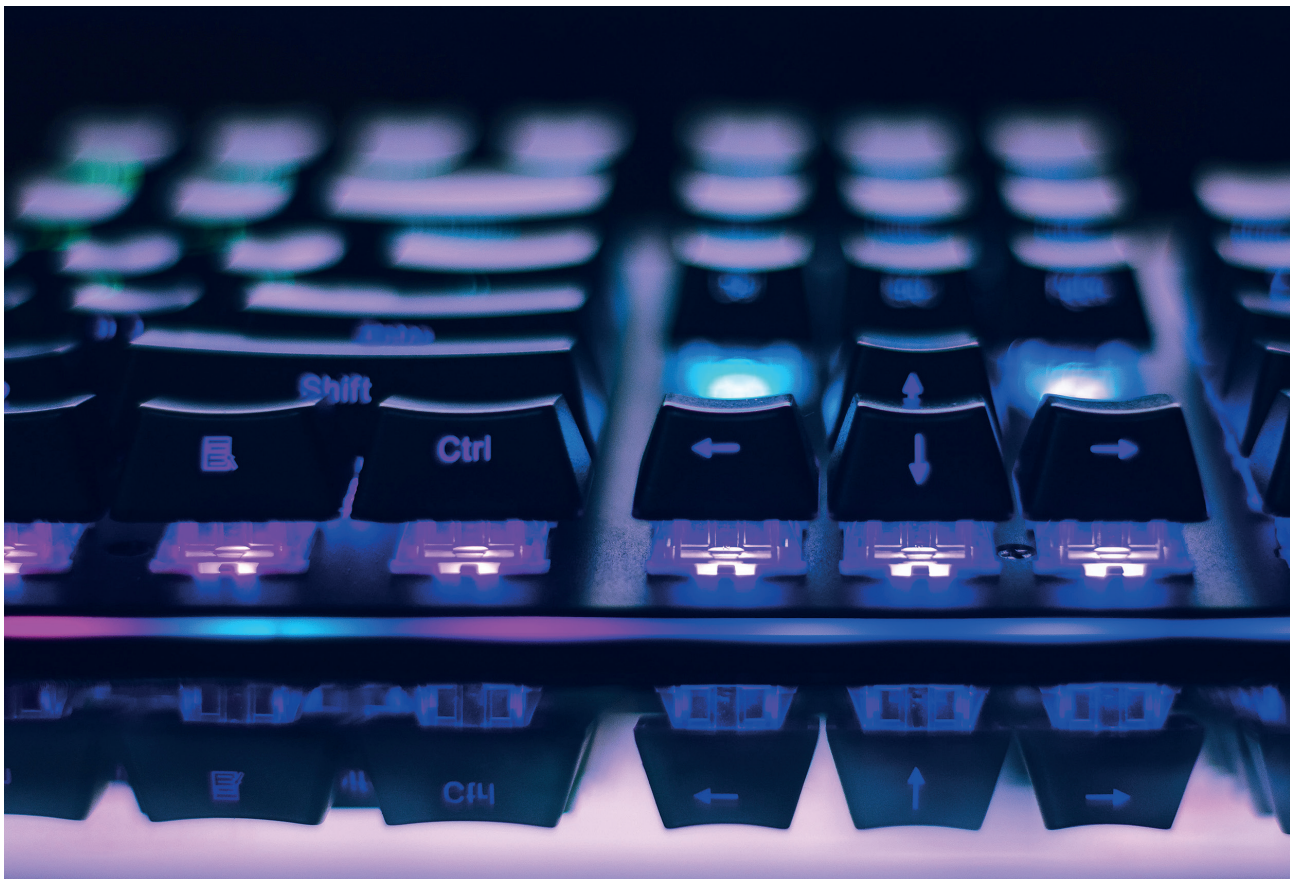
Własne czatowi dać słowo



Rośnie popularność globalnych rozwiązań z zakresu generatywnej sztucznej inteligencji. Jednocześnie pojawiają się inicjatywy tworzenia narzędzi lepiej dostosowanych do specyfiki poszczególnych języków. Duży model językowy o nazwie PLLuM powstaje również w Polsce.

W nauce o kulturze funkcjonuje pojęcie językowego obrazu świata. Oznacza ono obraz świata utrwalony w danym języku i determinujący sposób postrzegania rzeczywistości przez użytkowników tego języka. Posługując się tym czy innym językiem, człowiek dokonuje oglądu środowiska swojego życia według wzorców funkcjonujących w używanym języku. Język jest środkiem porozumiewania się, a zarazem narzędziem interpretacji otaczającego świata.

Te zależności znane są językoznawcom i kulturoznawcom od dawna. Badania prowadzone w latach trzydziestych XX w. przez Edwarda Sapira i Benjamina Lee Whorfa stały się podstawą do stworzenia teorii relatywizmu językowego. Według niej język, którym się posługujemy, wpływa na nasze

Andrzej Gontarz
ekspert ds. monitoringu rynku w zespole Sektorowej Rady ds. Kompetencji – Informatyka

myślenie o świecie. Z drugiej strony również rzeczywistość, w której żyjemy, oddziaływuje zwrotnie na kształt używanego przez nas języka.

Dzisiaj reguła ta nabiera nowego, szczególnego znaczenia w związku z pojawieniem się i wykorzystaniem na coraz większą skalę narzędzi generatywnej sztucznej inteligencji. W skali globalnej najbardziej znanym i powszechnie stosowanym rozwiązaniem wydaje się być ChatGPT. Problem jednak w tym, że model językowy, na którym się on opiera, trenowany był w głównej mierze na tekstach angielskojęzycznych. Siłą rzeczy odzwierciedla punkt widzenia charakterystyczny dla sposobu opisu świata w języku angielskim. Może więc nie uwzględniać wielu cech i składników specyficznych dla innych języków, w tym również języka polskiego. Podobnie wygląda sytuacja w przypadku innych popularnych modeli, jak Bard, Llama czy Bloom.

Dlatego w różnych krajach na świecie pojawiają się inicjatywy budowania własnych dużych modeli językowych LLM (*Large Language Model*). Przykładowo, o stworzeniu własnego modelu językowego Ăguila poinformowała w ubiegłym roku Hiszpania. Był on trenowany na mieszance tekstów z języka hiszpańskiego, katalońskiego i angielskiego. Finowie tworzą swój model Poro w ramach inicjatywy Silo AI, a Niemcy pracują nad własnymi rozwiązaniami w ramach projektu LAION. W Stanach Zjednoczonych trwają prace nad naukowym modelem językowym Aurora. We Francji firma Mistral AI stworzyła w ubiegłym roku model Mistral-7B, który jest dostępny na licencji *open source*. W tym roku wypuściła na rynek model Mistral Large, który jest już rozwiązaniem komercyjnym (podobnie jak jego odmiany Mistral Medium i Small). W Szwecji prace nad własnym modelem językowym rozpoczęły się już w 2021 r. To wtedy w ramach krajowego programu AI Sweden powstał zespół, którego celem miało być opracowanie szwedzkiego LLM-a o nazwie GPT-SW3. Został on ostatecznie upubliczniony w ub.r. i jest dostępny na licencji *open source*. W ramach rządowego programu AI Sweden współpracuje ponad 120 podmiotów, zarówno prywatnych firm, jak i instytucji publicznych oraz uczelni i instytutów badawczych.

Dobry, bo polski

Prace nad stworzeniem własnego, dużego modelu językowego prowadzone są również w Polsce. Odbywają się w ramach finansowanego ze środków Ministerstwa Cyfryzacji projektu PLLuM (*Polish Large Language Model*). Za jego realizację odpowiada konsorcjum, w skład którego wchodzą: Politechnika Wrocławska jako lider projektu, Państwowy Instytut Badawczy NASK, Ośrodek Przetwarzania Informacji – Państwowy Instytut Badawczy, Instytut Podstaw Informatyki PAN, Uniwersytet Łódzki oraz Instytut Slawistyki PAN.

Celem podjętych działań ma być opracowanie modelu, który będzie uwzględniał specyfikę języka polskiego, odzwierciedlał jego uwarunkowania i konteksty, a tym samym pozwalał jego rodzimym użytkownikom na bardziej precyzyjną, dostosowaną do ich różnorodnych potrzeb komunikację. Model ma mieć charakter otwarty. Będzie dostępny bezpłatnie dla wszystkich zainteresowanych jego wykorzystaniem w celach komercyjnych, edukacyjnych czy badawczych. Dzięki otwartej licencji użytkownicy będą mogli dostosowywać go do swoich specyficznych potrzeb i zadań, tworzyć na jego bazie własne rozwiązania.

Żeby móc spełnić pokładane w nim oczekiwania, PLLuM musi zostać wytrenowany na dużym zbiorze tekstów w języku polskim, zwanym korpusem. Zgromadzenie takiego zasobu wartościowych danych treningowych to jedno z najważniejszych wyzwań stojących przed konsorcjum realizującym projekt. Potrzebny jest dostęp do dużej puli różnorodnych tekstów w języku polskim – od literatury pięknej, przez dokumenty urzędowe i materiały techniczne, po mowę potoczną i wpisy internetowe. Chodzi o to, żeby w jak największym stopniu odzwierciedlały specyfikę naszego języka i były jak najbardziej dla niego reprezentatywne. Ważne jest też, aby korzystanie z tych danych było legalne, by były one dostępne zgodnie z obowiązującymi regulacjami prawnymi, w tym przepisami prawa autorskiego.

Dla zachowania reprezentatywności oraz zapewnienia jak najszerszego spektrum zawartej w naszym języku wiedzy, dane treningowe muszą pochodzić z wielu różnych źródeł. Powinny uwzględniać jak najwięcej aspektów rzeczywistego użycia polszczyzny oraz odzwierciedlać aktualne trendy czy procesy społeczne. Dla lepszego wytrenowania i uogólnienia modelu potrzebna też będzie jednak pewna pula tekstów anglojęzycznych oraz danych w językach słowiańskich i bałtyckich.

Budowę korpusu potrzebnych do uczenia przyszłego modelu tekstów zaczęto od przeglądu i skatalogowania zasobów znajdujących się w gestii podmiotów będących członkami konsorcjum realizującego projekt. Jednocześnie uruchomione zostało pozyskiwanie danych z innych źródeł, zarówno z publicznych, ogólnie dostępnych repozytoriów, jak i ze zbiorów będących w gestii poszczególnych osób, firm czy instytucji, które chciałyby się podzielić swoimi zasobami i udostępnić je do celów prac nad stworzeniem polskiego, dużego modelu językowego.

Twórcy PLLuM-a liczą m.in. na przychylność redakcji czy wydawnictw, ale zachęcają też do zgłaszania się wszystkich chętnych, którzy chcieliby wesprzeć projekt. Każdy może zgłosić chęć udostępnienia swoich zbiorów poprzez specjalny formularz na stronie internetowej projektu (https://pllum.org.pl/form).

Trudno na razie oszacować ostateczną wielkość zasobu treningowego. Jak spodziewają się przedstawiciele konsorcjum, będzie zapewne liczona w terabajtach danych. Kolejne, ważne zadanie to przekształcenie zgromadzonego zasobu danych, które często mają nieuporządkowany, surowy charakter w spójny, klarowny zbiór treningowy. Konieczne też będzie przygotowanie metadanych. Ich schemat ma być opracowany na bazie analizy istniejących już korpusów języka polskiego i dostosowany do specyfiki modelu.

Informatycy i językoznawcy razem

Budowa polskiego, dużego modelu językowego to nie pierwsze i nie jedyne przedsięwzięcie z dziedziny komputerowego przetwarzania języka naturalnego w naszym kraju. Do prac nad tworzeniem PLLuM-a przydatne mogą być również doświadczenia, rozwiązania i zasoby zgromadzone już przy realizacji wielu innych projektów.

Warto zauważyć, że większość członków konsorcjum PLLuM uczestniczy też w pracach działającego od ponad dziesięciu lat konsorcjum CLARIN-PL będącego częścią Europejskiej Infrastruktury Badawczej CLARIN (*Common Language Resources & Technology Infrastructure*). Jej celem jest udostępnianie badaczom, głównie z dziedziny nauk humanistycznych i społecznych, rozwiązań ułatwiających pracę z bardzo dużymi zbiorami tekstów. Powstają elektroniczne zasoby językowe i środki cyfrowe do przetwarzania dużych baz tekstów. Dostarczane są m.in. narzędzia do analizy języka naturalnego. Naukowcy mogą także korzystać z tworzonego w ramach projektu otwartego, publicznego archiwum zasobów językowych i materiałów źródłowych.

W konsorcjum CLARIN-PL skupionych jest sześć polskich jednostek naukowo-badawczych: Instytut Podstaw Informatyki PAN, Uniwersytet Łódzki, Instytut Slawistyki PAN, Uniwersytet Wrocławski, Polsko-Japońska Akademia Technik Komputerowych oraz Politechnika Wrocławska, która pełni funkcję lidera projektu. Jej zadaniem jest utrzymanie polskiego węzła infrastruktury w postaci Centrum Technologii Językowych CLARIN-PL. Koordynuje ono rozwój zasobów danych i narzędzi w postaci oprogramowania, zapewnia utrzymanie zaplecza technicznego (maszyn dostarczających potrzebnych mocy obliczeniowych) oraz prowadzi działania na rzecz upowszechniania rozwiązań inżynierii języka w środowisku akademickim. Badacze mogą korzystać z dostępnego zaplecza informatycznego bezpłatnie (otwarte licencje).

Infrastruktura i zasoby polskiego węzła – jak można przeczytać na stronie projektu – powstawały od początku z myślą o języku polskim, by zapewnić specjalne, autorskie rozwiązania pozwalające na osiąganie dokładnych wyników badań prowadzonych na tekstach w tym właśnie języku. Wśród zasobów językowych opracowanych w ramach infrastruktury CLARIN-PL znajdują się: Korpus języka polskiego Politechniki Wrocławskiej, Korpus dyskursu parlamentarnego, zbiór tekstów z posiedzeń plenarnych Sejmu i Senatu RP, dwujęzyczne korpusy równoległe tekstów współczesnych oraz wyszukiwarka polsko-angielskich korpusów równoległych. Są też takie narzędzia, jak: Platforma leksykalna, czyli system do przeszukiwania źródeł leksykograficznych, oraz Wyszukiwarka danych konwersacyjnych SpokesPL.

W zasobach CLARIN-PL znajduje się również Słowosieć, czyli wielki, relacyjny słownik semantyczny języka polskiego, oraz system o wdzięcznej nazwie Walenty, będący „słownikiem walencyjnym predykatów polskich". W uproszczeniu walencja oznacza w tym przypadku zależności określające, w jaki sposób poszczególne rodzaje wyrazów, głównie czasowniki, łączą się z wyrazami podrzędnymi (trochę na wzór wiązań chemicznych, stąd nazwa).

Jeden język, wiele korpusów

Do trenowania polskiego modelu językowego mogą być wykorzystywane też inne rozwiązania i zasoby stworzone i funkcjonujące poza siecią CLARIN. Jednym z nich jest udostępniony w 2012 r. Narodowy Korpus Języka Polskiego (NKJP). Stanowi on wspólne dzieło Instytutu Podstaw Informatyki PAN, jako koordynatora projektu, Instytutu Języka Polskiego PAN, Wydawnictwa Naukowego PWN oraz Zakładu Językoznawstwa Komputerowego i Korpusowego Uniwersytetu Łódzkiego. To zbiór tekstów, który pokazuje typowe użycia słów i konstrukcji językowych w naszym języku, określa częstości występowania poszczególnych form wyrazowych i konstrukcji składniowych, identyfikuje konteksty, w jakich pojawiają się dane wyrazy.

Zawiera on ponad półtora miliarda słów, a przypisane do niego wyszukiwarki pozwalają przeszukiwać jego zasoby z uwzględnieniem odmiany polskich wyrazów lub też wykonywać analizę budowy polskich zdań. Wśród źródeł wykorzystanych do stworzenia NKJP znalazła się zarówno klasyka literatury polskiej, jak też prasa codzienna i specjalistyczna, nagrania rozmów oraz teksty ulotne i pochodzące z internetu. Jego zawartość można przekształcić na przykład w instrukcje do modelu językowego, wykorzystać do oceny efektywności funkcjonowania stworzonego modelu czy zastosować do tzw. strojenia modelu.

W ramach programu DARIAH-PL (*Digital Research Infrastructure for the Arts and Humanities*) powstał również Korpus Współczesnego Języka Polskiego (KWJP) bazujący na tekstach z lat 2011–2020. Zawiera on ponad miliard

słów i stworzony został w Instytucie Podstaw Informatyki PAN w ramach projektu „Cyfrowa infrastruktura badawcza dla humanistyki i nauk o sztuce", prowadzonego w latach 2020–2023 przez konsorcjum naukowe DARIAH-PL. W skład tego konsorcjum wchodzi 18 wiodących w zakresie humanistyki cyfrowej instytucji naukowych w naszym kraju. Koordynatorem jest Uniwersytet Warszawski.

Swoje językowe korpusy narodowe, oprócz Polaków, mają też m.in. Brytyjczycy, Niemcy, Czesi, Amerykanie, Rosjanie. Oprócz tego, zarówno w Polsce jak i na świecie, tworzonych jest wiele innych, różnorodnych korpusów specjalistycznych i dziedzinowych, na przykład prawne, historyczne. Ciekawym przykładem takich obszarowych rozwiązań mogą być korpusy nierodzimych użytkowników danego języka, tzw. korpusy uczniowskie. Pozwalają one m.in. śledzić błędy popełniane przy używaniu języka, w tym i te wynikające z ograniczeń w ramach wspomnianej walencji. Jednym z pierwszych polskich korpusów uczniowskich jest *PELCRA Learner English Corpus* (PLEC) stworzony na Uniwersytecie Łódzkim. W innym obszarze na Uniwersytecie Warszawskim powstało Otwarte Repozytorium Korpusu Polskiego Języka Migowego.

Pasjonaci danych językowych

Pomocne w zdobywaniu danych tekstowych potrzebnych do opracowania modelu PLLuM, jak również i innych polskich dużych modeli językowych mogą być inicjatywy społeczne takie jak SpeakLeash. To stworzona oddolnie w ubiegłym roku nieformalnie funkcjonująca w modelu open science społeczność ludzi zainteresowanych ewidencjonowaniem i gromadzeniem danych, dzięki którym będą mogły powstawać duże modele językowe oferowane na różnych licencjach. Obecnie tworzy ją grupa około 300 osób, wśród których są przedstawiciele różnych zawodów i specjalności, pracownicy firm i instytucji, badacze z ośrodków naukowych oraz hobbyści i studenci. Docelowo projekt ma być realizowany w ramach powoływanej obecnie do życia fundacji. Osobowość prawna ułatwi współpracę z tymi podmiotami, które ze względów prawnych wymagają formalnej umowy.

„Naszym celem jest zbudowanie nowego i skatalogowanie istniejących zbiorów danych, aby zapewnić naukowcom możliwość prowadzenia najnowocześniejszych badań nad modelowaniem języka. Zbiory danych opracowane w ramach SpeakLeash są dostarczane z manifestami opisującymi licencjonowanie oraz zawierającymi statystyki, aby zapewnić lepsze dopasowanie do prowadzonych badań" – czytamy na stronie projektu, zwanego też przez jego inicjatorów Spichlerzem. W styczniu br. jego zasoby przekroczyły 1 TB danych i nieustannie się rozrastają. Jak zapewniają przedstawiciele opiekującego się danymi zespołu, są one już przygotowane i opisane pod kątem wymogów unijnego rozporządzenia AI Act.

Twórcy Spichlerza deklarują chęć współpracy i wymiany danych z konsorcjum realizującym PLLuM. Dotychczas wymieniano się już danymi z poszczególnymi instytucjami tworzącymi konsorcjum, współpracowano też z ośrodkiem CLARIN-PL. Nawiązano również współpracę z Akademickim Centrum Komputerowym Cyfronet AGH, które udostępnia moce obliczeniowe na swoich superkomputerach do realizacji zadań prowadzonych w ramach inicjatywy SpeakLeash.

Qra i Bielik

PLLuM ma być z założenia swoistym, polskim odpowiednikiem Chata GPT, modelem wszechstronnym, uniwersalnym, obejmującym całe spektrum możliwości języka polskiego. Nie oznacza to jednak, że w przestrzeni naukowej i badawczo-rozwojowej czy też biznesowej naszego kraju nie ma miejsca na tworzenie innych, dużych modeli językowych. Takie projekty są już realizowane.

Efektem współpracy Politechniki Gdańskiej i Ośrodka Przetwarzania Informacji (OPI) są trzy polskojęzyczne modele językowe o różnych poziomach zaawansowania i złożoności, udostępnione w tym roku pod wspólną nazwą Qra. W całości były trenowane na tekstach w języku polskim. Stanowią odpowiedniki otwartych narzędzi Mety czy Mistral AI, ale – jak zapewniają ich twórcy – Qra lepiej rozumie pytania zadawane w języku polskim i tworzy po polsku lepsze, bardziej spójne teksty. W testach *perplexity* modele Qra 7B i Qra 13B otrzymały lepsze wyniki w zakresie zdolności do modelowania języka polskiego niż modele Llama 2 i Mistral-7B.

Z kolei w ramach społecznościowego projektu Speakleash/Spichlerz, który realizowany jest w modelu *open science*, udało się stworzyć własny, polski model językowy Bielik. Jego twórcy zachęcają wszystkich chętnych do korzystania z wersji demonstracyjnej, by w ten sposób wspierać jego dalszy rozwój.

Te przykłady nie wyczerpują oczywiście możliwości działania w zakresie przetwarzania języka polskiego z użyciem narzędzi sztucznej inteligencji. Mogą powstawać modele mniejsze, lżejsze, ukierunkowane na obsługę konkretnych dziedzin czy wspomaganie realizacji konkretnych zadań, na przykład w bankowości, medycynie, marketingu. Mogą być one trenowane na mniejszej liczbie określonych rodzajów tekstów, na przykład dla obsługi codziennych potrzeb indywidualnych użytkowników, czy też uczone na zbiorach tekstów specjalistycznych, dziedzinowych na potrzeby biznesowe konkretnych firm i organizacji.