



iTWIZ

INFRASTRUKTURA BRANŻA IT SZTUCZNA INTELIGENCJA

RYNEK POLECANE TEMATY

Chcemy, żeby dzięki naszym danym w Polsce powstawały duże modele językowe



Mikołaj Marszycki • 20 marca 2024 8 minut(y) czytania

Z Sebastianem Kondrackim, Chief Innovation Officer w Deviniti i współtwórcą projektu SpeakLeash, rozmawiamy o: genezie projektu, tym jak jest zbudowany i jak można z niego korzystać, czy stanie się dużym modelem językowym; jak dostraja się i trenuje modele oraz jakie wiążą się z tym wyzwania; infrastrukturze, z której korzysta SpeakLeash; współpracy z Cyfronetem i PLLuM, a także najważniejszych celach i wyzwaniach stojących przed SpeakLeash.





Czym jest SpeakLeash, a może raczej Spichlerz? Jaka jest geneza tego projektu?

Półtora roku temu, przy pomocy LinkedIn, zacząłem prowadzić rozmowy z różnymi osobami z firm wchodzących w skład konsorcjum, które odpowiadało za stworzenie Blooma – modelu językowego będącego europejską odpowiedzią na GPT-3, który przetwarzał różne języki naturalne, ale nie miał wśród nich polskiego. Podczas tych rozmów – m.in. z przedstawicielami Big Science – otrzymywałem wiele ciekawych informacji np. na temat narzędzi do ewidencjonowania danych, sposobów pracy z nimi, licencji, praw autorskich itd. Jednocześnie, jak tylko w środowisku rozeszły się pogłoski, że jest taki szalony Polak, który wypytuje wszystkich o dane, skontaktowali się ze mną przedstawiciele EleutherAI – stowarzyszenia, które stworzyło m.in. open source'owy model GPT-NeoX. Zaproponowali, że jeśli zbiorę 1 TB danych, to być może udostępnią mi GPU, abyśmy razem wytrenowali dedykowany model polski. Uznałem więc, że warto stworzyć miejsce, w którym będziemy gromadzić i ewidencjonować wszelkie polskie dane tekstowe. I tak powstało nieformalne stowarzyszenie open source SpeakLeash, czyli Spichlerz.

W ujęciu strategicznym zależy nam na tym, aby dzięki danym ewidencjonowanym i gromadzonym w Polsce, powstawały duże modele językowe – rozwijane i oferowane na różnych licencjach. Ponadto zależy nam na rozwoju społeczności osób, które zajmują się LLM-ami i wspólnie wymieniają wiedzę. Obecnie tworzy ją ok. 300 osób. Mamy wśród nich automatyków, robotyków, przedstawicieli i przedstawicielki dużych banków, biznesu, naukowców np. z Polskiej Akademii Nauk, CLARIN-PL, ale i zwykłych hobbystów.

Obecnie rejestrujemy się jako fundacja, bo brak osobowości prawnej utrudnia nam rozmowy z archiwami państwowymi czy z przedstawicielami mediów, które chcą nas wesprzeć danymi, ale potrzebują np. podpisać umowę dotyczącą współpracy.

Chcemy przede wszystkim tworzyć ekosystem danych umożliwiających samodzielną budowę dużych modeli językowych. W tym ekosystemie są już dostępne czyste dane do uczenia modeli bazowych, ale także instrukcje, dialogi, czyli dane etykietowane i nadzorowane. Ponadto tworzymy system benchmarków polskich lub zagranicznych modeli, ale w warunkach polskich.

Nie ma w Polsce jednego scentralizowanego miejsca zbierania otwartych danych?

Danych otwartych w Polsce jest dużo, ale są one rozproszone. Przykładowo, mamy takie projekty jak CLARIN-PL, który dysponuje pokaźnymi zbiorami danych tekstowych. Instytut Podstaw Informatyki Polskiej Akademii Nauk – który m.in. na

Hugging Face publikuje różne zbiory danych. Mamy Korpus Dyskursu Parlamentarnego, tzw. zrzuty (dumps) Wikipedii itd. We wszystkich tych miejscach obowiązuje jednak inny format danych. Warto też przejrzeć licencje. Nam chodziło natomiast o to, żeby udostępnić jeden pakiet danych, w którym będą wszystkie ewidencje zbiorów, jakościowe metryki itp.

O ile jednak z dostępem do danych otwartych jest całkiem nieźle, tak z otwartymi danymi do celów uczenia maszynowego jest bardzo źle. Owszem, istnieje sporo modeli ośrodków naukowych, które udostępniają dane na otwartej licencji, ale bardzo często brakuje im otoczki biznesowej, a nawet dokumentacji użytkowej, administracyjnej, zamiast wyłącznie naukowej. To właśnie chcieliśmy zmienić tworząc Spichlerz.

W kulturystyce są takie pojęcia jak masa i rzeźba. Na początku budowaliśmy w Spichlerzu masę, czyli głównie zajmowaliśmy się danymi, zbieraliśmy je i opisywaliśmy każdy ich zestaw pod kątem licencji. Dziś dysponujemy zbiorem danych o wielkości rzędu 1,5 TB, więc możemy iść już w przysłowiową rzeźbę. Oznacza to, że bardzo dokładnie klasyfikujemy dane pod względem jakościowym, tematycznym, a także zaczynamy trenować i dostrajać duże modele językowe na tych właśnie danych.

Jak „zbudowany” jest Spichlerz i jak można z niego korzystać?

Wzorowaliśmy się na EleutherAI i ich zbiorze danych The Pile – od nich wzięliśmy sposób ewidencjonowania danych, metodę kompresji (Zstandard) czy składowania plików (LM

Dataformat). Dołożyliśmy do tego manifesty, w których dane są bardzo dokładnie opisane np. ile jest zdań, słów, znaków, kategoria, a także informacje do kogo należy licencja i gdzie jest jej główne źródło. Dodatkowo Spichlerz posiada własny dashboard, a także możliwość przejrzenia dowolnych próbek.

Przykładowo, badacz, który zamierza zbudować model językowy, określa, że chce zgromadzone teksty o wysokim poziomie jakości oraz te, które są np. na licencji creative commons, nawet do komercyjnego rozwiązania. Wtedy nasz system filtruje dane i udostępnia taki zbiór. Natomiast jeśli ktoś chciałby wytrenować model do rozwiązań niekomercyjnych, np. naukowych, może skorzystać z pełnej puli danych.

Jak przebiegał rozwój Spichlerza?

W kulturystyce są takie pojęcia jak masa i rzeźba. Na początku budowaliśmy w Spichlerzu masę, czyli głównie zajmowaliśmy się danymi, zbieraliśmy je i opisywaliśmy każdy ich zestaw pod kątem licencji. Dziś dysponujemy zbiorem danych o wielkości rzędu 1,5 TB, więc możemy iść już w przysłowiową rzeźbę. Oznacza to, że bardzo dokładnie klasyfikujemy dane pod względem jakościowym, tematycznym, a także zaczynamy trenować i dostrajać duże modele językowe na tych właśnie danych.

Do trenowania i dostrajania jeszcze wrócimy, ale najpierw pytanie – czy Spichlerz będzie osobnym, dużym modelem językowym? Czy bylibyście w stanie stworzyć np. drugiego Mistrała?

Nie mamy jeszcze takich planów. Chcemy przede wszystkim tworzyć ekosystem danych umożliwiających samodzielną budowę dużych modeli językowych. W tym ekosystemie są już dostępne czyste dane do uczenia modeli bazowych, ale także instrukcje, dialogi, czyli dane etykietowane i nadzorowane.

Ponadto tworzymy system benchmarków polskich lub zagranicznych modeli, ale w warunkach polskich. Innymi słowy, będziemy sprawdzać jak komercyjne modele – takie jak np. Mistral, Zephyr, Falcon lub OpenChat – radzą sobie z polskimi zadaniami.

Kolejną aktywnością jest samo trenowanie modeli, ale głównie po to, aby nasza społeczność wymieniała się wiedzą i danymi – aby powstał pewien framework, który będzie pozwalał lepiej budować modele i skrypty.

Co do pytania o Mistrala – biorąc pod uwagę, że wszyscy na rzecz Spichlerza pracują po godzinach, a w Mistralu na pełen etat i mając jeszcze ogromne dofinansowanie – uważam, że u nas ten stosunek budżetów i zasobów jest dość wysoki. Być może w przyszłości podejmiemy decyzję, żeby stworzyć np. prywatny ośrodek naukowy, wzorowany na francuskim Mistralu. Na razie jednak Spichlerz doskonale się rozwija w modelu open (science, source, data). Chcemy budować tę otwartą społeczność i otwarty dostęp do danych najdłużej jak się da.

Dzięki uprzejmości Cyfronetu możemy nasze dane trenować także na superkomputerze Helios. Robimy na nim bardzo dużo doświadczeń, testów i badań: jak dobierać dane, jak je klasyfikować, aby zaczęły lepiej generować tekst. To właśnie na Heliosie trenowaliśmy architekturę Mistrala, która ma 7 mld parametrów i jak na razie daje najlepsze efekty.

Jak wygląda wspomniane już

dostrajanie? Skąd można brać instrukcje do dostrajania modeli?

Do dostrajania wykorzystujemy architektury i modele bazowe z innych organizacji – trenujemy więc np. modele Mistrala, „bawimy się” architekturą LLaMa 2 lub tworzymy modele bazowe od podstaw. Cały czas skupiamy się też na stworzeniu odpowiednich narzędzi dla naszego ekosystemu.

Z kolei instrukcje do dostrajania najczęściej pozyskujemy z zadań egzaminacyjnych – przykładowo, dla prawników, lekarzy, marketerów czy dziennikarzy – z innych zestawów danych tzw. etykietowanych, a także tworzymy instrukcje syntetyczne. Brakuje nam natomiast – ze względu na brak budżetu – instrukcji tworzonych całkowicie przez ludzi.

Jakie problemy spotykacie przy dostrajaniu modeli?

Główne problemy dotyczą samych danych oraz ich dostępności. Przykładowo, jeśli do trenowania modelu wykorzystamy zbyt małą ilość tekstu, to jego język będzie przypominał nienaturalny i maszynowy. Zatem, im więcej tekstów, tym model zaczyna lepiej przewidywać słowa, a tekst jest czysty i gramatycznie poprawny.

Z kolei, jeśli mamy już model o odpowiedniej wielkości, to przechodzimy do drugiego etapu, czyli dostrajania go do specyficznych instrukcji i specjalistycznej wiedzy. Dzięki temu, model ten zyskuje nie tylko funkcję generowania następnego słowa z dużym prawdopodobieństwem, ale może też przekazywać wiedzę, czy konwersować. Drugi etap, wymaga więc znacznej ingerencji człowieka – danych przygotowanych i porządnie opisanych, a także wiedzy dziedzinowej.

Jak Wasze prace mają się do polskiego modelu PLLuM? Czy PLLuM pobiera od

Was dane?

Wzajemnie się wspieramy. Zaczęliśmy współpracować jeszcze przed ostatecznym kształtem konsorcjum PLLuM z poszczególnymi jego członkami. W momencie, kiedy PLLuM w pełni się już ukonstytuuje, zapewne nawiążemy oficjalną współpracę, zaproponujemy im nasze narzędzia czy benchmarki.

Z PLLuM nie wymieniamy się jeszcze danymi, ale np. z takimi ośrodkami jak CLARIN-PL, IPI PAN czy NASK – które są w konsorcjum firm tworzących PLLuM – współpracowaliśmy już w obszarze danych. Wspieramy ich również w różnych zadaniach. Przykładowo, NASK, który filtruje polski Internet, zajmuje się modelami do wykrywania różnych przemocowych czy wrażliwych tekstów. Dostarczamy im więc zestawy danych np. o hazardzie, przemocy i używkach, a oni wykorzystują je do trenowania swoich modeli. Podobnie rzecz ma się z CLARIN-PL – np. jeśli mamy dane z niezależnych portali publikujących poezję czy literaturę piękną, to je im przekazujemy.

Chcemy stworzyć kilka mniejszych modeli, przystosowanych do języka polskiego. Takich, które nie muszą posiadać sprawności modeli OpenAI, które potrafią prawie wszystko. Chcemy w ten sposób stworzyć modele, które np. banki, jednostki medyczne czy badawcze mogłyby uruchamiać na swojej infrastrukturze, a nie tylko w chmurze prywatnej – czego czasem zabraniają różne regulacje.

Przejdźmy teraz do infrastruktury, na

której pracujecie...

Z infrastrukturą wiążą się dwa aspekty. Pierwszy, to infrastruktura związana z udostępnianiem danych. Ponieważ nie mamy żadnego budżetu, korzystamy komercyjnie z prostych hostingów, które udostępniają nam swoje zasoby – są one jednak mało pojemne. Z pewnością, kiedy staniemy się już fundacją, na pewno nawiążemy współpracę z którymś z dostawców infrastruktury chmurowej i ten komercyjny hosting zamienimy na duże repozytoria w chmurze.

Drugi aspekt to GPU i możliwości obliczeniowe. W tej kwestii zmieniło się już bardzo dużo. Dotychczas korzystaliśmy bowiem z własnych domowych GPU. Nasz pierwszy model bazowy, stworzony na tekstach w języku polskim – bo inne modele, które pokazywały się w Polsce były tworzone głównie na tekstach angielskich – miał zaledwie 1 mld parametrów. Tymczasem literatura mówi, że dobry model, który można zacząć dostrajać, powinien mieć minimum 7 mld parametrów.

W ramach ciekawostki dodam, że ten pierwszy model liczył się przez 40 dni na domowych kartach graficznych, a rachunek za wykorzystany przy tym prąd sięgnął 1,5 tys. zł. (śmiech). Przy czym taki jedno miliardowy model mocno jeszcze w języku polskim bełkocze i nadaje się co najwyżej do tworzenia abstrakcyjnych memów. Dostrajanie go nic nie da. Natomiast fakt, że taki model stworzyliśmy i zebraliśmy początkowo 1 TB danych, uwiarygodnił nasze działania jako społeczności AI. Wówczas nawiązaliśmy współpracę z Cyfronetem, który zaproponował wspólne badania oraz udostępnił nam swoją infrastrukturę, a konkretnie superkomputer Atenę i czasami Heliosa.

Jakie możliwości daje Wam taka maszyna jak Atena?

Ogromne – nie tylko na poziomie GPU, ale też CPU. Mogliśmy więc zacząć oczyszczać teksty, klasyfikować je jakościowo,

pozwoić sobie wreszcie na stworzenie narzędzi do deduplikowania tekstów – zdarza się bowiem, że mamy np. ustawę w sześciu wersjach, które różnią się treścią tylko np. na poziomie „i/lub”, a wbrew pozorom są to mało istotne różnice na poziomie budowania modelu bazowego.

Sztuczna inteligencja będzie się w Polsce szybko rozwijać wtedy, gdy wszyscy będą się ze sobą wymieniać wiedzą, zgodnie z ideą słynnego eseju Erica Raymonda pt. The Cathedral & The Bazaar. I nie tylko taką na poziomie artykułów naukowych – choć to bardzo ważny aspekt – ale też na poziomie wiedzy praktycznej, dyskusji na temat działania różnych narzędzi.

A Helios?

To potężna maszyna obliczeniowa zbudowana na potrzeby polskiego środowiska naukowego, która może również wspierać duże projekty obliczeniowe realizowane przez krajowe przedsiębiorstwa. Posiada ponad 75 000 rdzeni procesorów x86_64 (CPU) ale także największy w Polsce zasób 440 nowoczesnych kart GPU NVIDIA GH200, które są podstawowym narzędziem treningowym dla AI. Całość wsparta jest ultra szybkimi i pojemnymi systemami dyskowymi oraz siecią komunikacyjną.

Dzięki uprzejmości Cyfronetu możemy nasze dane trenować także na tym superkomputerze. Robimy na nim bardzo dużo doświadczeń, testów i badań: jak dobierać dane, jak je klasyfikować, aby zaczęły lepiej generować tekst. To właśnie na Heliosie trenowaliśmy architekturę Mistrała, która ma 7 mld parametrów i jak na razie daje najlepsze efekty.

Czy współpracujecie z Cyfronetem jeszcze w jakimś innym zakresie?

Przygotowujemy również cały system benchmarków polskich zadań. Odświeżamy pewne gotowe już zadania, które wcześniej stworzyło m.in. Allegro (tzw. KLEJ Benchmark). Tworzymy też nowe, nowocześniejsze zadania. Taki system benchmarków na pewno przyda się innym badaczom. Ponadto testujemy też różnego rodzaju tokenizery i trenujemy własny. Wszystko to robimy po to, aby sprawdzić, przy której konfiguracji i w jakiej architekturze, polski model bazowy oraz polskie instrukcje – bo budujemy też duży zestaw danych instrukcji polskich – będą dawały najlepsze wyniki.

Co jest teraz głównym celem Spichlerza?

Chcemy stworzyć kilka mniejszych modeli, przystosowanych do języka polskiego. Takich, które nie muszą posiadać sprawności modeli OpenAI, które potrafią prawie wszystko. Chcemy w ten sposób stworzyć modele, które np. banki, jednostki medyczne czy badawcze mogłyby uruchamiać na swojej infrastrukturze, a nie tylko w chmurze prywatnej – czego czasem zabraniają różne regulacje. Natomiast, jak już będziemy mieć kilka modeli i cały ekosystem narzędzi, wtedy zdecydujemy co dalej.

Pogłębianie dalszej współpracy z Cyfronetem, PLLuMem...

Na pewno. W przestrzeni publicznej padło nawet takie pytanie czy PLLuM jest w ogóle potrzebny? Uważam, że potrzebnych jest jeszcze 5, 6, a może nawet 10 tego typu inicjatyw – chyba, że jedna z 10 razy większym budżetem. Chodzi bowiem o to, żeby one ze sobą wzajemnie współpracowały. Sztuczna inteligencja będzie się w Polsce szybko rozwijać wtedy, gdy wszyscy będą się ze sobą wymieniać wiedzą, zgodnie z ideą słynnego eseju Erica Raymonda pt. The Cathedral & The Bazaar. I nie tylko taką na poziomie artykułów naukowych – choć to

bardzo ważny aspekt – ale też na poziomie wiedzy praktycznej, dyskusji na temat działania różnych narzędzi.

Na razie nie wiemy jeszcze jak to się wszystko potoczy. Być może PLLuM będzie ogromnym modelem, który będzie ogólnodostępny za darmo, ale tylko poprzez API. Jednostki regulowane nie będą mogły z niego skorzystać ze względu na przepisy i być może wtedy sprawdzą się właśnie te mniejsze modele z możliwością odpalenia na własnej infrastrukturze.

Na pewno PLLuM zrobił pierwszy ważny krok – połączył wreszcie ośrodki naukowe. Jeśli zrobimy drugi krok i do tych ośrodków dołączy biznes oraz takie społeczności jak nasza, to będzie już prawdziwa rewolucja.

Co jest obecnie największym wyzwaniem dla Spichlerza?

Powiem tak – mamy dostęp do najlepszych polskich i europejskich superkomputerów oraz GPU. Ogromne infrastrukturalne, superkomputerowe, profesjonalne i naukowe wsparcie Cyfronetu. Mamy też 1,5 TB bardzo dobrze sklasyfikowanych danych. Mamy wreszcie wspaniałą społeczność, za którą stoi wielka wiedza. Można więc powiedzieć, że mamy prawie wszystko oprócz czasu i budżetu (śmiech).

To czego nam teraz potrzeba, to trochę organizacji – musimy się sprofesjonalizować. Obecnie wygląda to bowiem trochę tak, jak na boisku szkolnym podczas meczu – wszyscy chcą grać w ataku (śmiech). Potrzebujemy też jednak bramkarza czy ludzi, którzy zagrają w obronie i będą pełnili pewne funkcje pomocnicze. Tego właśnie nam brakuje, ale powoli zaczynamy się organizować.



Tagi

PLLuM [<https://itwiz.pl/tag/pllum/>]

SpeakLeash [<https://itwiz.pl/tag/speakleash/>]

Cyfronet [<https://itwiz.pl/tag/cyfronet/>]

Big Data [<https://itwiz.pl/tag/big-data/>]

open source [<https://itwiz.pl/tag/open-source/>]

dane [<https://itwiz.pl/tag/dane/>]

LLM [<https://itwiz.pl/tag/llm/>]

duże modele językowe [<https://itwiz.pl/tag/duze-modele-jezykowe/>]

REDAKCJA



Adam Jadczyk

O IT w biznesie pisze od 1994 roku. Specjalizuje się w zagadnieniach związanych z rynkiem IT oraz informatyką w zastosowaniach biznesowych.



Piotr Waszczuk

Związany z rynkiem nowych technologii od 2004 roku. Autor publikacji, wywiadów i opracowań dotyczących trendów i wydarzeń branżowych, a także technologii i rozwiązań IT.



Mikołaj Marszycki

Dziennikarz z kilkunastoletnim doświadczeniem. Autor opracowań i wywiadów dotyczących nowych technologii.



Szymon Augustyniak

Od 2000 roku pisze o burzliwej współpracy IT i biznesu kreśląc także jej szerszy, społeczny, naukowy i gospodarczy kontekst. Od 2009 animuje działalność społeczności CIO i CxO, obecnie w ramach ekosystemu ITwiz.