Entropia Wieży Babel

Już w czasach biblijnych rozmaitość języków utrudniała kontakty handlowe i była traktowana jako dopust Boży (Rdz 11,1–9). Współczesnym odpowiednikiem biblijnego przedsięwzięcia jest powstająca na naszych oczach globalna wioska (ciekawe, czy wynik będzie lepszy niż w Księdze Rodzaju). Choć po raz kolejny wykształcono *lingua franca*, to nie rozwiązuje to wszystkich problemów językowych (tym bardziej że, jak zwykle, wybór nie był specjalnie szczęśliwy). Dodatkowo, im dostęp do „informacji wszelakiej" staje się powszechniejszy, tym trudniej odnaleźć tę poszukiwaną.

Bardzo pomocna mogłaby się okazać możliwość automatycznego stwierdzania nie tylko, w jakim języku dana informacja została zapisana, ale również czego dotyczy, czy jej autorem jest ktoś powszechnie znany itd. O takim narzędziu marzą nie tylko internauci, ale również np. antyterroryści. Wydawałoby się, że trudno o taką metodę, zwłaszcza jeżeli miałaby być uniwersalna. Pogląd taki opiera się jednak na założeniu, że do selekcji informacji potrzebne jest jej rozumienie. Czy tak jest w rzeczywistości? Stosunkowo mało ludzi zna więcej niż kilka języków. Jednak większość z nas potrafi rozpoznać ich całkiem sporo. Bez trudu również dostrzegamy „obcy akcent" u osoby mówiącej językiem, który dobrze znamy. Umiejętności te opierają się na rozpoznawaniu dźwięków, melodii lub słów charakterystycznych dla danego języka. Wiemy, że ktoś mówi po włosku, francusku czy niemiecku, niekoniecznie wiedząc, co mówi. Nadal jednak nie widać sposobu na prostą algorytmizację takiej umiejętności. Specjaliści potrafią ocenić prawdziwość przypisania danego tekstu danemu, znanemu autorowi, ale wymaga to lat studiów i dobrego wyczucia.

A jednak taka prosta i uniwersalna metoda była od lat w zasięgu ręki. Wystarczyło sięgnąć po standardowy program kompresji danych. Dlaczego jednak artykuł [1] z doniesieniem na ten temat ukazał się w Physical Review Letters? Powód jest bardzo prosty. Autorzy użyli powszechnie znanego (przynajmniej tym, którzy używają unixa) gzipa do pomiaru względnej entropii urywków tekstu. W zasadzie nie odkryli niczego nowego. No, może poza jednym. Chyba nikt przed nimi nie przypuszczał, że akademicką wiedzę z tej dziedziny można tak łatwo i tak efektywnie zastosować.

Wiadomo powszechnie, że każde odkrycie staje się banalne, jak się już go dokona. Wystarczyło sięgnąć do źródłosłowiu nazwy biblijnej wieży, wywodzącej się od określenia „pomieszanie", żeby problem rozpoznawania języków skojarzyć z entropią, czyli miarą nieuporządkowania. Entropię informacji zakodowanej w postaci ciągu znaków można zdefiniować (Chaitin–Kołmogorow) jako minimalną wielkość programu (czyli innego ciągu znaków), który pozwala na odtworzenie pierwotnej informacji. Oczywiście, jak każda zwięzła definicja, jest ona niepraktyczna, gdyż nie sposób podać przepisu na najlepsze kodowanie. Jednak wiele z funkcjonujących na rynku programów kompresujących zbliża się do ideału tym lepiej, im ciąg znaków podlegających kodowaniu jest dłuższy.

Autorzy pracy użyli gzipa, ktory wykorzystuje algorytm LZ77 autorstwa Lempela i Ziva. Program ten szuka powtarzających się sekwencji znaków, zapisując dla nich (oprócz nich samych) odległości między poszczególnymi wystąpieniami. Dzięki temu często występujące sekwencje potrzebują mało miejsca do zakodowania ich pozycji, bo odległości między tymi pozycjami są małe.

W jaki sposób można wykorzystać tak działający program kompresujący do rozpoznawania tekstu? Wystarczy jego próbkę dołączyć na końcu tekstu referencyjnego. Jeżeli teksty są podobne, to program kompresujący „nie musi" uczyć się niczego nowego i zakodowana całość zajmuje mało miejsca – entropia nie rośnie znacząco. W przeciwnym przypadku program zaczyna pracę jakby od nowa i kompresja nie jest tak efektywna.

Autorzy pracy [1] przeprowadzili dwie analizy. W pierwszej starali się rozpoznać klasyków literatury włoskiej. Wybrali 90 tekstów 11 autorów i sprawdzali każdy z każdym, szukając najmniejszego przyrostu długości skompresowanego tekstu. Za sukces uznawali przypadek, w którym najbliższym tekstem okazywał się tekst tego samego autora. Udało się to w 84 przypadkach. Dodatkowo tylko raz zdarzyło się, że tekst tego samego autora nie znalazł się na co najmniej drugiej pozycji.

W drugiej analizie wzięli pod uwagę tekst „Powszechnej deklaracji praw człowieka" w pięćdziesięciu wersjach językowych (ograniczyli się do języków zapisywanych alfabetem łacińskim). Za pomocą (kosmetycznie) zmodyfikowanej metody z poprzedniej analizy wyznaczyli macierz odległości między językami, a następnie używając algorytmów wywodzących się z filogenetyki, zbudowali drzewo pokrewieństwa. Jako poszczególne konary tego drzewa otrzymali zasadniczo wszystkie podstawowe grupy językowe (wybranej pięćdziesiątki języków): romańską, celtycką, germańską, ugro-fińską, słowiańską, bałtyjską i ałtajską; a jako osobne gałęzie języki baskijski, maltański, ale również węgierski (który w tym ujęciu okazał się bliżej spokrewniony z tureckim niż fińskim i estońskim), angielski (najbliższy językom romańskim), bretoński (najbliższy germańskim) oraz języki albański i rumuński. Jak widać, choć metoda ta daje wyniki nie zawsze w pełni zgodne z filologią, to jak na procedurę całkowicie automatyczną sprawia się znakomicie.

Autorzy sugerują, że ich algorytm jest bardzo ogólny i może być zastosowany do dowolnych serii znaków, takich jak sekwencje genetyczne, szeregi czasowe, dane medyczne itp. Podkreślają, że żadna uprzednia wiedza o naturze analizowanych danych nie jest potrzebna.

A jakby tak obliczyć średnią entropię naszych wieszczów? Już dobrze ponad wiek trwa spór o to, „kto większym poetą był?" Jeszcze tylko należy się zdecydować, czy bardziej wartościowy jest utwór o jak najmniejszej, czy o jak największej entropii, ale potem będzie już ściśle i naukowo...

Piotr ZALEWSKI

[1] *Language Trees and Zipping*; D. Benedetto, E. Caglioti i V. Lereto *Phys. Rev. Lett.* **88**(2002)48702