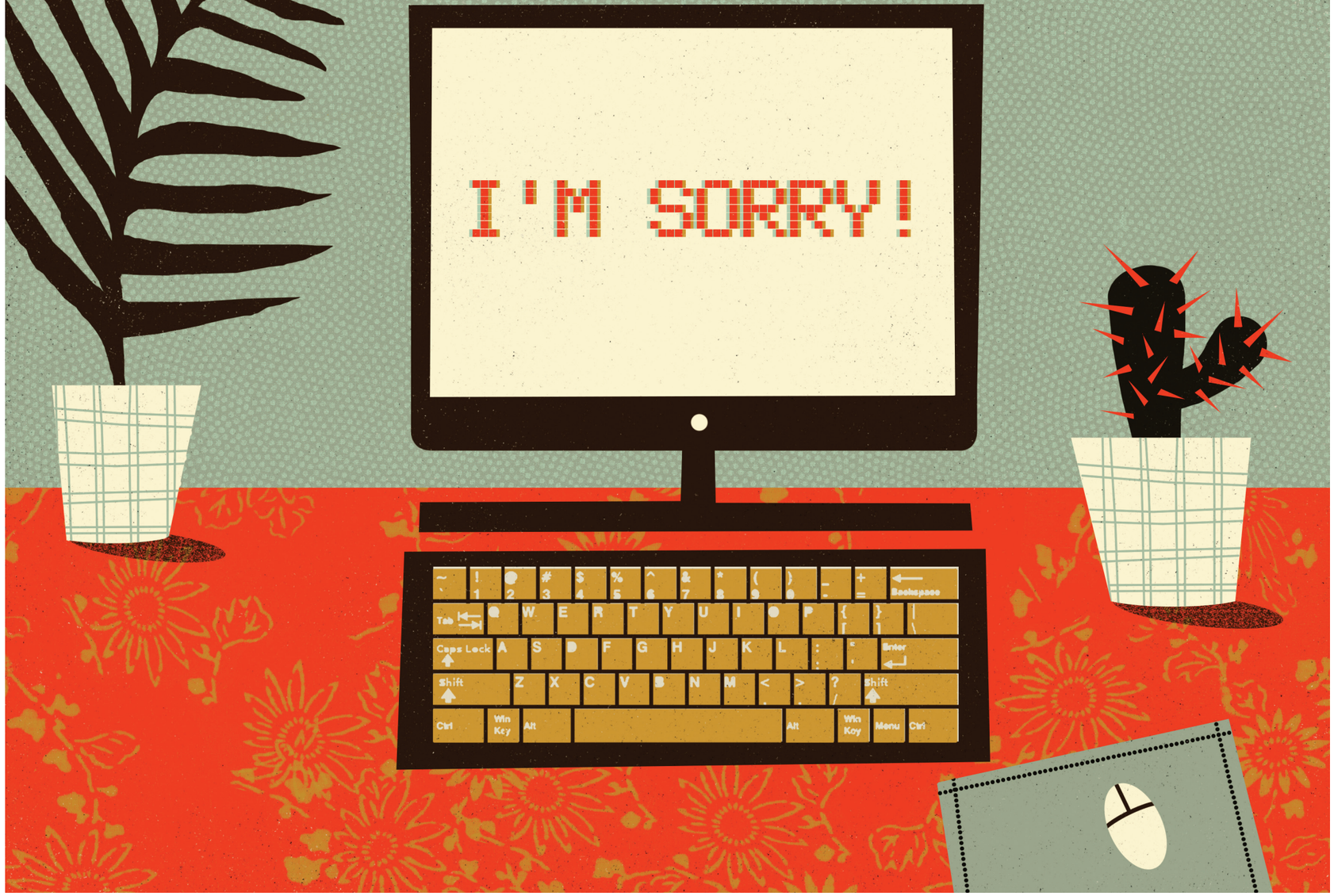
Narzędzia inżynierii lingwistycznej

Cyfrowi mówcy uczą się szybko

MACIEJ OGRODNICZUK
Instytut Podstaw Informatyki
Polska Akademia Nauk, Warszawa
maciej.ogrodniczuk@ipipan.waw.pl
Dr nauk humanistycznych Maciej Ogrodniczuk stopień magistra zdobył na Wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego. Pracuje w Zespole Inżynierii Lingwistycznej Zakładu Sztucznej Inteligencji IPI PAN. Jest kierownikiem projektu „Komputerowe metody identyfikacji nawiązań w tekstach polskich", realizowanego od 2011 do 2014 roku.



Komputerowe narzędzia lingwistyczne są tak powszechne, że już ich prawie nie zauważamy. Nie chcemy się jednak do nich dostosowywać – to one muszą zacząć rozumieć, co piszemy i mówimy, w dodatku nie po angielsku, ale we własnym języku. Do tego celu potrzebna jest sprawna analiza tekstu. Urządzenia radzą sobie z nią coraz lepiej

Wziąć do ręki telefon i podyktować e-mail lub SMS? To już nie science fiction. Automatyczne tłumaczenie, może jeszcze nie do końca dokładne, ale pozwalające zorientować się, o co chodzi w tekście napisanym w zupełnie obcym języku? Proszę bardzo. Synteza mowy? Czasem nawet nie wiemy, że na lotnisku czy dworcu mówi do nas komputer. Kiedy połączymy te narzędzia i wyposażymy w odpowiednią moc obliczeniową – dziś najpewniej wziętą z „chmury", czyli infrastruktury dostarczającej ją zdalnie na żądanie – możemy już sobie wyobrazić dostatecznie zrozumiałą rozmowę z Chińczykiem czy Tajem, w której pośrednikiem jest nasz smartfon. Wszystko dzięki narzędziom inżynierii lingwistycznej, coraz częściej dostępnym także dla języka polskiego. Proste? Tylko w teorii, bo cały proces wymaga rozwiązania wielu złożonych problemów.

Co zrobić ze „Środą Śląską"?

Gdy nasz komputerowy pomocnik upora się już z zamianą sygnału mowy na fonemy i spróbuje złożyć z nich tekst, otrzyma prawie zawsze wiele wariantów jego zapisu. Czy nasz rozmówca powiedział „Bóg", „Bug" czy „buk", „przednim" czy „przed nim"? Co miał na myśli, mówiąc o „zdjęciu kurtki"? Te kwestie należy rozwikłać. A do uruchomienia procesu tłumaczenia wciąż daleko – trzeba jeszcze głębiej przeanalizować treść: wykryć nazwy własne (choćby po to, by zachować w tekście Środę Śląską) czy związki składniowe między częściami zdań, nie zapominając o tradycyjnych właściwościach polszczyzny – zagnieżdżeniach, nieciągłościach, swobodzie szyku. Sprawy komplikują się jeszcze bardziej, gdy potrzebny jest zaawansowany sposób przekształcania tekstu, np. gdy chcemy automatycznie utworzyć jego streszczenie. Tu muszą zostać uwzględnione złożone relacje między częściami wypowiedzi, jak chociażby zbiory tekstowych odwołań do tego samego obiektu (tzw. koreferencja), co pozwoli zapewnić stylistyczną spójność wyniku (np. poprzez zastąpienie zaimka w streszczeniu jego pełnoznacznym odpowiednikiem ze zbioru). Synteza sygnału mowy z tak przeanalizowanego tekstu jest już zadaniem stosunkowo prostym, zresztą rozwiązanym przez jedną z polskich firm ponoć najlepiej na świecie.

Jak można zauważyć, źródłem wszystkich powyższych trudności jest wielopoziomowa wieloznaczność języka. Człowiek rozstrzyga ją kontekstowo za pomocą wszystkich środków, jakimi dysponuje – prawdopodobieństwa wystąpienia danej konstrukcji, wiedzy o świecie, sygnałów niewerbalnych. Dla komputera to zadanie jest niezwykle trudne, ale już teraz częściowo możliwe.

Już nie „szeptam", czyli poprawność

Umiemy już sporo. Lekarstwem na wiele poziomów niejednoznaczności jest wielowarstwowa analiza tekstu, od podziału na zdania i słowa przez ujednoznacznianie kategorii gramatycznych, sensów słów i struktur gramatycznych aż po analizę semantyczną.

Zacznijmy od segmentacji tekstu i odmiany polskich wyrazów. Są to zadania podstawowe, ale nawet one umożliwiają osiągnięcie ogromnych korzyści. Już dziś w wydawnictwach działają automatyczne weryfikatory poprawności, wykraczające poza to, co potrafią edytory z pakietów biurowych. Wyszukują formy błędnie uznawane za poprawne („bawoła”, „szeptam”), powtórzenia w szerszym kontekście niż tylko słowo po słowie („z całą klasą weszli do klasy”), a także formy o wspólnym rdzeniu („postawiła ją na szafce i ustawiła tak, żeby...”), niekonsekwencje w zapisie nazw („Hoffmann” vs. „Hoffman” w różnych częściach książki), błędy w interpunkcji (mała litera po kropce kończącej zdanie czy brak przecinka przed „bo”). Programy podpowiadają miejsca potencjalnie zawierające wystąpienie błędu („na prawdę”, „za zwyczaj”), oznaczają wyrazy identycznie brzmiące czy często mylone („rządzą” vs. „żądzą”).

Trwają wielopoziomowe prace nad przetwarzaniem składniowym polszczyzny – analizą składnikową, zależnościową, funkcyjną – wykorzystujące komputerowe implementacje różnych gramatyk języka polskiego i słowniki walencyjne (rejestrujące sposoby łączenia się czasowników z frazami innego typu). Związki semantyczne (takie jak bliskoznaczność, przeciwstawność czy podrzędność) modelujemy z powodzeniem m.in. w Słowosieci – sieci zależności między jednostkami leksykalnymi, która jest jednym z największych zasobów tego rodzaju na świecie. Tworzymy ontologie i inne modele wiedzy ogólnej – znów w postaci danych językowych.

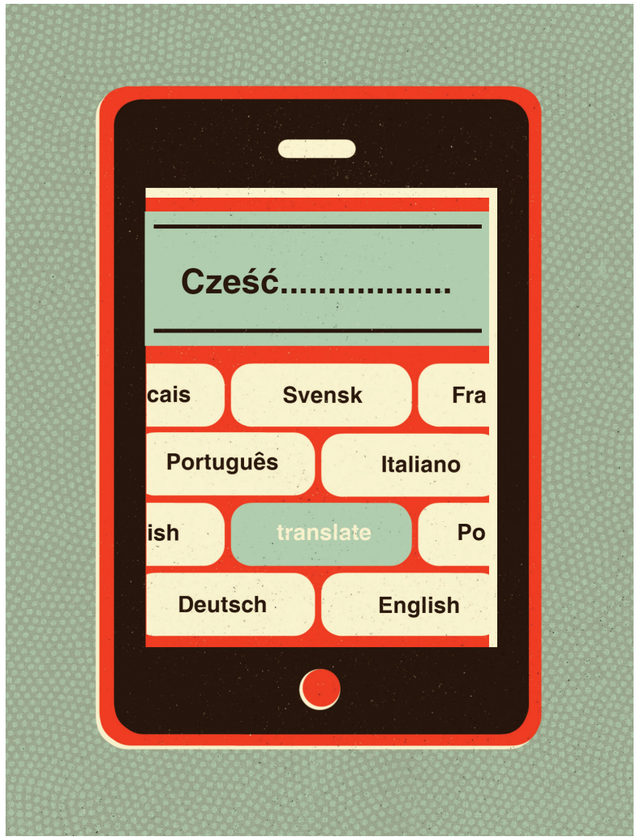
Równolegle powstają systemy wykorzystujące łącznie wszystkie omówione warstwy opisu językowego tekstu. Służą one np. do badań nad zrozumiałością tekstów, które mogą nam wszystkim poprawić kontakty z urzędami, zwiększyć przystępność instrukcji obsługi urządzeń czy podręczników. Oprócz tradycyjnych metod opartych na długości zdań i słów brane są w nich pod uwagę także wskaźniki leksykalne (fachowość, przestarzałość, wieloznaczność słownictwa), morfoskładniowe i składniowe (obecność imiesłowów, negacji, zdań wielokrotnie złożonych podrzędnie). Dynamicznie rozwija się także nurt łączenia narzędzi wielojęzycznych, np. w modelu wykorzystującym automatyczne tłumaczenie tekstu, jego przetworzenie bardziej zaawansowanym narzędziem dla języka obcego, a potem przeniesienie wykrytych własności na język źródłowy.

Narzędzia lingwistyczne są często elementami większych systemów (np. zarządzania treścią), oferują możliwość automatycznej klasyfikacji tekstów, generowanie list dokumentów treściowo podobnych czy ekstrakcję ważnych zwrotów i fraz. Użytkownicy systemów wspomagających sprzedaż aukcyjną mogą już korzystać np. z automatycznego tłumaczenia opisów oferowanych przedmiotów na inne języki w sposób umożliwiający ich prezentację w zagranicznych serwisach. Jedna po drugiej powstają witryny umożliwiające uzyskiwanie odpowiedzi na pytania zadawane w języku naturalnym w rodzaju „Kiedy została ukończona wieża Eiffla?” – znów dzięki inżynierii lingwistycznej, zaprzęgniętej do pracy nad analizą zawartości stron internetowych.

Niezgoda statystyczno-lingwistyczna

Od lat w procesie analizy języka konkurują ze sobą (choć ostatnio zaczynają też współpracować) dwa podejścia: teoretyczne i empiryczne. Celem pierwszego jest modelowanie języka poprzez zrozumienie jego struktury, drugie zakłada skupienie się na efektywności przetwarzania zarejestrowanych (np. w korpusie tekstów) danych językowych ze wszystkimi tego konsekwencjami – z uwzględnieniem ich złożoności, obecności wypowiedzeń niepoprawnych itp. Polem działania podejścia teoretycznego jest analiza lingwistyczna wykorzystująca metody i narzędzia regułowe, tworzone na podstawie idealnego modelu języka. Domeną podejścia empirycznego jest natomiast szeroko zakrojona analiza statystyczna zaobserwowanych zjawisk językowych.

Krytyka podejścia empirycznego wskazuje zwykle na potrzebę modelowania pełnej kompetencji językowej, nieznajdującej odzwierciedlenia w dowolnie dużym zbiorze przykładów, co wynika z nieograniczonej potencji języka. Lingwiści korpusowi zwykle odpowiadają na te argumenty, podając trudne przykłady rzeczywistych danych, których nie obejmują wyrafinowane teorie. Drogi obu filozofii cyklicznie schodzą się i rozchodzą od lat 50. XX wieku – i ostatnio, po złotej erze triumfu metod statystycznych, zdaje się, że obserwujemy powrót



do metod lingwistycznych, o czym świadczy chociażby poszerzanie statystycznych systemów tłumaczenia maszynowego o reguły gramatyczne, co znacznie poprawia jakość ich działania. Dobrym przykładem przenikania się obu podejść jest nowa metoda automatycznego uczenia się informacji lingwistycznej (morfologicznej) z danych nieprzetworzonych, co pozwala sądzić, że oba światy będą się coraz bardziej uzupełniać.

Kiedy komputer powie „I’m sorry”?

Znaczenie zasobów i narzędzi językowych wydaje się stale rosnąć, głównie ze względu na wykładniczo rosnący zalew informacji – sam Facebook i Twitter wytwarzają codziennie rękami swoich użytkowników setki terabajtów danych, które trzeba efektywnie udostępniać, ale także badać. Google nie zastąpi lingwistów nie tylko dlatego, że obecnie wyszukiwarki internetowe nie oferują lingwistycznego opisu treści, ale także ze względu na – wciąż jeszcze i na całe szczęście – niereprezentatywność internetowego modelu języka dla całości polszczyzny (i każdego innego języka). Do tego wciąż potrzebujemy referencyjnych zbiorów tekstów, takich jak Narodowy Korpus Języka Polskiego, oraz narzędzi i zasobów językowych – słowników, banków drzew składniowych, baz relacji semantycznych – coraz częściej tworzonych właśnie na podstawie danych korpusowych.

W dłuższej perspektywie przyszłość inżynierii lingwistycznej jest związana ze stopniowym wyposażaniem komputera w umiejętność rozumienia semantyki wypowiedzi, zdolność wnioskowania i agregacji informacji (np. w celu tworzenia zwięzłych podsumowań treści wielu dokumentów). HAL 9000, inteligentny komputer z książki Arthura Clarka „2001: Odyseja kosmiczna”, żeby poinformować astronautę, że nie wpuści go do statku kosmicznego zdaniem „I’m sorry Dave, I’m afraid I can’t do that”, musiał zrozumieć człowieka, odnieść jego słowa do rzeczywistości, podjąć decyzję (tu: odrzucić polecenie mimo możliwości wykonania go), sformułować odpowiedź (dobrać odpowiednio słowa – w tym przypadku uprzejmie odmówić, odnosząc się do poprzedniej części dialogu) i przekazać ją w zrozumiały sposób. To działanie wykracza poza samo przetwarzanie tekstu i prowadzi do pojęcia sztucznej inteligencji ogólnej, wymagającej zdolności odczytywania intencji, wydawania sądów, może także pewnego rodzaju samoświadomości. Na to jeszcze z pewnością trochę poczekamy, ale do sprawnego pomocnika w analizie informacji i wykonawcy poleceń przekazywanych za pomocą języka mamy już niedaleko. ■

Ilustracje **Dawid Ryski**

Chcesz wiedzieć więcej?

CLIP – Computational Linguistics in Poland. http://clip.ipipan.waw.pl

Jurafsky D., Martin J.H. (2008). *Speech and Language Processing* (wyd. II). Prentice Hall Series in Artificial Intelligence, Prentice Hall.

Przepiórkowski A., Bańko M., Górski R.L, Lewandowska-Tomaszczyk B. (red.) (2012). *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo Naukowe PWN. (tekst dostępny bezpłatnie w całości pod adresem: http://nkjp.pl/settings/papers/NKJP_ksiazka.pdf.).

Wilks Y. (2005). *The History of Natural Language Processing and Machine Translation. Encyclopedia of Language and Linguistics.*