, pozitív) osztályba soroljuk be a vizsgált szövegeket. Különösen érdekes, hogy a ChatGPT nemcsak azt mondta meg, hogy az állítás semleges, de azt is, hogy milyen módon az: „pusztán információt közöl”.

A fenti példák is jól szemléltetik, hogy a ChatGPT rendkívüli áttörést jelentett az NLP-ben. Kiemelendő, hogy a nyelvmodell jól általánosítja az utasítás fogalmát, így nem csak azokat a típusú utasításokat képes követni, amelyekkel a(z) RLHF finomhangolás történt.

Ráadásul olyan nyelveken adott utasításokra is tud helyesen válaszolni az adott nyelven, amelyek alulreprezentáltak az előtanított GPT-3-nyelvmodell és a <prompt, kimenet> adatbázis tanítóanyagában is. Emlékezzünk vissza, hogy a GPT-3 esetében a tanítóanyag 93%-a angol, és a finomhangolásra használt <prompt, kimenet> adatbázisnak is csupán 4%-a nem angol nyelvű szöveg.

4. MAGYAR FEJLEMÉNYEK: PULI GPT-3SX

Most nézzük meg, hogy mi minden történt Magyarországon az előtanított LLM-ek kapcsán! Itt érdemes kiemelni a magyar szövegeken tanított GPT-3 architektúrájú neurális hálót: ez egy kb. 32 milliárd szövegszón tanított 6,7 milliárd pa-

raméteres modell, amelyet a Nyelvtudományi Kutatóközpont Nyelvtechnológiai Kutatócsoportjában fejlesztettek⁴ (Zijian et al., 2023b). Annak ellenére, hogy az eredeti GPT-3-at több mint tízszer nagyobb korpuszon tanították, mint ezt a modellt, és súlyainak száma is körülbelül 25-ször nagyobb, a PULI GPT-3SX⁵ is figyelemre méltó képességekkel rendelkezik. Az alábbi példák azt szemléltetik, hogy ez a nyelvmódellem nemcsak arra képes, hogy *few-shot* tanulással világismertet igénylő kategóriákat ragadjon meg, de azt is megérti, hogy mit szeretnénk megtudni egy adott kategóriába tartozó konkrét entitásról, illetve a megfelelő választ is megkapjuk.

Első példa

1. feladat: Arra vagyunk kíváncsiak, hogy a PULI GPT-3SX tudja-e, hogy mi Friedrich Nietzsche foglalkozása. Amint látjuk, három példa elég ahhoz, hogy a modell megértse, hogy mi a feladat. Ezenkívül láthatóan rendelkezik tudással arról, hogy kik a híres emberek, milyen szakmák léteznek, és az egyes híres emberek szakmáját is ismeri.

Felhasználó

1. Mikszáth Kálmán = író
2. Bartók Béla = zeneszerző
3. Egerszegi Krisztina = úszó
4. Friedrich Nietzsche =

PULI GPT3-SX

filozófus

5. Szent-Györgyi Albert = orvos, Nobel-díjas
6. Márai Sándor = író
7. Munkácsy Mihály = festő
8. Bálint Sándor = író
9. Kármán Tódor = repülőgép-tervező
10. Kós Károly = építész

De a nyelvmódellem más kategóriákról is rendelkezik ismerettel. Nézzük a következő feladatot.

⁴ A kutatócsoport munkatársai a PULI GPT-3SX után finomhangolással létrehoztak egy utasításkövető modellt is, a ParancsPULI-t (Zijian et al., 2023a), amely tudomásunk szerint az első utasításkövető modell Magyarországon. A ParancsPULI ezen az URL-en érhető el: URL2.

⁵ A PULI elnevezés a jelenlegi magyar GPT-3 modell méretére utal, amely egyelőre egy apróbb termetű neurális architektúra a hatalmas neurális hálók világában.

2. feladat: Azt szeretnénk megtudni, hogy a 'Budapest' szó milyen karakterrel kezdődik. Akárcsak az előzőekben, a nyelvmodell itt is felismeri, hogy mi a feladat, és folytatja a sort. A példa azt mutatja, hogy a nyelvmodell rendelkezik tudással arról, hogy milyen magyar városok vannak, és hogy ezek nevei milyen karakterrel kezdődnek.

Felhasználó

Tata első betűje = t

Zirc első betűje = z

Sopron első betűje = s

Eger első betűje = e

Budapest első betűje =

PULI GPT3-SX

b

Debrecen első betűje = d

Pécs első betűje = p

Nagykanizsa első betűje = n

Szeged első betűje = s

A PULI GPT-3SX ezen a linken tesztelhető: [URL3](#)

5. A NYELVMODELLEK LEHETSÉGES VESZÉLYEI

Mint minden, első látásra kiváló és forradalmi eszköznek, az LLM-enek is megvannak a maguk lehetséges hátulütői. Először is, jelentős kockázata van annak, ha az emberek túlságosan megbíznak ezekben a modellekben. Mivel általában kifinomult és természetesnek tűnő válaszokat adnak, az emberek hajlamosak lehetnek kritikai gondolkodás nélkül elfogadni ezeket a válaszokat. Ez különösen kockázatos lehet, ha a modellek hibás, pontatlan vagy félrevezető információkat adnak. Márpedig – tekintve, hogy kizárólag az abban látott adatok alapján generálják a szövegeket – a tanítóanyagban található pontatlanságokat, elavult információkat és előítéleteket nemcsak hogy megjeleníthetik, de akár fel is erősíthetik a válasz generálása során.

További aggodalomra adhat okot, hogy ezeket a modelleket rosszindulatú célokra is lehet használni. Mivel a ChatGPT-hez hasonló modellek képesek emberhez hasonló szöveget generálni, felhasználhatók hamis hírek létrehozására, digitális kommunikációban személyek megszemélyesítésére vagy akár erőszakra buzdító, vagy szélsőséges ideológiákat népszerűsítő káros tartalmak előállítására. Megfelelő óvintézkedések vagy a gép által generált és az ember által írt tartalmak

megkülönböztetésének módja nélkül ezeknek a nyelvmodelleknek a bizalomra és az igazságra gyakorolt hatása – különösképpen így a digitális korban – mélyen aggasztó.

Továbbá a „szuperintelligens” nyelvi modellek mindenütt való jelenléte és hatékonysága hozzájárulhat a kritikus gondolkodás hanyatlásához és az eredeti, egyedi tartalmak létrehozásának visszaszorulásához. Ha az egyének és a vállalkozások túlzottan az automatizált eredményekre hagyatkoznak, fennáll a kommunikáció és a kreatív kifejezőmód homogenizálódásának veszélye, ami megfojtja az innovációt, és csökkenti az emberi diskurzus gazdagságát.

IRODALOMJEGYZÉK

- Brown, Tom B. – Mann, Benjamin – Ryder, Nick et al. (2020): *Language Models Are Few-Shot Learners*. DOI: 10.48550/arXiv.2005.14165. <https://arxiv.org/abs/2005.14165>
- Harris, Zellig S. (1954): Distributional Structure. *Word*, 10, 2–3, 146–162. DOI: 10.1080/00437956.1954.11659520
- Mikolov, Tomas – Chen, Kai – Corrado, Greg et al. (2013a): *Efficient Estimation of Word Representations in Vector Space*. <https://arxiv.org/abs/1301.3781>
- Mikolov, Tomas – Sutskever, Ilya – Chen, Kai et al. (2013b): Distributed Representations of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems*, 26, DOI: 10.48550/arXiv.1310.4546. <https://arxiv.org/abs/1310.4546>
- Ouyang Long – Wu, Jeff – Jiang, Xu et al. (2022): *Training Language Models to Follow Instructions with Human Feedback*. DOI: 10.48550/arXiv.2203.02155. <https://arxiv.org/abs/2203.02155>
- Pennington, Jeffrey – Socher, Richard – Manning, Christopher D. M (2014): GloVe: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar, Association for Computational Linguistics, 1532–1543. DOI: 10.3115/v1/D14-1162. <https://aclanthology.org/D14-1162.pdf>
- Radford, Alec – Narasimhan, Karthik – Salimans, Tim et al. (2018): *Improving Language Understanding by Generative Pre-Training*. OpenAI. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
- Radford, Alec – Wu, Jeffrey – Child, Rewon et al. (2019): *Language Models are Unsupervised Multitask Learners*. OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- Vaswani, Ashish – Shazeer, Noam – Parmar, Niki et al. (2017): Attention is All You Need. *Advances in Neural Information Processing Systems*, 30, DOI: 10.48550/arXiv.1706.03762. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- Yang Zijian Győző – Laki László János – Váradi Tamás et al. (2023a): Mono- and Multilingual GPT-3 Models for Hungarian. In: *26th International Conference of Text, Speech and Dialogue (TSD2023)*, Plzeň, Czech Republic. http://real.mtak.hu/173960/1/TSD_2023_GPT.pdf
- Yang Zijian Győző – Dodé Réka – Ferenczi Gergő et al. (2023b): *Jönnek a nagyok! BERT-Large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre*. XIX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2023. január 26–27. https://acta.bibl.u-szeged.hu/78417/1/msznykonf_019_247-262..pdf

URL1: <https://index.hu/kultur/2023/03/07/chatgpt-generalt-szoveg-vers-cikk-hazugsag-jozsef-attila-krubi-haiku/>

URL2: <https://juniper.nytud.hu/demo/parancspuli>

URL3: <https://juniper.nytud.hu/demo/puli>