Narzędzia komputerowe do analizy języków

W sieci słów

MARCIN MIŁKOWSKI
Instytut Filozofii i Socjologii
Polska Akademia Nauk, Warszawa
marcin.milkowski@gmail.com
Dr Marcin Miłkowski jest adiunktem w Zakładzie Logiki i Kognitywistyki IFiS PAN; zajmuje się filozofią umysłu i lingwistyką komputerową; stypendysta Fundacji na rzecz Nauki Polskiej i tygodnika „Polityka".

Do pisania i redagowania tekstów, liczenia i wyszukiwania. Do czytania, słuchania muzyki, przeglądania zdjęć i oglądania filmów. Do tego wszystkiego służą nam komputery. Nie zawsze jednak zdajemy sobie sprawę z tego, jak wiele ich możliwości wiąże się z przetwarzaniem języka naturalnego

Chcąc przeczytać najnowsze wiadomości z egzotycznych krajów, możemy skorzystać z internetowych serwisów tłumaczenia maszynowego. I choć wyniki mogą nas rozśmieszyć do łez, to takie serwisy cieszą się coraz większą popularnością, gdyż mimo wszystko dają jakieś przybliżenie znaczenia oryginału. Sukcesy lingwistyki komputerowej są niezaprzeczalne; nie jest przypadkiem, że Watson – superkomputer IBM przetwarzający gigantyczne ilości informacji, łączący je i wyciągający z nich wnioski – okazał się zwycięzcą w amerykańskim teleturnieju „Jeopardy".

Trzeba jednak przyznać, że najbardziej spektakularne sukcesy dotyczą języka angielskiego, bo choć rodzimych użytkowników angielskiego nie jest więcej niż 400 milionów, to ocenia się, że włada nim co najmniej półtora miliarda osób. Nic więc dziwnego, że inwestycje i badania naukowe od lat koncentrują się na angielszczyźnie.

Język językowi nierówny: polszczyzna różni się od angielskiego stosunkowo swobodnym szykiem, skomplikowaną odmianą, istnieniem rodzajów gramatycznych czy literami z ogonkami. Wszystko to sprawia, że zwykle nie można zastosować anglojęzycznych programów do obsługi polskiego. IBM Watson miałby znacznie większe problemy z pokonaniem polskich zawodników grających w „Va banque", odpowiedniku amerykańskiego „Jeopardy", prowadzonym przez kilka lat w TVP2 przez Kazimierza Kaczora. Tekst angielski Watson może przeanalizować gramatycznie, korzystając z dojrzałego formalizmu i słownika, który wskazuje na zależności między wyrażeniami w zdaniu (tzw. gramatyka zależnościowa). A taka gramatyka jest dla polszczyzny dopiero tworzona w Instytucie Podstaw Informatyki PAN.

Tysiąc reguł to za mało

Nie oznacza to, że wszystkie narzędzia obsługujące tekst lub mowę angielską są lepsze od tych dostępnych dla polszczyzny. Dzieje się tak między innymi dlatego, że komputery przetwarzają język w dwojaki sposób. Pierwszy to analiza statystyczna. Czasem udaje się znaleźć algorytmy, które pozwalają na skuteczną realizację analizy w sposób neutralny w stosunku do języka. Na takich algorytmach oparte są m.in. syntezatory mowy, które mogą występować w roli automatycznego lektora filmu czy udzielać głosu nawigacji GPS. Polski program IVONA należy do najlepszych na świecie i obsługuje bardzo wiele języków. Gorzej, niestety, z analizą mowy: tu zadanie jest znacznie trudniejsze i być może wymaga znacznie więcej wiedzy lingwistycznej niż budowa syntezatora. Anglojęzyczne systemy analizy mowy często opierają się także na regułach gramatycznych, które pozwalają na celniejsze wyodrębnianie wypowiadanych słów ze strumienia mowy. Stąd właśnie druga, historycznie wcześniejsza idea wyuczenia komputerów języka: przy użyciu reguł. A żeby uświadomić sobie, że rozpoznanie mowy dźwięku to zadanie niełatwe, wystarczy wsłuchać się w to, jak mówimy: w mowie wcale nie robimy przerw w tych samych miejscach, w jakich stawiamy spację, gdy piszemy na klawiaturze, a w dodatku mówimy niewyraźnie. Komputer może „usłyszeć" więc

IBM

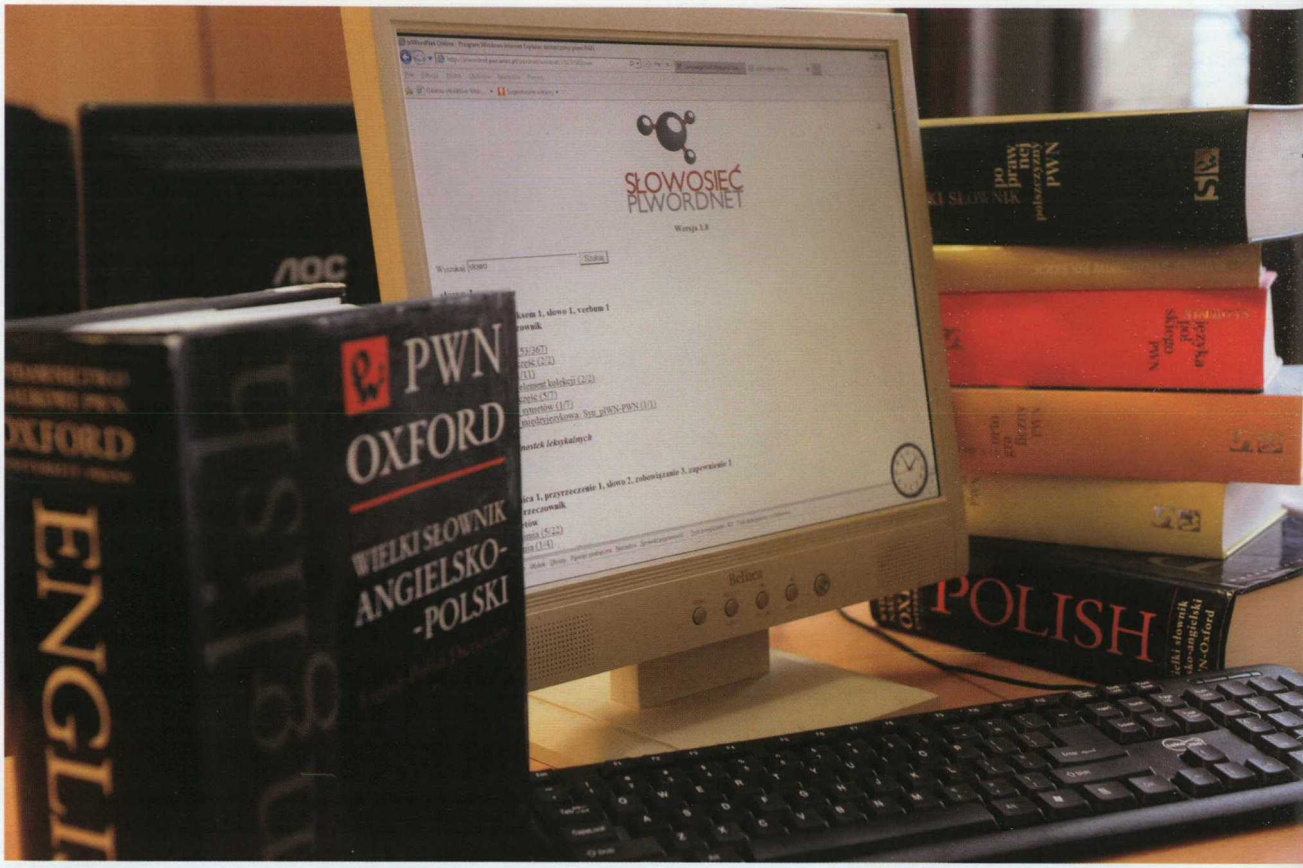
Watson ogrywa ludzi w *„Jeopardy"*

tekst, który wiernie trzeba byłoby zapisać jako „tojesjakiśtekswypwiadanyprzezczłieka". A my zrozumiemy go bez trudu, nawet jeśli w tle słychać muzykę.

Podobnie stosunkowo niezależne od języka są korektory pisowni, których używamy w edytorach tekstu, przeglądarkach internetowych czy telefonach. Bez trudu radzą sobie one z polszczyzną, choć mogą nas nie ostrzec, że w danym kontekście należy napisać raczej „żądny władzy" niż „rządny władzy" czy też „prosimy o niepalenie", a nie „prosimy o nie palenie". Korektory te sprawdzają każdy wyraz osobno w słowniku. Dopiero korektory gramatyczne czy stylistyczne, takie jak współtworzony przeze mnie korektor LanguageTool, są w stanie wykryć takie byki i zaproponować poprawki. I choć ma LanguageTool ponad 1000 reguł do sprawdzania tekstów w języku polskim, to daleko mu do kompletności. Automatyczne tworzenie reguł, choć w zasadzie możliwe, wymaga ogromnych zasobów tekstowych. Współczesne językoznawstwo opiera się na empirycznych badaniach uzusu językowego (od łac. *usus* - użycie, przyzwyczajenie) za pośrednictwem odpowiednio przygotowanych zbiorów tekstów lub innych zapisów języka, zwanych „korpusami".

Największe korpusy na świecie, na przykład stworzone na podstawie zeskanowanych książek Google Books, mogą osiągać wielkość setek miliardów wyrazów, choć to nie sama wielkość świadczy o ich jakości. Ważniejszy jest zrównoważony dobór próbek języka. Największym korpusem polszczyzny jest Narodowy Korpus Języka Polskiego, dostępny na stronie www.nkjp.pl, o którym „Academia" informowała już w numerze 2(22) w 2009 roku. Na stronie korpusu znajdują się wyszukiwarki, które pozwalają na badanie użycia pojedynczych słów czy złożonych wyrażeń, w tym słów, które się „lubią" (tzw. kolokacji). Zwłaszcza to ostatnie narzędzie może bardzo ułatwiać pracę redaktora czy tłumacza. Jeśli nie pamiętamy,

Jakub Ostałowski



Nowoczesne słowniki, takie jak Słowosieć, pomagają nie tylko człowiekowi, lecz także sztucznej inteligencji

czy mówi się „w porównaniu do" czy raczej „w porównaniu z", wystarczy sprawdzić, którym wyrażeniem chętniej posłużą się autorzy piszący staranną polszczyzną (przede wszystkim autorzy książek). A może okaże się, że oba wyrażenia są poprawne, lecz używane w różnych kontekstach? Polecam samodzielną analizę wyników.

Korpus byłby jednak nie tak przydatny, gdyby nie można było w nim wyszukiwać wyrażeń o określonej formie gramatycznej czy też znajdować w nim dowolnie odmienionych wystąpień danego słowa. Żeby to jednak było możliwe, konieczne jest odpowiednie przetworzenie tekstów, w tym zastosowanie słownika morfosyntaktycznego, czyli takiego, który zawiera formy wyrazów opisane gramatycznie. Opracowano ich już wiele. Do niedawna największymi takimi zasobami były Słownik Gramatyczny Języka Polskiego (SGJP), opracowany przez językoznawców Zygmunta Saloniego, Włodzimierza Gruszczyńskiego, Marcina Wolińskiego i Roberta Wołosza, oraz Morfologik, zasób, który utworzyłem i udostępniłem na swobodnej licencji w Internecie. Zaletą SGJP była oczywiście większa jakość opisu gramatycznego, a Morfologika – dostępność na liberalnej licencji, która umożliwiała wbudowanie go w bezpłatne programy. Naturalne więc wydawało się połączenie obu zasobów – i wbrew panującym czasem w nauce tendencjom do rozdrabniania projektów i dublowania wysiłków udało się stworzyć jednolity zasób o nazwie „PoliMorf" w ramach europejskiego projektu CESAR. Jest to obecnie największy słownik morfosyntaktyczny języka polskiego, zawierający ponad 400 tysięcy wyrazów (co daje ponad cztery miliony form odmienionych).

Innym rodzajem korpusu jest tzw. równoległy – zawierający teksty w wielu językach. Ponieważ wiele tekstów jest tłumaczonych, i to za pomocą narzędzi komputerowych, możliwe jest stosunkowo tanie tworzenie takich zasobów. I tak chociażby Unia Europejska nieodpłatnie udostępnia tłumaczenia swoich aktów prawnych – a także stworzone z nich korpusy. Dokumenty prawne nie są bowiem chronione prawem autorskim (w przeciwieństwie do innych tekstów, co stanowi istne utrapienie dla lingwistyki komputerowej; nie można po prostu skorzystać

ze zbiorów tekstów i opublikować korpusu z nich zbudowanego, jeśli teksty są dostępne na ograniczonej licencji). Z takich właśnie równoległych korpusów korzystają też programy służące do tłumaczenia maszynowego w sposób statystyczny, a więc uczące się nie na podstawie reguł, lecz obserwowanych w tekście prawidłowości statystycznych. Warto zresztą zauważyć, że usługa tłumaczenia automatycznego firmy Google jest wyraźnie lepsza podczas tłumaczenia tekstów ekonomicznych i prawnych. To zasługa właśnie korpusów unijnych, wykorzystanych przez Google do treningu.

„Pudło jest w piórze"

Warto zauważyć, że zadowalające tłumaczenie maszynowe miało powstać już bardzo dawno temu, jak wieścili optymiści. Już w latach pięćdziesiątych ubiegłego wieku inwestowano spore środki w tę dziedzinę sztucznej inteligencji, licząc na łatwe i szybkie tłumaczenie tekstów (zwłaszcza między rosyjskim i angielskim – nie trzeba przypominać, że były to często wydatki wojskowe). Lecz postępy przychodziły opieszale. Niektórzy wobec tego bardzo negatywnie odnosili się do samej idei tłumaczenia przez komputery. Na przykład wybitny logik i lingwista Yehoshua Bar-Hillel twierdził, że niemożliwe jest, aby komputer poprawnie przetłumaczył zdanie „The box is in the pen" (Pudełko jest w kojcu), gdyż wyrazy „box" i „pen" są bardzo wieloznaczne i komputer sobie z tą wieloznacznością nigdy nie poradzi. Tłumacz Google radzi sobie nie najgorzej („Skrzynka jest w zagrodzie"), bo jest systemem statystycznym, choć może czasem tworzyć zdania niegramatyczne. Nieco gorzej wypada tworzony w Poznaniu system regułowy Translatica („Pudło jest w piórze"). Zaletą Translatiki jest jednak ogromny zasób wyrazów (korzysta ona ze słowników wydawnictwa PWN, wzbogaconych o dane zgromadzone przez firmę PolEng), a także dobra obsługa gramatyki języka polskiego. Być może do budowy większej liczby reguł potrzebne będzie Translatice analizowanie dużych ilości tekstów w sposób statystyczny.

To właśnie statystyczna obróbka danych stanowi motor rozwoju praktycznych rozwiązań z lingwistyki komputerowej. Przykładem, oprócz tłumaczenia komputerowego czy korpusów, jest Słowosieć – słownik przedstawiający związki między polskimi wyrazami (synonimię, przeciwstawność, podrzędność i tak dalej), który dostępny jest na stronie http://plwordnet.pwr.wroc.pl/wordnet. To jeden z największych takich zasobów na świecie, choć utworzono go na Politechnice Wrocławskiej stosunkowo niedawno. Słowosieć, tworzona częściowo półautomatycznie – przy zastosowaniu specjalnych algorytmów do wykrywania związków między wyrazami w ogromnych zasobach tekstowych – jest odpowiednikiem amerykańskiego słownika WordNet, który stanowi wzorzec dla tego rodzaju prac. Słownik tego typu ukazuje wyrazy jako powiązane gęstą siecią relacji, co pozwala uchwycić ich znaczenie i stosunki wzajemne. Zastosowanie tego słownika jest nieco inne niż typowych słowników wyrazów bliskoznacznych, które mają pomóc w znalezieniu lepszego odpowiednika osobie piszącej jakiś tekst. Nie, nie chodzi o unikanie monotonii stylistycznej, lecz o możliwość zautomatyzowanego wnioskowania na temat tekstu. Wyposażenie wyszukiwarki sieciowej w taki słownik pozwala jej dostarczać wyniki, które obejmują nie tylko dosłownie sformułowane słowa kluczowe, ale także ich synonimy lub inne pokrewne wyniki. Mówiąc krótko, Słowosieć może być stosowana do tworzenia bardziej semantycznego Internetu.

Przed lingwistyką komputerową stoi wiele wyzwań. Choć dzięki licznym projektom naukowym i zastosowaniom komercyjnym w zakresie istniejących narzędzi i zasobów dla polszczyzny dokonały się ogromne postępy, nie ma mowy, aby komputer mógł dziś lub jutro w pełni rozumieć czy wytwarzać wypowiedzi w języku polskim na poziomie typowym dla Polaka. Sporo wody upłynie w Wiśle, zanim komputery dobrze przetłumaczą podstępne przykłady wymyślone przez Bar-Hillela. I to nie tylko dlatego, że brak środków finansowych i wsparcia największych korporacji. Też dlatego, że pełne rozumienie języka przez komputery to Święty Graal lingwistyki. Kto wie, czy osiągalny. ■

Chcesz wiedzieć więcej?

Miłkowski M., *The Polish Language in the Digital Age. Język polski w erze cyfrowej*, (red. Rehm G. i Uszkoreit H.), Springer, Berlin, Heidelberg, 2012 (dostępny bezpłatnie w całości pod adresem http://www.meta-net.eu/whitepapers/volumes/polish).

Przepiórkowski A., Bańko M., Górski R.L., Lewandowska-Tomaszczyk B. (red.). *Narodowy korpus języka polskiego*, Wydawnictwo Naukowe PWN, Warszawa, 2012