dr inż. Karol Grzegorczyk

Rozmowa z autorem pracy:
"Vector representations of text data in deep learning"

Co zainspirowało Pana do zaangażowania się w badania dotyczące uczenia maszynowego i sztucznej inteligencji? Skąd zainteresowanie tymi dziedzinami?

Na wybór obszaru badawczego wpłynęli moi promotorzy. Pasja, z jaką prowadzili badania w obszarze sztucznej inteligencji i uczenia maszynowego, skłoniła mnie do zgłębienia tego tematu w doktoracie.

Dlaczego Pana zdaniem warto „uczyć" komputery przetwarzania języka naturalnego? Jak w tym zakresie ocenia Pan rolę algorytmów sieci neuronowych?

Przetwarzanie języka naturalnego jest skomplikowanym problemem. Jawne zakodowanie wszystkich reguł, jakimi rządzi się język, jest trudne. Dodatkową komplikacją jest fakt, że język naturalny jest żywym tworem. Niektóre słowa są dodawane do słownika, a inne wychodzą z użycia. Dlatego też w przetwarzaniu języka naturalnego najlepiej sprawdzają się systemy uczące się.

Jakie są zalety reprezentacji wektorowej w analizie dużych zbiorów danych tekstowych?

Zaletą ukrytych reprezentacji wektorowych jest ich zdolność do modelowania podobieństwa. W oryginalnej przestrzeni każdy kawałek tekstu (np. słowo, zdanie albo dokument) jest reprezentowany przez identyfikator numeryczny. Na przykład algorytm nie wie o tym, że podobne słowa są podobne, bo każde słowo jest reprezentowane przez jakąś liczbę porządkową ze słownika. Gdy korzystamy z ukrytych reprezentacji wektorowych, to algorytm jest w stanie dowiedzieć się, że dane dwa słowa są podobne do siebie, a inne są niepodobne. Właściwość ta okazuje się kluczowa w wielu zadaniach przetwarzania języka naturalnego.

Jakich osiągnięć w analizowanej dziedzinie spodziewa się Pan w najbliższych latach?

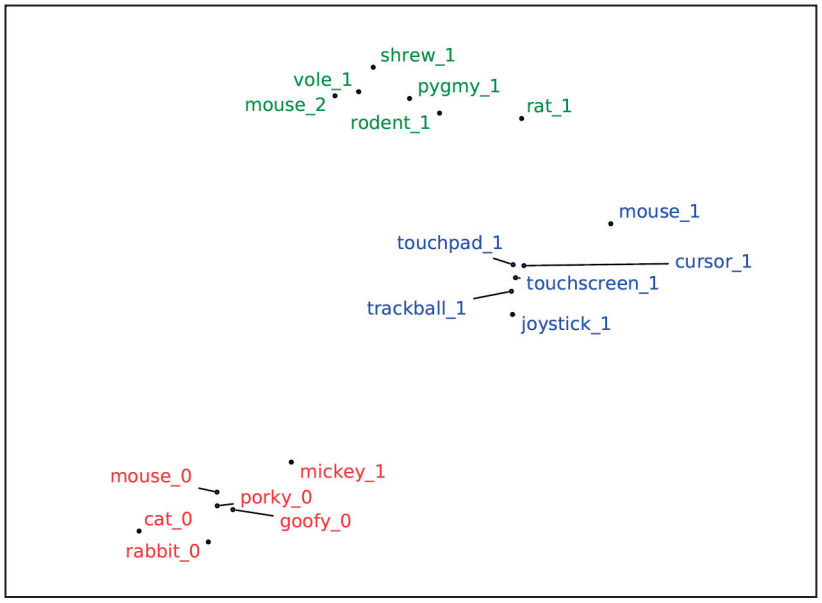
Podejrzewam, że w najbliższych latach powstaną systemy dialogowe (ang. *chatbot*) umożliwiające prowadzenie długotrwałych rozbudowanych konwersacji. Obecnie większość systemów dialogowych sprowadza się do stosunkowo prostych systemów odpowiadających na pytania. Możliwość wykorzystania informacji przekazanej przez użytkownika na przestrzeni długiej konwersacji wydaje się kluczowa do utworzenia inteligentnych asystentów imitujących ludzi.

Z jakich zasobów Cyfronetu i w jakim zakresie Pan korzystał?

Korzystałem z klastra obliczeniowego Prometheus. Klaster ten posiada wiele mocnych węzłów obliczeniowych wyposażonych zarówno w tradycyjne rdzenie obliczeniowe, jak i karty graficzne ogólnego przeznaczenia.

Jakich porad mógłby Pan udzielić osobom rozpoczynającym ścieżkę naukową? Na co te osoby powinny zwrócić uwagę?

Wydaje mi się, że przy wyborze obszaru badawczego warto kierować się autentyczną ciekawością i chęcią poznania świata. Na początku kariery naukowej warto możliwie dużo czasu poświęcić na naukę, samorozwój i poszerzanie horyzontów. Nie warto zbyt wcześnie decydować się na temat doktoratu. Dodatkowo myślę, że na początku kariery warto spróbować zająć się trudnymi zagadnieniami i wyzwaniami w danej dziedzinie. Daje to w dalszej karierze więcej wolności, gdyż łatwiej przejść od trudniejszych zagadnień do prostszych, niż w drugą stronę.



Dwuwymiarowa wizualizacja analizy głównych składowych najbliższych sąsiadów trzech znaczeń angielskiego słowa „mouse”