O pożyczaniu innych światów czyli po co nam polski LLM

Michał Dulemba
Egnyte
Podcaster - nieliniowy.pl
https://www.linkedin.com/in/michal-dulemba/

dr Maria Filipkowska
SpeakLeash
https://www.linkedin.com/in/maria-filipkowska/

dr Dominika Kaczorowska-Spychalska
Centrum Inteligentnych Technologii,
Wydział Zarządzania
https://www.linkedin.com/in/dominika-kaczorowska-spychalska-25916a172/

Sebastian Kondracki
Deviniti sp. z o.o., SpeakLeash
SoDA AI Research Group
https://www.linkedin.com/in/sebastian-kondracki/

Jan Maria Kowalski
Bank Pekao S.A., SpeakLeash
https://www.linkedin.com/in/janmariakowalski/

dr Inez Okulska
NASK Państwowy Instytut Badawczy
Ministerstwo Cyfryzacji
https://www.linkedin.com/in/inezokulska/

Emilia Wiśnios
NASK Państwowy Instytut Badawczy
https://www.linkedin.com/in/emilia-wisnios/

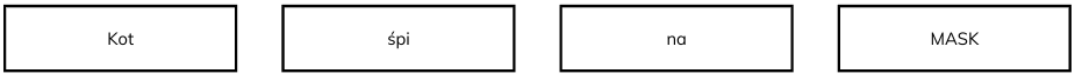
W ostatnich miesiącach byliśmy świadkami wyścigu gigantów w biegu o najlepszy duży model językowy (ang. LLM). Sęk, w tym, że obecnie modele te są albo zamknięte – tak jak ChatGPT (Microsoft) czy Bard (Google) – i nie można ich ani modyfikować, ani w pełni zrozumieć czy kontrolować, albo, nawet jeśli ich źródła są publicznie dostępne – tak jak w przypadku Llama2 (Meta) – to nie obsługują języka polskiego (w tym modelu udział polskiego wynosi 0,09%, co wystarcza, by zrozumieć polecenie, ale już niekoniecznie napisać sensowną odpowiedź). Zbiory treningowe użyte do tworzenia tych modeli też nie są publicznie dostępne ani choćby jawnie opisane. Ciekawą odpowiedzią na działania big-techów i pięknym przykładem współpracy świata nauki i biznesu jest Bloom – w pełni transparentny i otwarty model (zarówno pod kątem algorytmu, jak i danych). Niestety, pozostaje znów to samo „ale": wśród wielu języków, które obsługuje, nie ma języka polskiego. Otwarty, dostępny i całkowicie transparentny duży model językowy dla języka polskiego to zatem luka, którą należy wypełnić. I to nie rękoma jednego giganta, a wspólnie, kolektywnie, łącząc siły najlepszych ze świata nauki, biznesu i administracji publicznej.

Ostatnio zaczynają pojawiać się regulacje w obszarze m.in. generatywnej AI. 14 czerwca 2023 roku Parlament Europejski przyjął tak zwany AI ACT, będący pierwszym europejskim aktem prawnym dotyczącym sztucznej inteligencji. Jak można było przewidzieć, zawiera on elementy związane z przejrzystością modeli generatywnej sztucznej inteligencji, zwłaszcza w kontekście praw własności zestawów danych służących do ich trenowania. W związku z tym wiele firm może zacząć obawiać się wykorzystywania na skalę biznesową dużych modeli językowych ze względu na brak transparencji w zakresie danych użytych do ich uczenia.

Stąd wynika konieczność rozpoczęcia dialogu w obszarze szeroko pojmowanej sztucznej inteligencji. Kwestia, którą powinniśmy poddać dyskusji, dotyczy ewentualnego połączenia potencjału naukowego oraz biznesowego w celu stworzenia w pełni polskiego, otwartego i transparentnego modelu języka na dużą skalę (LLM). Powinien on służyć nie tylko naukowcom i przedsiębiorcom, ale także, co nie mniej ważne, polskiemu społeczeństwu, jako ostatecznemu odbiorcy innowacji wynikłych z wykorzystania takiego modelu.

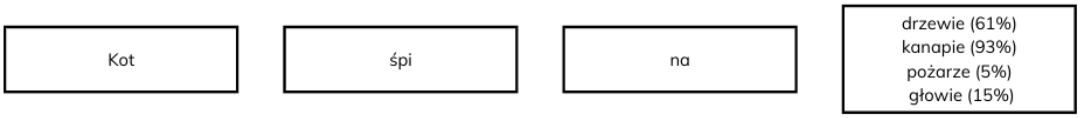
Czym jest LLM?

Model językowy (LM) to model, którego celem jest nauczenie się struktury języka. Trening takiego modelu odbywa się poprzez uczenie się przewidywania zamaskowanego słowa (tak zwany paradygmat MLM, czyli Masked Language Model). Przeanalizujmy krótki przykład, żeby zrozumieć ideę takiego modelu. Załóżmy, że nasz model dostał zdanie *Kot śpi na MASK*. Jego zadaniem jest przewidzenie, jakie słowo powinno znajdować się pod maską.



Oczywiście istnieje wiele słów, które pasowałyby do tego zdania, ale niektóre są bardziej prawdopodobne niż inne. Właśnie tego prawdopodobieństwa uczy się model podczas

treningu.



Im więcej danych widział model, tym lepszy staje się w przewidywaniu słów. Tak właśnie narodziła się idea **dużych modeli językowych (LLM)**. Taką skrzynkę, która nauczyła się danego języka (lub języków w przypadku modeli wielojęzycznych), możemy potem douczać pod konkretne zadania, np. klasyfikację tekstu czy odpowiadanie na pytania. Obecnie największą popularność zdobyły tak zwane konwersacyjne modele językowe, do których wytrenowania potrzeba zbioru instrukcji, czyli polecenia (promptu) i odpowiedzi na nie.

Dla języka polskiego dostępnych jest kilka modeli wytrenowanych w paradygmacie MLM, jak RoBERTa[1], HerBERT[2], papuGaPT[3], Polbert[4], mT5[5]. Istnieje też polski model konwersacyjny oparty na wielojęzycznym modelu i dostrojony na polskich instrukcjach[6].

Nie istnieje natomiast prawdziwie duży model nauczony konkretnie na języku polskim i dotrenowany na zbiorze polskich instrukcji.

Doskonałym globalnym przykładem współpracy między światem nauki a sektorem biznesu jest projekt **BigScience** i jego model **Bloom**, mający 176 miliardów parametrów. Model ten został uruchomiony w lipcu 2022 roku jako europejska odpowiedź na model **GPT-3**. W ciągu zaledwie 117 dni ponad 1000 naukowców z 250 zaangażowanych w projekt instytucji z 70 krajów opracowało najbardziej efektywną alternatywę dla narzędzia **OpenAI**.

Bloom działa w 46 językach naturalnych i 13 językach programowania. W realizację projektu najbardziej zaangażowane były francuskie instytuty badawcze, takie jak **CNRS** (fr. Centre National de la Recherche Scientifique, czyli Narodowe Centrum Badań Naukowych) oraz **GENCI** (Grand Equipement National de Calcul Intensif – Główny Krajowy Ośrodek Superkomputerowy), a także amerykańska firma **Hugging Face**. Wkrótce po rozpoczęciu prac nad projektem dołączyły do niego różne inne uczelnie oraz firmy, a także wolontariusze rekrutowani podczas hackathonów programistycznych. Jedyne, czego w nim brakuje, to możliwości pracy w języku polskim. Ten istotny dla Polski niedobór zainspirował założycieli projektu **SpeakLeash**, znanego również pod nazwą **Spichlerz**. Jest to polski projekt w ramach otwartej nauki (open science), którego celem jest gromadzenie zasobów tekstowych w języku polskim i tworzenie na ich bazie obszernych korpusów językowych. Korpusy te mają za zadanie przygotować grunt pod trenowanie dużych modeli językowych, podobnych do modelu Bloom.

---

[1] https://huggingface.co/sdadas/polish-roberta-base-v2
[2] https://huggingface.co/allegro/herbert-base-cased
[3] https://huggingface.co/flax-community/papuGaPT2
[4] https://github.com/kldarek/polbert
[5] https://huggingface.co/google/mt5-base
[6] https://voicelab.ai/

Zarówno **GRAI** (Grupa Robocza ds. Sztucznej Inteligencji), jak i **OPI-PIB** oraz **Izba Łukasiewicz** aktywnie wspierają tworzenie i rozwój projektu Spichlerz. **IPI PAN** dostarcza dane do tego przedsięwzięcia. Dodatkowo, od ostatniego **Data Science Summit** w Warszawie, projekt **Spichlerz** rozpoczął współpracę z **NASK PIB**, m.in. w kontekście wykorzystania pakietu **StyloMetrix NASK** do jakościowej oceny dokumentów tekstowych oraz pozyskania kolejnych obszernych zbiorów danych.

Polish Big Science ma spory potencjał, aby w istotny sposób zwiększyć zakres i dynamikę realnej współpracy między biznesem a światem nauki, generując znaczne korzyści dla każdej ze stron. Nowe, bardziej efektywne modele biznesowe, nowe źródła przewagi konkurencyjnej i większe zdolności do proaktywnej reakcji na światowe trendy to tylko jedne z wielu potencjalnych korzyści. Nie bez znaczenia jest także wzrost rozpoznawalności polskiej nauki na arenie międzynarodowej.

Być może to właśnie dzięki polskiemu LLM mamy szansę znaleźć się w czołówce pierwszych 25% gospodarek, które wytwarzają innowacyjne rozwiązania AI, zgodnie z zapisami *„Polityki dla rozwoju sztucznej inteligencji w Polsce od roku 2020” .*

dr Dominika Kaczorowska-Spychalska
Dyrektor Centrum Inteligentnych Technologii
Wydziału Zarządzania Uniwersytetu Łódzkiego

Projekt Spichlerz wyróżnia się również tym, że łączy w sobie przedstawicieli sektora biznesowego na co najmniej dwa sposoby. Z jednej strony, założyciele i uczestnicy projektu są zatrudnieni w dużych polskich firmach technologicznych. Z drugiej strony, zaczynają dołączać do nich firmy, które pragną wspierać projekt jako sponsorzy. Co jest istotne, zarówno uczestnicy, jak i sponsorzy są świadomi potrzeb w obszarze dużych modeli językowych dla biznesu.

Jeśli do tego wszystkiego dołączymy projekt **CLARIN-PL** (Common Language Resources & Technology Infrastructure) czyli fundament polskiego NLP, oddział ogólnoeuropejskiej infrastruktury naukowej, która umożliwia badaczom z dziedziny nauk humanistycznych i społecznych wygodną pracę z bardzo dużymi zbiorami tekstów, to czy nie jest to początek **Polish Big Science**? Łączy nas jeden cel: stworzenie dużego polskiego modelu językowego, który będzie otwarty, dostępny i transparentny. W tym artykule postaramy się pokazać, dlaczego to tak ważne, i ocenić, czy jest to możliwe.

Czy polski LLM na pewno jest potrzebny?

Czy rzeczywiście dla języka polskiego jest potrzebny duży model językowy (LLM) na otwartej licencji? Wokół LLM-ów rozwija się sieć firm dostarczających innowacyjne narzędzia. Zaczynając od generatorów opisów produktów, przez stylizatory i kreatory artykułów, postów na blogi czy e-maili biznesowych, a kończąc na chatbotach. Niemniej jednak polscy przedsiębiorcy obecnie mogą korzystać wyłącznie z rozwiązań płatnych, takich jak te oferowane przez **OpenAI**.

Niestety, nawet dostrajanie modeli do indywidualnych potrzeb – proces, który i tak wymaga znacznych nakładów czasu i wiedzy – jest dodatkowo obciążony opłatami, co zauważalnie ogranicza możliwości konkurencyjne polskich start-upów. O ile łatwiej byłoby pobrać otwarty model z językiem polskim i dostosować go odpowiednio na własnej infrastrukturze!

Na portalu **Hugging Face** znajduje się wiele modeli językowych na otwartych licencjach, jednak wśród nich nadal brakuje języka polskiego. Oczywiście czasami zdarzają się modele, które mają kilkuprocentowy udział polskich tekstów w procesie trenowania lub są dostrojone do języka polskiego, jednak pod kątem jakości i zakresu realizowanych zadań NLP są to bardzo przeciętne modele. Wobec tego priorytetem powinno być wyrównanie szans w wyścigu dla polskich przedsiębiorców poprzez umożliwienie dostępu do modeli z językiem polskim na bezpłatnej licencji.

Inną grupą przedsiębiorstw, które wyraźnie wykazują potrzebę posiadania polskiego dużego modelu językowego (LLM), są podmioty regulowane, takie jak banki, ubezpieczalnie czy np. służba zdrowia. Te instytucje również widzą potencjał w generatywnej sztucznej inteligencji, jednak wymagają pełnej kontroli nad nią.

Przede wszystkim ważne jest dla nich to, aby model działał na lokalnej infrastrukturze, bez konieczności wysyłania danych do chmury, szczególnie zagranicznej. Po drugie, chcą mieć możliwość czuwania nad budową samego modelu: oznacza to, że dane, na podstawie których model był trenowany, architektura modelu i jego parametry powinny być otwarte, bo to zapewni pełną ich transparentność. Taki model musi być w pełni zgodny z wszelkimi aktami prawnymi dotyczącymi danych i sztucznej inteligencji obowiązującymi w Unii Europejskiej.

	Generatywna Sztuczna Inteligencja znajduje wiele zastosowań w bankowości, rewolucjonizuje procesy obsługi klienta, pozwala lepiej zrozumieć jego oczekiwania i tworzyć oferty dopasowane do jego potrzeb. Dlatego polski Duży Model Językowy, możliwy do uruchomienia we własnej infrastrukturze, zapewniający bezpieczeństwo, minimalizujący ryzyka związane z przetwarzaniem danych czy złamaniem warunków licencyjnych, a do tego możliwy do dostrojenia do własnych potrzeb, to będzie prawdziwy przełom. Przełom w automatyzacji procesów i budowie rozwiązań z

	wykorzystaniem AI w zakresie przetwarzania języka naturalnego nie tylko w tak bardzo regulowanym sektorze, jak bankowość. Jan Maria Kowalski Dyrektor Biura Nowych Technologii Bank Pekao

Ostatni aspekt, równie istotny jak pozostałe, ma wyraźnie naukowy charakter. Wskazujemy tu na konieczność prowadzenia badań w obszarze dużych modeli językowych, aktywnego udziału naszych naukowców na tym polu, a także tworzenia dedykowanych wykładów i specjalizacji. Te działania zostały dokładnie opisane w „Polityce dla rozwoju sztucznej inteligencji w Polsce od roku 2020". Dokument ten podkreśla potrzebę wspierania polskiego środowiska naukowego i badawczego w projektowaniu interdyscyplinarnych wyzwań lub rozwiązań w obszarze AI, z uwzględnieniem nauk humanistycznych i społecznych. Dodatkowo kładzie nacisk na tworzenie katedr AI, kształcenie doktorantów, przyznawanie grantów dla badaczy oraz inne działania mające na celu przygotowanie ekspertów zdolnych do tworzenia rozwiązań AI. Wszystko to powinno być realizowane z uwzględnieniem etycznych i bezpiecznych ram wykorzystania tej technologii, a przy tym przynosić korzyści gospodarce i zwiększać dobrobyt obywateli.[7]

Warto podkreślić, że mamy niezwykle utalentowanych polskich naukowców, zarówno w kraju, jak i za granicą. Przyjrzyjmy się bliżej kilku najważniejszym badaczom w obszarze dużych modeli językowych.

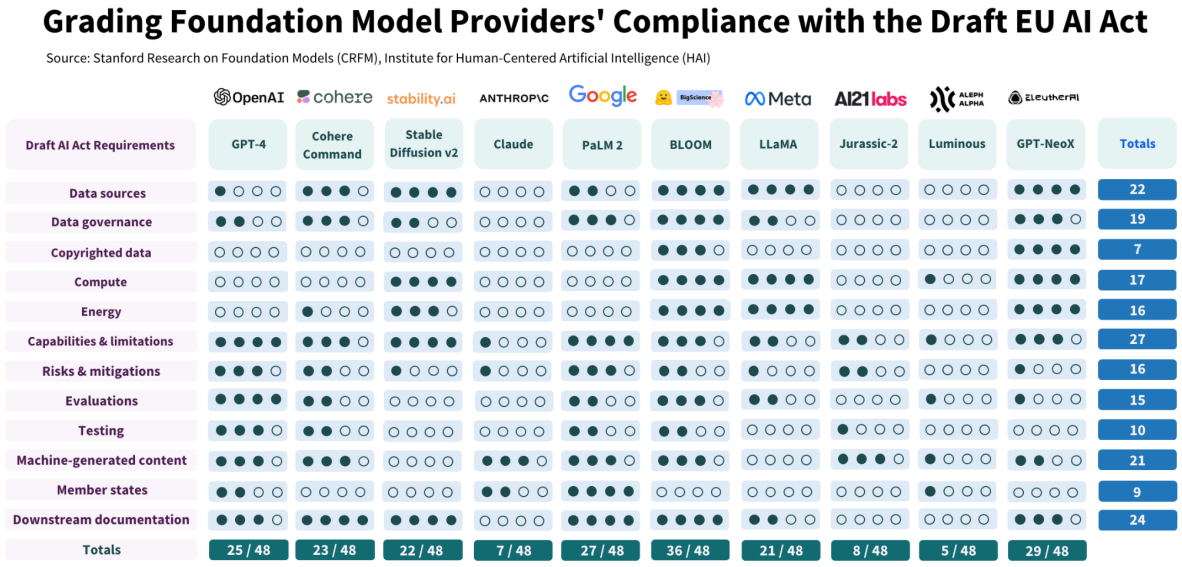
Rewolucja, która zapoczątkowała rozwój dużych modeli językowych, rozpoczęła się od publikacji artykułu pt. ***Attention Is All You Need***. Artykuł ten wprowadził architekturę typu **Transformer**. Wśród autorów tego przełomowego dokumentu znalazł się **Łukasz Kaiser**, pochodzący z **Wrocławia**, obecnie pracujący w firmie **OpenAI**. Wielu czytelników nie potrzebuje przedstawienia tej renomowanej firmy, podobnie jak jednego z jej współzałożycieli, **Wojciecha Zaremby**. Jest tam również zatrudnionych wielu innych inżynierów urodzonych w Polsce, takich jak **Jakub Pachocki** czy **Szymon Sidor**. Ten ostatni miał niedawno zauważyć, że ***„w początkowej fazie istnienia OpenAI, spośród około 50 zatrudnionych osób, aż 10 pochodziło z Polski!"***. Oznacza to, że w Polsce jesteśmy w stanie kształcić inżynierów oraz badaczy o wysokim poziomie kompetencji.

Czas powtórzyć sukces BigScience

Wytrenowanie dużego modelu językowego wymaga trzech ważnych elementów tj. danych, wiedzy i mocy obliczeniowej. My do tej trójcy chcielibyśmy dodać jeszcze dobrze zarządzany kapitał społeczny. Dlaczego ten element jest ważny, a wręcz całkowicie niezbędny, pokazał

[7] https://www.gov.pl/web/ai/polityka-dla-rozwoju-sztucznej-inteligencji-w-polsce-od-roku-2020 [dostęp: 18.07.2023]

opisywany wyżej projekt **BigScience** i jego model **Bloom**. Po ukazaniu się pierwszej europejskiej regulacji dotyczącej AI (tzw. **AI ACT**) naukowcy przyjrzeli się dużym modelom językowym pod względem zgodności z regulacją i to właśnie **Bloom** uzyskał najwyższą możliwą notę (tj. 36 na 48 punktów).



Grading Foundation Model Providers' Compliance with the Draft EU AI Act

Source: Stanford Research on Foundation Models (CRFM), Institute for Human-Centered Artificial Intelligence (HAI)

Draft AI Act Requirements	GPT-4	Cohere Command	Stable Diffusion v2	Claude	PaLM 2	BLOOM	LLaMA	Jurassic-2	Luminous	GPT-NeoX	Totals
Data sources	●○○○	●●●○	●●●●	○○○○	●●○○	●●●●	●●●●	○○○○	○○○○	●●●●	22
Data governance	●●○○	●●●○	●●○○	○○○○	●●●○	●●●●	●●○○	○○○○	○○○○	●●●○	19
Copyrighted data	○○○○	○○○○	○○○○	○○○○	○○○○	●●●○	○○○○	○○○○	○○○○	●●●●	7
Compute	○○○○	○○○○	●●●●	○○○○	○○○○	●●●●	●●●●	○○○○	●○○○	●●●●	17
Energy	○○○○	●○○○	●●●○	○○○○	○○○○	●●●●	●●●●	○○○○	○○○○	●●●●	16
Capabilities & limitations	●●●●	●●●○	●●●●	●○○○	●●●●	●●●○	●●○○	●●○○	●○○○	●●●○	27
Risks & mitigations	●●●○	●●○○	●○○○	●○○○	●●●○	●●○○	●○○○	●●○○	○○○○	●○○○	16
Evaluations	●●●●	●●○○	○○○○	○○○○	●●○○	●●●○	●●○○	○○○○	●○○○	●○○○	15
Testing	●●●○	●●○○	○○○○	○○○○	●●○○	●●○○	○○○○	●○○○	○○○○	○○○○	10
Machine-generated content	●●●○	●●●○	○○○○	●●●○	●●●○	●●●○	○○○○	●●●○	●○○○	●●○○	21
Member states	●●○○	○○○○	○○○○	●●○○	●●●●	○○○○	○○○○	○○○○	●○○○	○○○○	9
Downstream documentation	●●●○	●●●●	●●●●	○○○○	●●●●	●●●●	●●○○	○○○○	○○○○	●●●○	24
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48	29 / 48	

Źródło: https://crfm.stanford.edu/2023/06/15/eu-ai-act.html [dostęp: 19.07.2023]

Model Bloom otrzymał również wysokie oceny pod względem precyzji, co potwierdza, że jest to narzędzie nie tylko zgodne z regulacjami, lecz również efektywne. Mimo to, brak wsparcia dla języka polskiego skłania do przytoczenia słów noblistki, Toni Morrison, która niegdyś stwierdziła: **„If you find a book you really want to read but it hasn't been written yet, then you must write it"**[8].

Korzystając z doświadczeń **EleutherAI** (projektu open science, twórcy modelu **GTP-NEOX**), **BigScience,** polskiego **SpeakLeasha,** a także polskich ośrodków naukowych i akademickich, warto połączyć siły w celu stworzenia Polish BigScience. Przejdźmy do szczegółów wspomnianych trzech głównych filarów tego przedsięwzięcia.

Dane

1. **CLARIN-PL** (Common Language Resources & Technology Infrastructure) to polski oddział ogólnoeuropejskiej infrastruktury naukowej, która umożliwia badaczom z dziedziny nauk humanistycznych i społecznych wygodną pracę z bardzo dużymi zbiorami tekstów. Nie tylko bada duże zbiory, ale je przede wszystkim posiada.
2. **SpeakLeash a.k.a Spichlerz** to projekt open science, wzorowany na zestawach danych wykorzystywanych do uczenia modeli **GPT-NeoX** (**The Pile**), **Bloom** (**The**

[8] „Jeśli książka, którą bardzo chciałbyś przeczytać, jeszcze nie istnieje, to ją napisz" – tłum. IO, przedruk na tej stronie: https://www.newspapers.com/article/21863475/tonimorrison/ [dostęp: 20.07.2023]

Catalogue/Roots) oraz **Falcon** (**The RefinedWeb**), ale obecnie dużo bardziej zaawansowany. Oprócz surowych danych do każdego dokumentu tekstowego dodawane są metryki (językowe i jakościowe), każdy dokument jest klasyfikowany, dokładnie opisywane są też prawa do zbioru. **Spichlerz** posiada zaawansowany dashboard, w którym można przeglądać status projektu i metryki danych, a także pakiet programistyczny do natychmiastowego użycia. Na ten moment zbiór danych obejmuje już ponad 350 GB. Jeśli do tego dodamy polskie prace naukowe (wkład uczelni), archiwalne księgozbiory, informację publiczną czy też transkrypcje telewizyjne, to projekt polskiego LLM w ciągu tygodni będzie miał największy na świecie zestaw zróżnicowanych tekstów w jednym języku. Należy podkreślić, że przekazywanie danych przez uczelnie, biblioteki czy instytucje odbywałoby się na zasadzie dobrowolności, z zachowaniem wysokiego poziomu precyzji w opisie praw i autorstwa poszczególnych prac czy publikacji.

3. **Aya Project (Cohere)**[9] to otwarty projekt naukowy, mający na celu wyrównanie dysproporcji w liczebności instrukcji do dostrajania modeli językowych w różnych językach. Wśród egzotycznych dla nas języków, takich jak shona, marathi czy tamilski, znalazł się również język polski. Zbiór instrukcji tworzony jest przez ochotników posługujących się danym językiem i będzie udostępniony publicznie. W momencie pisania artykułu Polska mierzy się z kolejnym wyzwaniem, czyli zebraniem 1000 instrukcji. Gorąco zachęcamy do dołączenia się do tej inicjatywy, która znacznie wzbogaci dostępne zbiory danych dla naszego rodzimego języka.

	Polski nie stać na to, żeby nie mieć własnych modeli LLM. Widać, że stają się one kluczowe dla powstawania nowych firm, usług czy nawet całego sposobu interakcji człowieka z elektroniką. Druga sprawa – od wielu lat środowisko open source zapewnia zbliżonej jakości narzędzia dla tych, którzy z jakiegoś powodu nie mogą korzystać z wersji komercyjnej – np. LibreOffice vs. MS Office. Podobnie powinno być ze zbiorami danych czy modelami LLM. Możliwość wyboru zawsze wpływa pozytywnie na cały rynek. Cieszę się, że *Nieliniowy* pomaga się spotkać ludziom branży Machine Learning w Polsce. Michał Dulemba Senior Machine Learning Engineer (Egnyte) Podcaster (Nieliniowy.pl)

Moc obliczeniowa

[9] https://sites.google.com/cohere.com/aya-en/home

Potrzeba mocy obliczeniowej zależy od liczby parametrów, długości okna kontekstowego oraz szczegółów architektury modelu. Wartości te dobiera się m.in. z uwzględnieniem wielkości zbioru treningowego. W przypadku modelu Bloom proces treningowy trwał około 3,5 miesiąca i wykorzystano do tego 384 karty GPU A100 (80 GB)[10]. Część kosztów pokrył Hugging Face, dodatkowo w projekcie uczestniczyły francuskie ośrodki obliczeniowe wraz z superkomputerem Jean Zay.

Aby pokazać skalę kosztów: jedna karta NVIDIA® A100 GPU (80 GB) kosztuje około 15 000 USD. Oczywiście nie trzeba od razu kupować takiej liczby kart, można skorzystać z rozwiązań chmurowych – np. model Falcon był trenowany na AWS SageMaker.[11] Przed nami jeszcze czas na podjęcie decyzji dotyczącej optymalnej infrastruktury obliczeniowej. Wśród dostępnych opcji, oprócz korzystania z komercyjnych chmur obliczeniowych wyposażonych w jednostki GPU, warto rozważyć możliwość skorzystania z usług naukowych centrów obliczeniowych.

Na przykład superkomputer Athena zainstalowany w 2021 roku w Cyfronecie (Akademickie Centrum Komputerowe Cyfronet AGH) dostarcza polskiemu środowisku naukowemu i gospodarce najnowocześniejszych zasobów obliczeniowych o dużej mocy. Konfiguracja Atheny obejmuje: 48 serwerów z procesorami AMD EPYC i 1 TB pamięci RAM (w sumie 6144 rdzenie obliczeniowe CPU) oraz 384 karty GPU NVIDIA A100 (wow!). Athena osiąga teoretyczną moc obliczeniową ponad 7,7 PetaFlopsów (7709 TeraFlopsów), co zapewniło maszynie zajęcie 105 miejsca na liście TOP500 superkomputerów na świecie oraz sprawia, że jest to obecnie najszybszy superkomputer w Polsce.[12] Kolejka obliczeniowa do Atheny jest zapewne długa, jednak może warto już zapisać się na listę oczekujących.

Sztuczna inteligencja, a w szczególności duże modele językowe (LLM) to jeden z najbardziej ekscytujących kierunków rozwoju technologii na świecie. Technologia ta staje się fundamentem transformacji przedsiębiorstw oraz tworzenia nowych firm, wpływając na ich konkurencyjność.

Polska ma szansę stać się nie tylko konsumentem, ale także producentem tej technologii. Wymaga to dużych nakładów kapitałowych oraz udostępnienia infrastruktury obliczeniowej jednostkom akademickim oraz przedsiębiorstwom z ambicjami rozwoju LLM. Wytrenowanie polskiego LLM jest przykładem takiej ambicji, która zasługuje na zdecydowane poparcie.

Dariusz Kłeczek
Machine Learning Engineer
Weights & Biases

[10] S. Bekman (2022) The Technology Behind BLOOM Training, Hugging Face, https://huggingface.co/blog/bloom-megatron-deepspeed [dostęp: 07.07.2023]
[11] https://huggingface.co/justinpinkney/falcon-7b [dostęp: 07.07.2023]
[12] https://www.cyfronet.pl/komputery/19030,artykul,athena.html [dostęp: 07.07.2023]

Zespół i wiedza

Dane, maszyny, a co z ludźmi? Czy potrafimy stworzyć interdyscyplinarny zespół i powtórzyć sukces **Blooma**? Główne założenia przełomowych architektur są dostępne w formie artykułów naukowych i modeli na otwartych licencjach – możemy z nich swobodnie korzystać. Architektura **Transformer** stworzona przez naukowców głównie z **Google Brain**, **Google Research**[13] została właśnie wykorzystana w modelu **GPT** (**OpenAI**). Oczywiście diabeł tkwi w szczegółach, ale tutaj nasze zespoły świetnych rodzimych badaczy w obszarze LLM już zacierają ręce.

Wspominaliśmy o modelu **Falcon**, który cechuje wysoka wydajność przy dość niskich wymaganiach związanych z procesem trenowania. Według autorów modelu udało się to uzyskać dzięki przemyślanym potokom związanym z jakością danych. Jednak już w 2020 roku **Sławomir Dadas**, **Michał Perełkiewicz** i **Rafał Poświata** w artykule ***„Pre-training Polish Transformer-based Language Models at Scale”*** pokazali metody filtrowania i oczyszczania danych Common Crawl w celu uzyskania wysokiej jakości korpusu internetowego. To wypisz wymaluj metody wykorzystywane przy tworzeniu zestawu danych **The RefinedWeb,** dzięki któremu w 2023 roku **Falcon LLM** odniósł taki sukces.[14] Zresztą **OPI PIB,** w którym pracują na co dzień cytowani badacze, jest autorem jednego z największych neuronowych modeli języka polskiego opartego na architekturze **BERT (Polish RoBERTa)**.

Ośrodek wrocławski pod kierunkiem **profesora Przemysława Kazieńki** z kolei oprócz wielu prac związanych z metodami przetwarzania dużych ilości tekstów ma w swoim dorobku także zaawansowane testy oraz metodykę oceny narzędzia **ChatGPT** i modelu **GPT-4** – to ponad 25 zadań i ponad 48 tysięcy podpowiedzi. Badania w tym obszarze odpowiednio ujęte w cyklu treningowym mogą dać rewelacyjne wyniki podczas trenowania własnego LLM. Dodajmy, że **Clarin-PL** to nie tylko ośrodek **Politechniki Wrocławskiej** (choć są tu formalnie liderem), ale także **Instytut Podstaw Informatyki PAN**, **Uniwersytet Łódzki**, **Instytut Slawistyki PAN**, **Uniwersytet Wrocławski** oraz **Polsko-Japońska Akademia Technik Komputerowych**.[15]

Państwowy Instytut Badawczy NASK to z kolei z jednej strony strażnik polskiego internetu (wymieniony w Ustawie o Krajowym Systemie Cyberbezpieczeństwa jeden z trzech krajowych CSIRT-ów), a z drugiej strony dynamicznie działający ośrodek badawczy z prężnie działającym Zakładem Inżynierii Lingwistycznej i Analizy Tekstu. Dzięki narzędziom tworzonym w **NASK PIB** możemy wzmocnić procesy odpowiedniego filtrowania i oczyszczania danych trenujących, a także poszerzyć zestaw metryk opisujących te dane. Jesteśmy w stanie identyfikować dane wrażliwe (np. osobowe) lub nieodpowiednie do trenowania (przemoc, rasizm, treści pornograficzne). Dodatkowo możemy przeprowadzać ekstrakcje tzw. instrukcji z nieustrukturalizowanego tekstu.

---

[13] Wcześniej działające jako oddzielne jednostki: Google Brain i Google Research, obecnie funkcjonują razem jako Google Deepmind.

[14] S. Dadas, M. Perełkiewicz, R. Poświata (2020) Pre-training Polish Transformer-based Language Models at Scale, National Information Processing Institute, https://arxiv.org/abs/2006.04229 [dostęp: 07.07.2023]

[15] https://clarin-pl.eu/index.php/o-nas/ [dostęp: 19.07.2023]

	Rok temu w podcaście „Nieliniowy” wyszło nam z Michałem Dulembą, że NASK PIB to takie „małe Stanford”. Dzisiaj, wciąż na obranym kursie, szturmem bierzemy standardy prowadzenia badań. We współpracy z najlepszymi (m.in. MIT, UTS), na froncie rozwiązań prawdziwych problemów (AI dla cyberbezpieczeństwa, medycyny, administracji publicznej i innych), okrzyknięci mianem „Moonshot” przez Mathworks na ich ogólnoświatowym Expo – tacy są naukowcy w NASK PIB. Jeśli tworzy się tak ważna i ciekawa inicjatywa, jak Big Polish Science, która w dodatku ma zapewniać transparencję i dostępność, to po prostu nie mogło nas tu zabraknąć. dr Inez Okulska Kierownik Zakładu Inżynierii Lingwistycznej i Analizy Tekstu Państwowy Instytut Badawczy NASK

Współpraca

Na koniec mały wtręt lingwistyczno-antropologiczny, stawiający problem braku polskiego LLM-a w innym świetle.
Od czasów Wilhelma von Humboldta wiemy, że język jest dla człowieka narzędziem poznania rzeczywistości. Kolejni kontynuujący tę myśl badacze, tacy jak Edward Sapir i Benjamin Lee Whorf, zwracają uwagę, że język kształtuje nasz sposób postrzegania świata[16], pomaga go porządkować, kategoryzować, zwracać uwagę na to, co ważne, pomijać nieistotne itp.[17] Co więcej, ponieważ to w języku formułujemy myśli, samo pisanie lub mówienie często utożsamiane jest z myśleniem – potwierdza to fenomen botów konwersacyjnych, którym nierzadko przypisuje się osobowość i inteligencję, mimo że są tylko narzędziami do generowania tekstu (którego wartość merytoryczna jest często dyskusyjna – bo też nie w celu edukacyjnym te boty stworzono). Jeśli zatem dla człowieka język tak silnie jest związany z myśleniem, to przeszczepiając (z konieczności) do rodzimej mowy rozwiązania stworzone oryginalnie dla języka angielskiego, pożyczamy wraz z nimi anglosaski dorobek kulturowy i obcy nam dotąd sposób myślenia. Nie najlepiej, jeśli zgodzimy się co do tego, że to w różnorodności tkwi piękno tego świata.[18]

Te wszystkie argumenty przemawiają za koniecznością stworzenia w Polsce własnego, zaawansowanego modelu językowego. Jesteśmy już gotowi, aby pójść śladem takich projektów jak Bloom, Falcon czy GPT-NeoX, ale skupić się wyłącznie na danych w języku polskim. Chcemy zbudować model, który stanie się uniwersalnym narzędziem, dostępnym

[16] Por. m.in. Edward Sapir: „Kultura, język, osobowość”; Benjamin Lee Whorf: „Język, myśl i rzeczywistość”.
[17] Konsekwencje tego procesu są obecnie argumentem m.in. w dyskusji o feminatywach, por. https://krytykapolityczna.pl/kraj/zenskie-koncowki-zycie-smierc-feminatywy/
[18] Przykładowe rtykuły w prasie dotyczące sytuacji mniej popularnych języków [dostęp: 20.07.2023]: https://www.wired.com/story/chatgpt-non-english-languages-ai-revolution/
https://slator.com/can-low-resource-languages-catch-up-multilingual-model-race/
https://slator.com/meta-warns-large-language-model-may-not-be-suitable-non-english-use/

dla każdej firmy, uczelni czy też zainteresowanego obywatela, niezależnie od profilu działalności. Tworząc otwarty model języka polskiego, chcemy przyczynić się do stymulowania innowacji i nauki w kraju. Umożliwienie szerokiego dostępu do zaawansowanych modeli generatywnych, takiego jak ten, którego budowę planujemy, może w przyszłości wspierać rozwój wielu sektorów nauki i gospodarki.

Dlatego też zwracamy się z gorącym apelem do wszystkich, którzy mogliby przyczynić się do naszego przedsięwzięcia przez udostępnienie znacznej ilości danych, do których posiadają prawa lub które są zaliczane do kategorii danych otwartych. Szczególnie cenne są dla nas dane tekstowe w języku polskim, które mogą służyć do dalszego trenowania i doskonalenia naszego modelu.

Osoby zainteresowane współpracą lub chcące wesprzeć projekt SpeakLeash w jakikolwiek inny sposób zapraszamy do kontaktu pod adresem e-mail: team@speakleash.org. Wspólnie możemy przyczynić się do rozwoju polskiej sztucznej inteligencji.

Modele wykorzystujące głębokie uczenie (ang. deep learning), takie jak LLM, z impetem zdobywają coraz większe obszary naszego życia, co stanowi zmianę porównywalną z rewolucją przemysłową. Wraz z tym procesem zwiększa się jednak bariera wejścia – niezbędne są ogromne ilości danych, potężna moc obliczeniowa, wykwalifikowani inżynierowie oraz interdyscyplinarna współpraca.

Aby Polska nie została w tyle, konieczne jest zintegrowanie środowisk nauki, biznesu oraz społeczności open science i open source. Ta synergia jest warunkiem sine qua non dla innowacyjności naszej cyfrowej gospodarki oraz konkurencyjności naszych produktów i usług.

W związku z tym powstał projekt Spichlerz, którego celem jest nie tylko przygotowywanie danych wysokiej jakości, ale również tworzenie innowacyjnych narzędzi, aktywizowanie polskiej społeczności AI oraz stanowienie łącznika między nauką a biznesem. Rosnąca liczba członków, partnerów technologicznych oraz z dnia na dzień większy zasób danych świadczą o słuszności tego kierunku.

Sebastian Kondracki
Chief Innovation Officer
Deviniti
Twórca projektu SpeakLeash a.k.a Spichlerz