Adam Karolewski
Smartstock Sp. z o.o., Grupa AIrrival
akarolewski@o2.pl

Sztuczna inteligencja w poszukiwaniu wiedzy

Streszczenie (wygenerowane przez ChatGPT 3.5): Tekst przedstawia rozwój sztucznej inteligencji (SI) i jej wpływ na różne obszary życia, szczególnie skupiając się na obszarze języka. Zaczyna od wczesnych algorytmów analizy danych, przechodzi przez uczenie maszynowe i sieci neuronowe, aż do dużych modeli językowych. Omawia różne etapy uczenia tych modeli, w tym wykorzystanie danych tekstowych i technik generatywnych. Autor analizuje również implikacje rozwoju SI dla przekazywania wiedzy, zarządzania wiedzą naukową oraz możliwe ryzyka i korzyści związane z automatyzacją syntezy tekstu naukowego. Podkreśla także rosnące znaczenie dużych modeli językowych w przetwarzaniu języka naturalnego i wyszukiwaniu informacji. W końcowej części autor wspomina o polskich modelach językowych i potencjalnych korzyściach z ich wykorzystania w zarządzaniu wiedzą naukową.

Słowa kluczowe: sztuczna inteligencja, duże modele językowe, sieci neuronowe, uczenie maszynowe

Sztuczna inteligencja (*artificial intelligence*, AI) pojawiała się stopniowo. Jej rozwiązania były z nami już od dawna – od jakichś 15, może nawet 20 lat. Przynajmniej odkąd algorytmy prostej regresji liniowej zasilały działania jakichkolwiek aplikacji internetowych. Aż nastała era smartfonów i aplikacji mobilnych. Tak, sztuczna inteligencja i uczenie maszynowe są z nami od dawna. Do tej pory po prostu pracowały pod powierzchnią interfejsów i były niezauważalne, a jednak od wielu lat sterują naszym życiem. Tak, sterują. Może częścią naszego życia i może tą mniejszą, a może większą. Może to tylko rozrywka, a może coś więcej. Algorytmizacja życia postępuje i dzisiaj możemy jej dotknąć tak, jak ona dotyka nas. Możemy z nią nawet porozmawiać. Z nią …?

W dawnych czasach mówiło się o analityce zaawansowanej, o *data mining* lub *text mining*. Pojawiły się nowe algorytmy: lasy losowe, wektory nośne i metoda gradientów wzmacnianych. Informatyzacja analityki doprowadziła do automatyzacji, skalowalności i powtarzalności złożonych procesów analitycznych. I tak powstało uczenie maszynowe. Proces uczenia: dane cyfrowe jako horyzont doświadczeń, statystyka jako silnik wnioskowania, tworzenie wiedzy i prognoza – adaptacja do nowej rzeczywistości – kwintesencja uczenia. Stało się.

Potem przyszły sieci neuronowe. Okazało się, że nawet na laptopach można je uruchomić. W końcu dane tabelaryczne są tanie i nie zajmują dużo miejsca. To cyfry ułożone w kolumnach. Nie to, co grafika i wymagania nowoczesnych gier wideo. No właśnie, grafika i super wydajne karty graficzne. Kto ich nie ma – ten „nie jeździ”. I proszę, w małym laptopie gamingowym można „odpalić” całkiem wydajną architekturę sieci i to głęboką na 50 milionów rekordów. Z użyciem karty graficznej proces uczenia modelu sieci neuronowych trwa krócej. Głównym obszarem było rozpoznawanie obrazów i procesowanie języka naturalnego. Ludzkie twarze, co do ich architektury, nie zmieniają się od setek tysięcy lat, więc i reguły są niezmienione. Tak samo z kwadratami, kołami – architektura przestrzeni rządzi się stałymi od milionów lat regułami. Podobnie z językami i mową. Choć się zmieniają i mówi się, że język „żyje”, to jednak dzieje się to powoli. Czasami słowa nabierają nowego znaczenia, detronizując inne, czasem powstają nowe słowa, najczęściej po kolejnej rewolucji.

A reguły gramatyczne zmieniają się jeszcze wolniej, jeśli w ogóle. Zdanie jako równanie, nawet jeśli mamy 100 tysięcy x-ów, to dla sieci neuronowej jest to właśnie równanie. Sieci neuronowe okazały się być bardzo pojemne i radzą sobie z takimi wielkimi równaniami – wystarczy słowa lub ich części zamienić na cyfry. Modele sieci neuronowej to nieskończone macierze wag przypisanych do x-ów. I ta nieskończoność daje im wielką siłę.

Historia modelowania danych za pomocą algorytmów przebiegała w trzech wyraźnych krokach. Do modelowania szukało się idealnej postaci równania: wielomianu, funkcji wykładniczej – takiej postaci, która najlepiej dopasowywała się do danych. Pojawiły się także pakiety statystyczne i implementacje tych algorytmów. Parametry modeli były wyszukiwane automatycznie, iteracyjnie, często heurystycznie. To był duży postęp i duże ułatwienie.

Następnie pojawiły się algorytmy uczenia maszynowego, algorytmy drzewiaste, z ogromną liczbą drzew losowych, tworzących lasy (stąd np. las losowy – *random forest*). Te algorytmy nie dbały o to, czy rzeczywistość jest liniowa czy kwadratowa. A może wykładnicza. A może tu liniowa, a tam logarytmiczna. Łamały przestrzeń liczbową, wrzucając kolejne drzewa. Miało to swoją cenę, jak choćby ryzyko nadmiernego dopasowania się do danych i tworzenie licznych hiperparametrów, ale znosiło potrzebę znajomości natury problemu. Nie nadawały się jednak dobrze do analizy obrazu i języka. Były zbyt płytkie…

Te same słowa łącząc się z innymi słowami, tworzą różne znaczenia. Łuki łącząc się z liniami prostymi, tworzą różne kształty. Nauczenie się takich zależności było możliwe dzięki zastosowaniu wielowarstwowych sieci neuronowych. Wielowarstwowych – stąd głębokich. Powstały różne typy sieci i różne architektury.

I tak oto znaleźliśmy się w epoce sztucznej inteligencji – jednego z wielu dzieci rewolucji informacyjnej. Sztucznej nie tylko dlatego, że niebiałkowej, powstającej poza mózgami istot żywych. Nie tylko dlatego, że powstałej według przepisu miernie naśladującego proces uczenia się człowieka i istotę ludzkiego umysłu. Sztucznej przede wszystkim dlatego, że cel uczenia jest podawany maszynie z zewnątrz. Cały kontekst i opis sytuacji jest tworzony przez architekta, pomysłodawcę, twórcę i jednocześnie zamawiającego. Inteligencja prawdziwa, jaka by nie była, rozwiązuje SWÓJ problem. W tym przypadku inteligencja jest sztuczna, bo rozwiązuje problem postawiony z zewnątrz. I nawet nie chodzi o to, że jest to narzucony problem, ale dlatego, że sama nie jest w stanie stworzyć własnego (no bo jakże?). Ciekawe jest zatem pytanie: jakie problemy rozwiązywałaby inteligencja zamknięta w krzemowych serwerach, gdyby była zdolna do stawiania sobie własnych celów?

Stąd synonimem dla sztucznej inteligencji mogłaby być inteligencja zewnętrzna lub inteligencja wspomagająca.

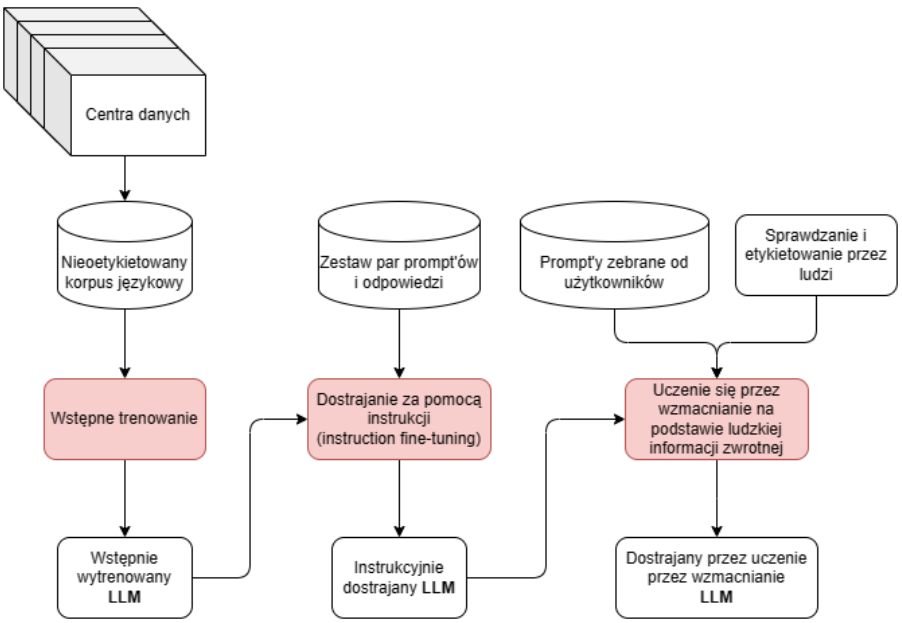
Czy to jest inteligencja? Można się zgodzić, ponieważ w wielu przypadkach dobrze naśladuje proces rozwiązywania problemu i go rozwiązuje. Ma pewną pamięć przeszłości. Posługuje się algorytmami kompresującymi zbiory danych w model rzeczywistości – zbiór poprawnych wniosków.

A czy operowanie językiem wymaga inteligencji? A rozpoznawanie otoczenia jest przejawem inteligencji? Co do języka, nie mam oporów z przyznaniem tego, a co do obrazów – nie jestem pewien.

Jak się okazuje zarówno język, jak i obraz w rozumieniu zastosowań AI dotyczy nadawania abstrakcyjnych znaczeń. Słowo reprezentowane w modelu przez zestaw cyfr wyraża pewien abstrakt w pewnej rzeczywistości, czyli kontekście. Kontekst ten też jest opisany słowami reprezentowanymi przez zestawy cyfr. Podobnie z obrazami. Linie, kolory, kształty w perspektywie tworzą obiekty, którym przypisujemy pewną reprezentację, często wielowarstwową, a warstwy zależą od kontekstów, więc im większa złożoność tych znaczeń tym większej wymaga inteligencji. Czym jest zatem funkcja celu? Trafnym odgadywaniem znaczeń słów i obrazów w określonych kontekstach. A na podstawie tego można dopasować odpowiednie zachowanie i odpowiedź.

I tu dochodzimy do momentu powstania dużych modeli językowych (*large language model*, LLM). Internet zgromadził olbrzymie zasoby słowa pisanego w różnych językach. Zrobiono coś, co wydawało się niemożliwe. Zbudowano olbrzymi model sieci neuronowych, które uczyły się i uczą zależności pomiędzy słowami. Ich celem jest przewidywanie kolejnego, najbardziej prawdopodobnego słowa. I kolejnego słowa. I kolejnych słów... Dzięki odpowiedniej architekturze uczenia (rys. 1) oraz interfejsom uzyskano aplikacje o zdolnościach konwersacyjnych. Ten etap potrafi być bardzo zasobo- i energochłonny. Dla przykładu, trenowanie jednego z dużych modeli językowych z 65 miliardami parametrów wymagało 21 dni pracy na 2048 nowoczesnych procesorach graficznych dedykowanych AI. Każdy z tych procesorów ma 80 GB pamięci operacyjnej (RAM). Szacowany koszt tego trenowania przekroczył 4 miliony dolarów (Zi, El Asri i Prince, 2023).

W pierwszym etapie uczenia tworzymy model bazowy na dużym zbiorze tekstów. Nie muszą być oetykietowane – ważne są zależności między słowami. Stąd, w pewnym sensie, takie modele nie są inteligentne. Nie rozumieją kontekstu, w którym się wypowiadają, ale znają tak wiele kontekstów (wszystkie?) i tak dogłębnie, że przestaje to być istotne z punktu widzenia skuteczności. W drugim etapie model jest douczany za pomocą oetykietowanych par: polecenie-odpowiedź. W ostatnim etapie uczenie jest nadzorowane przez ludzi celem uzyskania odpowiedzi nieszkodliwych dla użytkowników. Ten etap dotyczy choćby ChatGPT.



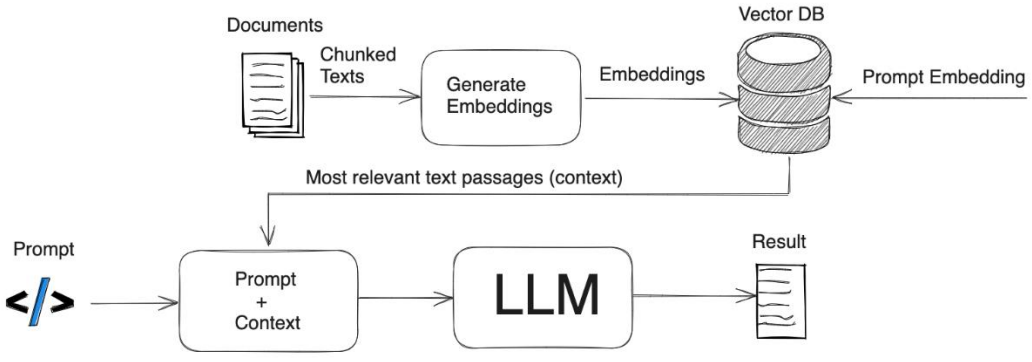
Rys. 1 Fazy uczenia LLM
Źródło: opracowanie własne.

W ciągu ostatnich dwóch lat powstało bardzo dużo aplikacji związanych z tą nową klasą AI. Pojęcie AI ugruntowało się i zepchnęło inne, jak *machine learning*, a nawet *deep learning* w szary kąt. Dzisiaj wszystko jest AI, a AI to ChatGPT i podobne. Tak nie jest. Trzeba pamiętać, że jest to obszar powiązany z językiem i specyficzną architekturą uczenia. AI to bardzo obszerny zbiór różnych algorytmów i zastosowań.

Wraz z aplikacjami bazującymi głównie na języku *text2text* (tekst na tekst) powstały aplikacje przekształcające tekst na grafikę, wideo lub mowę (odpowiednio: *text2image*, *text2video*, *text2speach*). Pojawiło się nowe pojęcie *generative AI* – generatywnej sztucznej inteligencji. I to nie koniec możliwych przekształceń.

Pewne pojęcie typów i różnorodności aplikacji z obszaru *generative AI* mogą dać graficzne zestawienia z firmy CBInsights[1]. Można zauważyć, że jednym z (oczywistych) zastosowań jest wyszukiwanie i zarządzanie wiedzą (*general search/knowledge*). Zastosowanie to idealnie łączy się z tym, czego możemy oczekiwać, myśląc o wyszukiwaniu informacji na dany temat, w danej dziedzinie. Generatywna AI, poza pełnieniem roli klasycznej wyszukiwarki, potrafi dokonywać syntez i streszczeń – w końcu ma wszystkie strony przejrzane i zsyntezowane, skompresowane w postaci tzw. *embeddings* (czyli cyfrowych reprezentacji słów i zdań).

Duże modele językowe mogą służyć (i służą) do wyszukiwania najnowszej wiedzy naukowej. Dobrym przykładem jest istniejąca aplikacja OpenEvidence, której zadaniem jest dostarczanie streszczeń najnowszych artykułów naukowych lekarzom, co umożliwia im zapoznawanie się z bieżącymi odkryciami i ustaleniami ze świata medycyny. Wykorzystywana jest tu technika RAG (*retrieval augmented generation*), która polega na wykorzystaniu bazowego LLM i pokazaniu mu pewnych tekstów (rys. 2). Model językowy pełni wówczas funkcję konwersacyjną, ale bazuje na danych zawartych w pokazanych mu tekstach. Może takie dokumenty streszczać i kompresować do najważniejszych informacji określonych w zapytaniu (*prompt*). Trzeba mieć odpowiednią wiedzę ekspercką, żeby wiedzieć, o co pytać.



Rys. 2. Proces RAG
Źródło: SAFJAN, K. (2023). Understanding Retrieval-Augmented Generation (RAG) empowering LLMs. W: *Krystian Safjan's Blog* [online]. [Dostęp 15.04.2024]. Dostępny w: https://safjan.com/understanding-retrieval-augmented-generation-rag-empowering-llms/.

[1] Wszystkie odesłania do stron internetowych przedstawiają wersję aktualną w dn. 15.04.2024 r.

Za taką praktyką kryje się pewne ryzyko, a może nawet niebezpieczeństwo. Będąc ekspertem w danej dziedzinie, czytamy publikację i znajdujemy inspirację lub rozwiązanie ukryte w jednym lub kilku zdaniach spośród setek, tysięcy oczywistych. Czasami jest to refleksja natury ogólnej wyrażona w filozoficzny sposób. Streszczenie lub synteza może nam tego nigdy nie pokazać. Można określić to zjawisko jako spłaszczenie (uśrednianie) wiedzy poprzez pomijanie niuansów, być może kluczowych i istotnych do dalszego rozwoju danej dziedziny.

Syntezy odzierają teksty z wielu ozdobników, form erudycyjnych, którymi autor uwierzytelnia się jako znawca formy, np. w dziedzinie naukowej. Z drugiej strony można liczyć na pewne ograniczenie dodatków niemerytorycznych wynikających z barokowego stylu wypowiedzi charakterystycznego dla polskiego kręgu kulturowego, jak to wykazywał w swoich badaniach Janusz Hryniewicz (2007).

Naukowcy w trakcie pracy badawczej zapoznają się z literaturą przedmiotu, przeglądając dziesiątki lub setki książek i artykułów naukowych. Destylują z nich potrzebne informacje i syntetyzują je jako punkt wyjścia do swoich badań. Oszczędność czasu i precyzja syntezy może być bardzo cenna z punktu widzenia badacza i całego środowiska. Zmniejszenie wysiłku na tym etapie pomoże przyśpieszyć procesy naukowe i pozwoli skupić uwagę na właściwej pracy badawczej. Z drugiej strony może to doprowadzić do braku czytania artykułów źródłowych. A jeśli prace naukowe trafiałyby do baz wiedzy jako *embeddings*, to autorzy zaczęliby używać nowego języka podobnie jak twórcy stron zaszywają słowa kluczowe. Mógłby powstać nowy język, a prace przestałyby przypominać formę literacką. W tym sensie modele językowe są nieludzkie i mogą nas pozbawić form języka, które spełniają funkcje estetyczne, formalne, komunikacyjne lub rozrywkowe. W tym sensie modele językowe mogą wykształcić swój język, co parafrazując autora *Upadku Hyperiona* (Simmons, 2015) może przypominać formę haiku lub rebusów jako wysokopoziomowe, skompresowane metafory.

Takie rozwiązanie do zarządzania wiedzą jest w zasięgu ręki. Powstają pierwsze polskie duże modele językowe: Bielik oraz Qra. Może powstać PLLuM. Istnieją już bazy artykułów naukowych. Po dokonaniu przekształcenia i umieszczeniu danych w bazie wektorowej otrzymamy nową jakość w przeszukiwaniu źródeł naukowych. Taki proces może okazać się kosztowny, ale wart swojej ceny.

Podsumowując, generatywna sztuczna inteligencja z wykorzystaniem dużych modeli językowych i baz wektorowych może stać się narzędziem służącym do syntezy obecnej wiedzy naukowej. Pomoże to usprawnić (skrócić, polepszyć lub umożliwić) przegląd treści dostępnych w bazach artykułów i innych publikacji naukowych. Stworzone możliwości konwersacyjne pozwalają na prowadzenie rozmowy, a nawet dyskusji z bazą wiedzy, zgodnie z wytyczonym przez pytającego kierunkiem. Można w ten sposób sprawdzać poziom istniejącej wiedzy na dany temat z uwzględnieniem interesujących (np. problematycznych) niuansów. Z drugiej strony trzeba przewidywać zagrożenia wynikające z takiego podejścia, jak np. upraszczanie zagadnień lub usuwanie szczegółów. Dlatego prowadzenie konwersacji wymaga odpowiedniego przygotowania teoretycznego – należy znać pytania i mieć plan takiej rozmowy. Idealne rozwiązanie dla naukowców! I trzeba nauczyć się promptowania…

Bibliografia:

1. HRYNIEWICZ, J. (2007). *Stosunki pracy w polskich organizacjach.* Warszawa: Wydaw. Nauk. Scholar.
2. SIMMONS, D. (2015). *Upadek Hyperiona.* Warszawa: Wydaw. MAG.
3. SAFJAN, K. (2023). Understanding Retrieval-Augmented Generation (RAG) empowering LLMs. W: *Krystian Safjan's Blog* [online]. [Dostęp 15.04.2024]. Dostępny w: https://safjan.com/understanding-retrieval-augmented-generation-rag-empowering-llms/.
4. ZI, W., EL ASRI, L., PRINCE, S. (2023). A High-level Overview of Large Language Models. W: *Borealis AI* [online]. [Dostęp 15.04.2024]. Dostępny w: https://www.borealisai.com/research-blogs/a-high-level-overview-of-large-language-models/.

KAROLEWSKI, A. (2024). Sztuczna inteligencja w poszukiwaniu wiedzy. *Biuletyn EBIB* [online]. Nr 2(213), Sztuczna inteligencja w bibliotekach – rewolucja czy ewolucja. [Dostęp 22.05.2024]. ISSN 1507-7187. Dostępny w: https://ebibojs.pl/index.php/ebib/article/view/909.