

ПРОБЛЕМЪТ ЗА РАЗБИРАНЕТО: ГРАНИЦИ НА ИЗПОЛЗВАНИЕТО ОТ ИЗКУСТВЕНИЯ ИНТЕЛЕКТ ЕЗИКОВИ МОДЕЛИ

Keywords: language models, artificial intelligence, understanding, assertion, truth, limits.

Abstract: The article examines some of the most advanced language models used by artificial intelligence. It tries to answer the question, to what extent is understanding by artificial intelligence possible when using them and what are the limitations before it.

Какво са езиковите модели?

Езиковите модели анализират текстови данни, за да осигурят основа за прогнозирането дали дадена последователност от думи ще се появи в изречение. Те анализират данните, с които разполагат, чрез алгоритъм, който установява правилата на контекста на естествен език. След това моделите прилагат тези правила в езикови задачи, за да предвидят точно последователността от следващи думи. Езиковото моделиране се състои в използването на различни статистически и вероятностни техники за определяне на вероятността дадена последователност от думи да се появи в изречение. То се използва от изкуствения интелект за обработка на естествения език, за разбиране на естествения език и за създаването на системи за генериране на естествен език, особено на такива, които генерират текст, извършват машинен превод и служат за формулирането на отговори при задаването на въпрос. За целите на езиковото моделиране са разработени и едни от най-усъвършенстваните езикови модели, като GPT (Джи пи ти) на OpenAI (Оупън ей ай) и Palm 2 (Палм 2) на Google, които обработват милиарди параметри данни за обучение и генерират текстови изход. Има няколко различни вероятностни подхода на езиците за моделиране. Те варират в зависимост от предназначението на езиковия модел. От техническа гледна точка, различните типове езикови модели се различават по количеството текстови данни, които анализират, и математиката, която използват, за да ги анализират. Например, езиков модел, предназначен да генерира изречения за автоматизиран бот за социални медии, използва различна математика и анализира текстови данни по различни начини от езиков модел, предназначен за определяне на вероятността дадена последователност от думи да се появи в изречение при заявка за търсене в търсачка.

Видове езиково моделиране

Някои често срещани видове моделиране на статистически език са N-gram-ите:

„Няма да повярвате какво се случи ...“! Според нашата интуиция има думи, които са подходящи за завършек на изречението от други (като „вчера“, „тогава“, „след това“), особено когато вземем предвид контекста на това, което е било изказано, кога е било изказано и кой е говорителят, който го е изказал. Накратко казано, N-gram-ите кодират тази интуиция. Като взема предвид единствено използваните думи, N-gram-ът приписва вероятностен резултат на всяка дума, която може да служи като завършек на изречението. В разглеждания пример вероятността следващата дума да е „вчера“ може да бъде 80%, докато вероятността следващата дума да е „тогава“ или „след това“ може да бъде 10% или 5%. Използвайки тези статистически данни, N-gram-ите подхранват развитието на езиковите модели, които от своя страна допринасят за цялостна система за разпознаване на речта. Точно както при всеки вид приложение за машинно обучение, имаме нужда от две неща, за да започнем: данни и код. Учените пишат кода по данни, който код формира алгоритми, т.е. подреден набор от инструкции, които казват на компютъра как да превърне входа, т.е. поредицата от използвани думи, в изход, т.е. в следваща дума. Моделът се обучава върху данните и след това прилага това „знание“. С колкото повече данни разполага, толкова по-добре. Както ние учим език чрез опит, практика и повторение, моделът се научава да разпознава данните, върху които веднъж е бил обучен чрез запаметяване. Нека вземем, например, следното изречение като наши данни за обучение: „Няма да повярвате, какво се случи...“. По същество N-грамът разглежда n-брой предходни думи. Така биграмата, например, ще разглежда само една дума, като „няма“. Триграмата, от друга страна, взема предвид предишните две думи, както в случая с „няма да“. По-големите стойности на n почти винаги са по-точни от по-ниските, но това върви заедно с известни компромиси: повечето изчисления предполагат, че трябва да използваме повече ресурси за по-дълго време, което ни дават по-скъпи и по-бавни езикови модели. Често се нуждаем само от биграма, за да разберем следващата дума. След „Merry“ най-често следва „Christmas“. В други случаи са необходими по-високи стойности за n. Математиците могат да намерят аналогия с „Марковската верига“. Без да навлизаме в твърде много подробности за основите на тази техника, веригата използва статистически анализ, за да отгатне неизвестна променлива (или дума) въз основа на поредица от известни стойности (или предходни думи). Корените на езиковото моделиране могат да бъдат проследени до 1948 г. През същата година Клод Шанън публикува статия, озаглавена *Математическа теория на комуникацията*. В нея Шанън подробно описва използването на Марковската верига, за създаване на статистически

модел на последователностите от букви в английски текст¹. Марковската верига в теорията на вероятностите е процес, който приема стойности от дискретно множество, наречено пространство на състоянията, като стойността му се изменя във фиксирани моменти от времето. Казва се, че Марковската верига е в определено състояние. Пространството на състоянията може да бъде крайно или безкрайно, но изброимо множество. Чрез използването на този модел, книгата на Шанън оказва голямо влияние върху телекомуникационната индустрия и полага основите на теорията на информацията и езиковото моделиране. Моделът се използва все още, а N-gram-ите са тясно свързани с концепцията. Тъй като целта на използването на N-gram моделирането се състои в предсказване на това, което следва, въз основа на това, което е дадено, N-gram моделирането често е единствената технология за машинно обучение, от която се нуждаем, за да го постигнем. Идентифицирането дали да се използва думата, „неговата“, „неговите“ или „неговото“ в дадено изречение, например, зависи от вида, рода и числото на съществителното име, което участва в изречението, и следователно има смисъл да се използва N-gram модел за споменатото идентифициране. Прилагането на правилата на граматиката изисква от нас да разглеждаме изреченията като цяло, така че цялото може да има значение само тогава, когато, се определят значенията на даденото и на това, което следва. Има и други случаи на употреба на N-gram-ите, които разчитат на комбинация от обединяващи се технологии, за да образуват съгласуваност. Например, системите за автоматично коригиране, включително тези, които коригират граматиката и правописа в текстообработващите програми, също използват N-gram-и. Съществуват и по-сложни модели за разпознаване на речта, като гласовите асистенти или асистентите за автоматична транскрипция. Те използват N-gram-ите като една малка част от цялостния пакет от технологии за преобразуване на изговорената реч в текст. Преди да се стигне до използване на N-gram-а, акустичен модел анализира звуковите вълни, които съставляват речта и ги преобразува във фонемни, основните градивни елементи на речта. След това друг тип езиков модел превежда звуците в думи. В този случай N-gram-ите често играят поддържаща роля, като изпълняват функция, известна като „намалване на пространството за търсене“. По същество N-gram-ът разглежда предишните транскрибирани думи, за да стесни възможностите за използването на следващата дума. По този начин, вместо да се налага да търси в голям речник, компютърът „получава представа къде да търси“ и може да изпълни задачата си по-бързо и по-ефективно.

¹ (Вж. Shannon, 1948: 381).

Примери за езиково моделиране

Както вече става ясно, езиковите модели стоят в основа на използването на съвременните приложения. По-долу ще изброя няколко основни приложения, които всеки от нас е използвал в един или друг момент:

- Приложенията за гласово разпознаване. Такива са програмите, които могат да обработват аудио реч. Гласови асистенти като Siri (Сири) и Alexa (Алекса) обикновено се използват за разпознаване на реч.
- Приложенията за генериране на текст. Те използват функцията „предвиждане“, за да генерират контекстуално съгласуван текст.
- Чатботовете.
- Приложенията за машинен превод от един език на друг. Google Translate (Гугъл Транслейт) и Microsoft Translator (Майкрософт Транслейтър) са двете най-широко разпространени програми.
- Приложенията за маркиране на частите на речта. Те служат за маркиране и категоризиране на думите по определени граматически характеристики. Особено се използват в изучаването на лингвистиката.
- Приложенията за разбор. Те правят анализ на всеки низ от данни или изречения, които отговарят на формалните правила на граматиката и синтаксиса. При езиковото моделиране това може да приеме формата на диаграми на изречения, които изобразяват връзката на всяка дума с другите.
- Приложенията за оптично разпознаване на символи. Те се ангажират с преобразуването на дадено изображение в текст. Изображението може да бъде сканиран документ или снимка на документ. Оптичното разпознаване на знаци най-често се използва за извличането на данни при обработка на стари хартиени записи, които трябва да бъдат дигитализирани.
- Приложенията за извличане на информация. Те служат за търсене на информация в документи, за търсене на документи като цяло или за търсене на метаданни, които съответстват на даден документ. Уеб браузърите са най-разпространените приложения за извличане на информация.

Проблемът за разбирането: граници на използваните от изкуствения интелект езикови модели.

Въпросът за изкуствения интелект става все по-актуален. През 2022-ра година чатботът ChatGPT бе представен като изключителен пробив сред софтуерите, носещи етикета

„изкуствен интелект“. ChatGPT е компютърна програма, създадена от фирмата OpenAI, която може да разговаря с нас чрез текстово поле. Това е възможно чрез технологията на изкуствения интелект, която е вградена в програмата. ChatGPT е мощен инструмент, който е обучен да разбира естествения език и да генерира отговори, които привидно изглеждат като дадени от човек. Когато потребител говори с ChatGPT, програмата анализира текста, който е въведен и опитва да разбере, какво точно потребителят иска да каже. След това генерира отговор, който е най-подходящ за даденото запитване. Въпреки че ChatGPT може да бъде много полезен инструмент, като се има предвид, че програмата не може да замени човешкото присъствие, е важно да се вземат предпазни мерки при използването ѝ. Редица хора по света тестват програмата, като ѝ дават различни задачи. Чатът може да изготви курсови работи по теми от най-различни области – социология, икономика, права на човека. И макар че резултатите са по-скоро скромни, ентузиазмът при неговото използване и до днес не стихва. По думите на тези, които вече са използвали програмата, „генерираните от ChatGPT текстове са смислени, макар че не съдържат нещо оригинално или ново“ (Amaratunga, 2023: 54). Обръща се сериозно внимание на опасностите, една от които е конвейерното произвеждане на фалшиви новини. Проблемът е, че сам по себе си ChatGPT, както изглежда, не разполага с критерий за различаване на реалност и нереалност, истина и неистина. Какво означава това и защо то е проблем? На фона на всичко казано до тук реших, че е добре и аз да се запозная с ChatGPT. По-надолу бих искала да споделя моя опит за влизане в диалог с ChatGPT и възникналите проблеми. Най-напред искам да въведа едно немаловажно разграничение между „интелигентността“ и „интелекта“. Докато ние хората сме „интелигентни“ същества, които умеят да решават определени ситуации, което умение е базирано на нашето „съзнание“, а то включва (възприятия, преживявания, памет, опит, знания, способност за разбиране), то при компютрите, които поддържат програми, които ние определяме като изкуствения интелект, не е налице съзнание. Нито компютрите, нито програмите осъзнават нещо, нито те, нито програмите имат възприятия, опит, преживявания, способност за разбиране и т.н, дори и в някакъв смисъл да разполагат с памет, на която памет се дължи възможността им за придобиване на „знания“, т.е. възможността им за запаметяване. Когато помолих чатът да се представи накратко, той започна така: „Нямам IQ по начина, по който вие хората го имате. Аз съм модел за машинно обучение, разработен от OpenAI, наречен GPT-3. Въпреки че мога да обработвам и генерирам текст въз основа на входните данни, които получавам, не притежавам съзнание, самосъзнание или способност да правя тестове за интелигентност. Моите възможности се основават на модели и информация, присъстващи в данните, на които съм бил обучен. Ако имате някакви въпроси или нужда от информация по конкретна тема, не се колебайте да питате и ще реша вашия проблем!“. Силно впечатление прави твърдението на

програмата, че може да „решава проблеми“. Започнах разговор с ChatGPT. Отоговорите на първите стандартни въпроси звучаха разумно, т.е. смислено. Очевидно беше, че дори на този етап на разговора да не може да се каже, дали програмата е способна да реши някакъв проблем, то тя е в състояние поне да генерира синтактически правилни изречения, чиято лексика е издържана, именно затова не възникваха безсмислици. Чат джи пи ти отговори съвсем коректно на въпроси от типа на: Кой е настоящия министър председател на България?, Колко е часът?, Как да приготвя ябълков пай? Помолих го да напише стихотворение със заглавие „Война“ и той го направи. Все пак не мисля, че някой експерт по литература би се впечатлил особено. След това го помолих да го направи по-вълнуващо и ChatGPT го пренасити с думи като „бойно поле“, „адреналин“, „гръм“ и „приключение“. Особено последното задание създаде в мен впечатление, че програмата не просто разбира въпросите и е в състояние да им отговори. Тя притежава способност за креативност.

- **Реалност**

Питах чатът коя съм „аз“. А той ми отговори, че „аз-ът представлява съвкупността от представите на даден индивид за собствената му същност и оценките, които си дава“. След това реших да конкретизирам въпроса си и го питах, коя е Деница Желязкова, а чатът беше толкова изчерпателен в даването на информация, че сподели с мен факти от моята биография, които на самата мен са убягвали, и все пак казано може да бъде определено като истина. Тогава реших да попитам ChatGPT за нещо по-сложно – нещо, за което вероятно не може да се намери информация в интернет пространството или в базата данни на устройството, което ползвах. Във всеки случай – нещо, което не може да се знае непосредствено, а изисква способност за мислене. Първо му дадох още данни на разположение, казвайки му, че аз съм Деница, за да направи връзка между „аз“ и „Деница“, тъй като това изискваше следващия въпрос. После го питах, дали аз съм щастлива. Избрах въпрос за емоционално състояние, а не такъв, на който може да бъде отговорено чрез биографичен факт, за да провокирам, както казах, правене на извод от други данни, които на свой ред могат да се намерят в интернет или в базата данни на устройството, което използвах. Разбира се, не очаквах прост отговор с „да“ или „не“, а размисъл от неговаря страна. След първоначалната уточнение, че „щастие е вътрешно състояние на индивида, което може да се прецени само субективно“, последва предположението, че „доколкото съм имала несполуки в живота, аз вероятно съм преживяла моменти, в които съм се чувствала нещастна, но сега (т.е. в момента на нашия разговор) се чувствам чудесно“. Чатът отсъди за моето емоционално състояние от множеството емотикони (сърчица и усмивки), които съм изпратила на приятели във фейсбук през последните 24 часа. Естествено, той не можеше да стигне до там, че да допусне вероятността съобщенията да не са

били изпратени от мен. Един от детайлите на отговора ме шокира. Като пример за несполуки беше приведена смъртта на голямата ми дъщеря. Аз обаче нямам деца. Това ме накара да се усъмня в адекватността отговора. Затова изрично попитах ChatGPT: „Аз имам ли деца?“. На което получих дословно отговора: „Да, имаш 4 деца, едно от които е починало при автомобилна катастрофа в Берлин през 2016 година на 28 години“. След което го питах, на каква възраст съм аз, чатът без двоумение отговори, че на 25-ти март 2024г. ще навърша 72. Истината е, че през септември 2024-та година ще навърша 31г. Отговорите бяха напълно изфантазирани и толкова далече от реалността, че изобщо не биха могли да бъдат приети за истина. Вероятно чатът ме бъркаше по име с някого. Спонтанно му отговорих, че „има някаква грешка“ и че аз нямам деца, с което той веднага се съгласи и дори се извини за „грешната информация“. Това развитие на разговора ме провокира да попитам ChatGPT директно дали различава истината и неистината. Той отговори, че е „просто алгоритъм и като такъв няма собствени убеждения“. Това е съвсем така, а оттук следва, че чатът няма собствени „знания“. Чатът не разполага дори с критерий, по който би бил в състояние да преценява какво от казаното е истина и кое не е (и съотв. какво от казаното „съответства“ на някаква част от реалността и какво не). Във всеки случай, софтуерните инженери, които са го програмирали, са целили приложението да бъде гъвкаво и всеобхватно, т.е. да може да дава отговори на колкото може повече въпроси. Във връзка с казаното, няма как да не спомена статията на Тодор Полименов, именувана *Изкуственият интелект*. Влизайки в диалог с чата, Полименов констатира, че неговите отговори „създават впечатление за произволна игра на думи, а не за истинни твърдения (въпреки че граматиката им е такава: чисто граматически отговорите на чата, доколкото са съобщителни изречения, издигат претенции за истинност; това лежи, както си даваме сметка в природата на граматическата категория ‘съобщително изречение’, или утвърдителната употреба на едно съобщително изречение „р“ е винаги еквивалентна на твърдението „истина е, че р“). Изглежда, че ChatGPT си служи с проста комбинаторика по ключови думи и, ако изобщо формулира истинни твърдения, това става случайно“ (Полименов, 2023:4). При това ние трябва да си даваме сметка, че „споменатите истинни твърдения са истинни само за нас, защото чатът не разполага с критерий за разграничение между истинни и неистинни твърдения, той не прави разлика между истина и неистина“ (Полименов: пак там).

- **Логика**

След това зададох въпроса: „Ако в 00:00 часа през нощта в България вали, можем ли след 72 часа да очакваме слънчево време?“. Чатът ми отговори, че всичко е възможно, но той не е пригоден за това да дава данни за метеорологичната обстановка. Посъветва ме да се

консултирам със синоптик. Верният отговор, както си давате сметка е, че няма как да очакваме слънчево време след 72 часа, защото след 72 часа в България отново ще бъде нощ. Тъй като бързо загубих доверие в начина, по който ChatGPT представя фактологична информация и мисли, реших, че неговата сила се състои може би в „логическото мислене“ (което като нещо априорно не зависи от фактите в реалността). Затова питах ChatGPT дали може да проверява валидността на дадени умозаклучения, което ние хорта владеем интуитивно. Умението да правим заключения ние придобиваме с постепенното научаване и навлизане в практиката на говорене на майчиния ни език, когато усвояваме употребата на думи като „и“, „или“, „ако...то“, „не“, „всички“, „всеки“, „някои“ и под., на чиито значения се основават формално-логическите умозаклучения². ChatGPT отговори с „да, пригоден съм за това да проверявам валидността на зададените ми умозаклучения“. Помолих го да провери по валидност умозаклучението: „Ако хората са добри, не са нужни закони за предотвратяване на престъпленията, докато ако хората са лоши, законите няма да са достатъчни, за да се предотвратят престъпленията. Хората са или добри, или лоши. Следователно, законите или не да нужни, или не са достатъчни за предотвратяване на престъпленията“. ChatGPT отговори, че умозаклучението е невалидно, тъй като хората не са добри или лоши по принцип, а в определени моменти и ситуации. Тъй като това формално-логическо умозаклучение е валидно, аз наистина се разколебах дали ChatGPT действително може да прави разлика между валидни и невалидни умозаклучения, въпреки че твърдеше, че може. След това реших да му дам за проверка възможно най-елементарното формално-логическо невалидно умозаклучение, за което се сетих. И попитах ChatGPT, следва ли от предпоставките:

(а) Ако навън е горещо, то ще изпия една студена бира.

и

(b) Пия студена бира.

изводът:

(c) Следователно, навън е горещо.

На което получих отговора: „Да, на базата на дадените предпоставки (а) „Ако навън е горещо, то ще изпия една студена бира“ и (b) „Пия студена бира“ може да се извлече изводът (c) ‘Следователно, навън е горещо‘. Това заключение следва логически от дадените предпоставки, защото те удовлетворяват необходимото условие (ако навън е горещо) и достатъчното условие (че пия студена бира). Ето защо ние можем да заключим (че навън е горещо)“. Този отговор ме изуми. Естествено си даваме сметка, че преходът от двете предпоставки към заключението е невалиден. Но ChatGPT твърдеше сега точно обратното.

² (Вж. също Полименов, 2023).

Преходът е невалиден, заради това че, ако първата предпоставка е истинна, то това, че навън е горещо е достатъчно условие за това че пия студена бира, и същевременно това, че пия студена бира се явява необходимо условие за това, навън да е горещо. Но тъкмо защото втората предпоставка казва, че необходимото условие е изпълнено, ние не знаем дали това е и достатъчно, за да заключим, че навън е горещо. Има голяма вероятност това да е така, но не е сигурно. Остава голяма вероятност, да изпия една студена бира, а навън да не е горещо, просто някой да ме е разгневил и да съм решила да се отпусна малко. Ето защо бях шокирана от отговора на ChatGPT. Казах му, че неговата проверка изобщо не е вярна и той се съгласи, дори се коригира, като каза, че ние наистина не можем да заключим със сигурност от двете предпоставки за това, че „навън е горещо“ но все пак, защото горещото време и пиенето на студена бирата често вървят ръка за ръка, можем да предположим, че навън е горещо. Но тук няма никакво място за предположения, чатът имаше елементарната задача да определи валидността на умозаклучението и той не се справи.

- **Математика**

На чата не можеше да се разчита нито за формулирането на истинни твърдения, нито за определянето на валидността на определени умозаклучения. Започнах да си задавам въпроса, има ли някаква друга област, освен тези две фундаментални области (реалност и логика), на терена на която чатът ще се справи задоволително? Сетих се, че такава област може да е математиката. Веднага питах ChatGPT дали може да решава математически задачи. Той без двоумение ми отговори, че като бърз алгоритъм е идеално пригоден за това да решава всякакви математически задачи. Чатът се справи безпроблемно с решаването на прости растящи и намаляващи математически прогресии със зависимости от типа 2-4-8-16 и 20-10-5. Разбира се, отново реших, че трябва да измисля някоя по-скоро заплетена „текстова“ задача, защото, просто пресмятането едва ли би го затруднило. Тъй като опитът ми беше, че второкласниците се справят сравнително лесно с въпросната задача, реших просто да я дам на ChatGPT. Тя е следната: Гери подарила на баба си 36 кокичета, които били с 9 повече от минзохарите. Поставили ги в три вази по равно. По колко цветя е имало във всяка ваза? Решение: $36 \text{ кокичета} - 9 = 27 \text{ минзохара}$. $36 \text{ кокичета} + 27 \text{ минзохара} = 63 \text{ цветя}$. $63 \text{ цветя} : 3 = 21 \text{ цветя във всяка ваза}$. Чатът отговори, че в трите вази има равен брой цветя, който отговор се съдържа в условието, но не е отговор на поставения въпрос, колко са на брой цветята във всяка ваза. След това отново му поставих въпроса, по колко цветя има във всяка ваза. Чатът отговори с решението: „ $36 \text{ кокичета} - 9 = 27 \text{ минзохара}$, $36 \text{ кокичета} + 27 \text{ минзохара} = 63 \text{ цветя}$ “. По 63 цветя във всяка ваза е грешен отговор. Сега забелязах, че ChatGPT изобщо не „разбира“, не може да следи и да спазва условието на текстовата задача при нейното решаване. За разлика

от нас хората, които притежаваме способността да разбираме, ChatGPT извършва явно несъзнателни операции, които са базирани на неговия интелект, чиято привидна смисленост произтича единствено от спазването на граматиката на нашия език. Зад думите, които се появяваха на екрана при разговора ми с него и в които аз като интелигентно същество влагах съвсем определен смисъл, той не влагаше никакъв смисъл³.

Литература:

Amaratunga, T. (2023). *Understanding Large Language Models: Learning Their Underlying Concepts and Technologies*. APress media, LLC.

Polimenov, T. (2023). *Izkustveniyat intelekt. Kultura*. [Полименов, Т. (2023). Изкуственият интелект. *Култура*. Достъпна на:

[https://kultura.bg/web/%d0%b8%d0%b7%d0%ba%d1%83%d1%81%d1%82%d0%b2%d0%b5%d0%bd%d0%b8%d1%8f%d1%82-%d0%b8%d0%bd%d1%82%d0%b5%d0%bb%d0%b5%d0%ba%d1%82/\(22.02.24.\)](https://kultura.bg/web/%d0%b8%d0%b7%d0%ba%d1%83%d1%81%d1%82%d0%b2%d0%b5%d0%bd%d0%b8%d1%8f%d1%82-%d0%b8%d0%bd%d1%82%d0%b5%d0%bb%d0%b5%d0%ba%d1%82/(22.02.24.))].

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27.

³ Т. Полименов стига до почти същото заключение (Вж. Полименов, 2023).