JĘZYK NATURALNY W WYSZUKIWANIU INFORMACJI

Wiesław Babik
Instytut Informacji Naukowej
i Bibliotekoznawstwa
Uniwersytet Jagielloński

Systemy wyszukiwania informacji, przetwarzanie języka naturalnego, wyszukiwanie informacji

Wprowadzenie

Przedmiotem moich rozważań są wybrane problemy przetwarzania języka naturalnego (ang. *Natural Language Processing* – NLP) w komputerowych systemach wyszukiwania informacji. Dotyczą one tworzenia charakterystyk wyszukiwawczych dla tych systemów.

To interesujący badawczo, a zarazem złożony i do tej pory niezbadany dogłębnie, problem m.in. nauki o informacji. Jest aktualny i ważny nie tylko ze względów teoretycznych, z uwagi na potrzebę analizy dotychczasowej i obecnej praktyki w tym zakresie, lecz także praktycznych, gdyż wymaga tego nowa sytuacja, powstała w wyniku pojawienia się Internetu, w którym podstawowym lingwistycznym narzędziem wyszukiwawczym stały się słowa i wyrażenia kluczowe będące elementami języka naturalnego.

W polskiej nauce o informacji problematyka ta stosunkowo rzadko była przedmiotem badań oraz teoretycznych refleksji i pogłębionych analiz [m.in. 1, 4, 11]. W zagranicznym piśmiennictwie sytuacja jest lepsza [5, 7, 14]. Oprócz licznych opracowań teoretycznych można znaleźć (także w Internecie) opisy wielu badań przeprowadzonych na tekstach z różnych dziedzin, głównie w języku angielskim.

Impulsem do ponownego zajęcia się tym tematem i badaniami jest próba weryfikacji przekonania, że dzięki wykorzystaniu nowoczesnych metod przetwarzania języka naturalnego komputer w sferze indeksowania i wyszukiwania informacji może całkowicie zastąpić, a nawet wyeliminować człowieka. Wydaje się to możliwe dzięki zastosowaniu metod automatycznych, których skuteczność nie ustępuje, a nawet czasami przewyższa skuteczność klasycznych metod stosowanych przez człowieka, nazywanych metodami kognitywnymi. Możliwość zautomatyzowania tego procesu daje niewątpliwe korzyści w postaci oszczędności kosztów pracy oraz czasu związanych z opracowaniem odpowiedniej

charakterystyki. Dlatego badania przybliżające praktyczne wdrożenie takich możliwości są cenne i użyteczne.

Artykuł ma na celu, w oparciu o istniejące piśmiennictwo, udzielenie odpowiedzi na pytanie o stan zaawansowania i perspektywy dalszych prac nad wykorzystaniem języka naturalnego w procesach wyszukiwania informacji w systemach wyszukiwawczych, w tym w Internecie. Odpowiedź na to pytanie może stanowić punkt wyjścia do rozważań nad szerszym włączeniem się polskich informatologów w ten szeroki nurt badawczy i aplikacyjny.

1. Kontekst i wielowymiarowość problemu

Problematyka wyszukiwania informacji w języku naturalnym ma charakter wieloaspektowy i mieści się w ramach komputerowego przetwarzania języka naturalnego. Jest to ważny nurt nie tylko we współczesnej informatyce, ale i w nauce o informacji. Ma on duże znaczenie naukowe, ale też istotne zastosowania praktyczne. Dlatego warto zająć się, obszarami aktywności intelektualnej badaczy i praktyków rozwijających nowe idee i usługi w tym zakresie. Chodzi o to, aby usprawnić „inteligentny" kontakt człowieka z komputerem w wyszukiwaniu informacji, przy czym powinien on odbywać się za pomocą bardzo prostych środków i metod, niewymagających od użytkowników większej wiedzy i zaangażowania się. Najlepiej, żeby był to język naturalny.

Nowoczesne rozwiązania programistyczne w tym zakresie wymagają dobrego zaplecza lingwistycznego, gdyż wdrożenie naturalnej komunikacji językowej człowieka z komputerem jest znacznie trudniejsze niż się wydaje. Swobodna konwersacja w języku naturalnym musi się opierać na wynikach badań prowadzonych na styku informatyki, językoznawstwa (filologii) i nauki o informacji (informacji naukowej).

Pospolita czynność komunikacyjna, jaką jest wyszukiwanie informacji, będąca w istocie dialogiem człowieka z systemem komputerowym/informacyjnym, w rzeczywistości okazuje się bardzo złożonym procesem informacyjnym. Wymaga więc także badań i zaplecza informatologicznego. Na gruncie polskim zwrócili na to uwagę m.in. Piotr Nowak i Paweł Nowakowski w artykule pt. *Infolingwistyka jako forma integracji językoznawstwa z nauką o informacji* [15], w którym teorię języków informacyjno-wyszukiwawczych, zagadnienia metadanych, języki programowania, kompleks badań kwantytatywno-statystycznych nad strukturą tekstów oraz bibliolingwistykę słusznie potraktowali jako obecnie bardzo ważne pola badawcze nauki o informacji. Dziedziny te mogą stanowić mocny fundament dla prac aplikacyjnych związanych z tworzeniem systemów informacyjno-wyszukiwawczych zdolnych do rozumienia wypowiedzi formułowanych w języku naturalnym. W badaniach nad przetwarzaniem języka polskiego nikt nas nie zastąpi.

Wykorzystanie języka naturalnego w wyszukiwaniu informacji zawsze trzeba rozpatrywać w kontekście konkretnego języka etnicznego. Język naturalny jest nośnikiem informacji silnie związanych z narodem i kulturą użytkownika, a poszczególne języki etniczne generują różne problemy związane z ich wykorzystaniem w systemach informacyjno-wyszukiwawczych.

Proces komunikowania się człowieka z komputerowym systemem informacyjno-wyszukiwawczym musi być sformalizowany. Do tego jest niezbędna:

– analiza leksykalna, zmierzająca do ustalenia, co dane słowa znaczą;

– analiza syntaktyczna (gramatyczna), która pozwala określić, jaką rolę pełnią poszczególne słowa w strukturze zdania i/lub tekstu;

– analiza semantyczna, zmierzająca do określenia znaczenia całej wypowiedzi;

– pragmatyka uwzględniająca wiedzę pozajęzykową.

Niezbędne są więc metody formalnego opisu składni języka naturalnego, metody analizy syntaktycznej i semantycznej zdań oraz wiedza pozajęzykowa systemu o świecie.

Pod wpływem Internetu w systemach informacyjno-wyszukiwawczych uległy znacznym zmianom sposoby formułowania pytań wyszukiwawczych (kwerend). Wyszukiwanie informacji w sieci różni się pod wieloma względami od wyszukiwania w tradycyjnym środowisku wyszukiwawczym. Cechuje je – w większym stopniu – tzw. zasada najmniejszego wysiłku (*the principle of least effort*). Użytkownicy obniżają standardy jakości informacji na rzecz łatwości i szybkości jej wyszukania. Są też niecierpliwi i „niewyrozumiali" dla serwisów trudnych w obsłudze. Raczej „skaczą" między stronami WWW niż wchodzą głębiej w zawartość takich serwisów. Jak pokazują liczne badania, większość użytkowników systemów informacji elektronicznej z reguły nie stosuje zaawansowanych technik wyszukiwawczych, nie formułuje złożonych pytań i nie wchodzi w głębsze interakcje z systemem ani nie wykorzystuje wszystkich jego możliwości, niezależnie od tego, czy jest to wyszukiwarka internetowa, czy biblioteczny OPAC. Niewielki procent kwerend zawiera operatory boolowskie. Zdecydowana większość sesji wyszukiwawczych w Web opiera się na pytaniach składających się średnio z dwóch terminów. Więcej niż połowa użytkowników przegląda jedynie pierwsze dziesięć (a nawet mniej) „wydanych" przez wyszukiwarki opisów dokumentów. Użytkownicy informacji doby Internetu stają się „informacyjnymi graczami", dla których wyszukiwanie informacji, niezależnie od celu, zawiera elementy zabawy, gry, konkurencji [6]. Coraz częściej wyszukiwanie dotyczy kilku informacji jednocześnie (*multitasking information behavior & information taskswitching*). Użytkownicy uważają wyszukiwanie informacji za pomocą Google i innych uniwersalnych serwisów jako łatwe, szybkie i przyjemne, natomiast wyszukiwanie za pomocą narzędzi wyspecjalizowanych, np. bibliotecznych – jako trudne, powolne i wymagające zbyt dużego (w stosunku do przewidywanych rezultatów) wysiłku intelektualnego, co sprawia, że rezygnują z tych ostatnich. Czasami to postępowanie jest pragmatycznie uzasadnione,

z komputerami w języku naturalnym, w szczególności z komputerowymi bazami danych, a także z automatycznym tworzeniem baz danych na podstawie tekstów języka naturalnego. Ważne są tu także analiza i synteza mowy niezbędne dla komputerów komunikujących się z użytkownikiem w subkodzie akustycznym języka naturalnego [4]. W towarzystwie tego terminu zwykle pojawiają się takie terminy, jak: inżynieria lingwistyczna (ang. *language engineering*, LE), lingwistyka komputerowa lub lingwistyka informatyczna (ang. *computational linguistic*, CL), inżynieria języka naturalnego (ang. *natural language engineering*, NLE), technologia języka (ang. *language technology*, LT lub *human language technology,* HLT) [20, 19, 16]. Tradycyjnie za dziedziny przetwarzania języka naturalnego uważa się: wyszukiwanie informacji w dokumentach (pełnotekstowych), grupowanie dokumentów (klasteryzację), klasyfikację opartą na wzorcach oraz klasyfikację bezwzorcową [12].

Obecnie zagadnienia wchodzące w skład problematyki komputerowego przetwarzania języka naturalnego są w znacznej mierze związane z badaniami nad sztuczną inteligencją i dotyczą rozumienia języka naturalnego przez komputer, komunikacji człowieka z maszyną przy użyciu języka naturalnego (w języku naturalnym i nie tylko), inżynierii (technologii) języka naturalnego (pozyskiwanie zasobów i narzędzi badawczych), formalnego opisu języka naturalnego (algorytmy parsingu, metody heurystyczne), ze szczególnym uwzględnieniem specyfiki języka polskiego.

Komputerowe przetwarzanie tekstów języka naturalnego jest więc dziedziną interdyscyplinarną, z pogranicza lingwistyki, sztucznej inteligencji, informatyki oraz kognitywistyki. Z lingwistyki klasycznej czerpie metody operowania danymi językowymi, modele języka wykorzystywane m.in. do przybliżania treści i znaczenia analizowanego tekstu oraz prawa językowe, głównie statystyczne, wykorzystywane m.in. w wyszukiwaniu informacji oraz w automatycznym klasyfikowaniu dokumentów, a także formalne metody tagowania poszczególnych elementów języka. Informatyka, a szczególnie sztuczna inteligencja, dostarcza metod i narzędzi automatycznego przetwarzania i analizowania dużych ilości danych językowych, algorytmów wyszukiwania podobieństw bądź prawidłowości statystycznych w dużych zbiorach oraz mechanizmów przechowywania i operowania na danych oraz metadanych. Nauki kognitywne oferują metody przybliżania znaczenia tekstu, pomagając w tworzeniu systemów rozumiejących treść i kontekst (ang. *natural language understanding*, NLU). Od ponad 50 lat niezrealizowanym wyzwaniem jest wyposażenie w kompetencję językową urządzeń wytworzonych przez człowieka, w tym komputera [20].

3. Geneza i rozwój badań nad przetwarzaniem języka naturalnego

Genezy przetwarzania języka naturalnego[1] można doszukiwać się już w latach 40. XX w., kiedy to w USA podjęto pierwsze (jakkolwiek nieskuteczne) próby automatycznego tłumaczenia tekstów. W latach 50. XX w. rozpoczęto przetwa-

[1] Obszerne omówienie rozwoju badań nad przetwarzaniem języka naturalnego zawiera artykuł Piotra Malaka: *Indeksowanie treści. Porównanie skuteczności metod tradycyjnych i automatycznych*. Warszawa 2012.

rzanie danych w postaci wyrażeń języka naturalnego dla celów wyszukiwania informacji, klasyfikacji i selekcji informacji w dużych zbiorach. Do końca lat 80. XX w. rozwijały się dwa niezależne nurty przetwarzania języka naturalnego: analiza statystyczna oraz gramatyki generatywne. W pierwszym nurcie mieści się wyszukiwanie informacji i dokumentów (ang. *Information Retrieval*, IR) spełniających zadane kryteria treściowe. IR jest jednym z najstarszych zastosowań automatycznego przetwarzania danych językowych i jedną z tzw. metod statystycznego nurtu NLP, polegającego na opracowywaniu frekwencyjnym tekstu. Obecne metody statystyczne stają się niewystarczające. Język naturalny nie jest przecież językiem logicznym, stąd niezbędne są języki formalne. Charakterystyczne dla kierunku formalnego gramatyki generatywne bazują głównie na teorii automatów Alana Turinga oraz pracach Noama Chomsky'ego dotyczących gramatyk formalnych i generatywnych.

Od końca lat 80. XX w. duże znaczenie mają metody inżynierii języka NLP, oparte najczęściej na wcześniej odpowiednio przygotowanych korpusach reprezentatywnych tekstów dla poszczególnych języków. Buduje się odpowiednie algorytmy do wykrywania znaczenia w tekście wykorzystujące słowniki rozpoznające wzorce i analizujące częstotliwość wystąpień wyrazów w tekście.

Pytania o rolę i możliwości wykorzystania języka naturalnego w wyszukiwaniu informacji były stawiane za granicą od dawna (np. W. J. Hutchins, F.W. Lancaster, G. Salton, K. Jones Spärck), ale i w Polsce, na Uniwersytecie Warszawskim: B. Bojar, O. A. Wojtasiewicz, J. S. Bień, S. Szpakowicz, K. Szafran; w IPI PAN Warszawa: A. Przepiórkowski, A. Kupiść, A. Marciniak, A. Mykowiecka; UJ/AGH: W. Lubaszewski; Politechnika Wrocławska: M. Piasecki; UAM w Poznaniu: Z. Vetulani, J. Martinek, G. Vetulani, J. Marciniak, J. Daciuk, T. Obrębski oraz A. Wakulicz-Deja, M. A. Kłopotek,

Warto zwrócić uwagę na istnienie polskich opracowań nurtu informatologicznego. Opracowania te powstały głównie w ramach ówczesnego IINTE, na przykład opracowania I. Szymanowskiej, H. Dryzek czy J. Solaka. Większość tych prac jest już wprawdzie przestarzała, ale ilustrują one wkład nauki o informacji w Polsce do problematyki przetwarzania języka naturalnego, stanowiąc zarazem ich dokumentację. Na te problemy ostatnio w nauce o informacji zwracali uwagę m.in. S. Kurek-Kokocińska, B. Bojar, W. Babik, A. Pawłowski, P. Malak, P. Nowak i P. Nowakowski[2].

W Polsce już działają korpusy tekstów języka polskiego: Korpus języka polskiego IPI PAN, Korpus referencyjny języka polskiego PELCRA, Narodowy Korpus Języka Polskiego, Korpus Języka Polskiego Wydawnictwa Naukowego PWN. Znajdują one wykorzystanie w takich projektach realizowanych przez polskich badaczy jak: Słowosieć (plWordNet)[3] – sieć relacji semantycznych, struktury wykorzystywane do odtwarzania składni zdania (parsery, czyli ana-

[2] Por. P. Nowak, P. Nowakowski: *Infolingwistyka jako forma integracji językoznawstwa z nauką o informacji*. W: *Studia nad językiem, informacją i komunikacją*. Pod red. W. Krzemińskiej i P. Nowaka. Poznań 2003, s. 193-203.

[3] Słownik został stworzony przez badaczy z Politechniki Wrocławskiej z Grupy Technologii Językowych G4.19. Zawiera 160 000 jednostek leksykalnych i 350 000 relacji leksykalnych. Każda relacja jest opisana linkiem, dzięki czemu Słowosieć jest słownikiem interaktywnym zarówno dla użytkowników, jak i dla programów komputerowych.

lizatory składniowe), analizatory morfologiczne języka polskiego: SAM, LEM, GRAM, AMOR, PoMOR, Xelda, Morfeusz SIAT, Morfologik.

Obecnie najważniejsze dla informacji problemy przetwarzania języka naturalnego to:

– Automatyczna ekstrakcja informacji z dużych zbiorów tekstów. Podstawową trudność stanowi m.in. definicja informacji niesionej przez ludzką wypowiedź; przydatne tu są różne koncepcje wartości informacyjnej słowa (wartość informacyjna słowa H. P. Luhna czy wartość informacyjna dokumentu).

– Metody automatycznego indeksowania dokumentów:
  – indeksowanie statystyczne – wykorzystujące statystyczne właściwości wyrazów lub wyrażeń występujących w tekście dokumentu w aspekcie danego dokumentu lub danego korpusu; wykorzystywana jest tzw. wartość informacyjna słowa H. P. Luhna, która jest funkcją jego częstości (krzywa Gaussa opisuje gęstość prawdopodobieństwa zdarzeń w rozkładzie normalnym);
  – indeksowanie probabilistyczne – wykorzystujące rachunek prawdopodobieństwa w celu określenia prawdopodobieństwa wyszukania dokumentu relewantnego oraz wykorzystujące rozkład częstości terminów w celu określenia tego prawdopodobieństwa;
  – indeksowanie lingwistyczne/syntaktyczne – wykorzystujące automatyczną analizę językową w celu wyróżnienia w tekście dokumentu wyrażeń informacyjnie ważnych, znaczących dla jego treści; podstawową metodą jest tu teoria języków formalnych, zwłaszcza tzw. zbiór znaczników frazowych N. Chomsky'ego.

Przy wykorzystywaniu języka naturalnego w wyszukiwaniu informacji podstawowym wymaganiem jest możliwość stosunkowo prostej automatyzacji tego procesu. Tu reprezentacje wyrazów muszą pochodzić z naturalnego kontekstu ich użycia, a nie z sytuacji laboratoryjnej. Zawsze istnieje jakiś wpływ związków asocjacyjnych na reprezentację tych wyrazów. Innym problemem jest interpretacja wyrazów mających wiele znaczeń. Bliskość semantyczna zależy bowiem od wielu czynników, m.in. od niejęzykowych struktur poznawczych. Tymczasem dane z korpusu językowego mogą/i różnią się zazwyczaj zawsze od doświadczenia językowego. Stworzona na podstawie korpusu przestrzeń (informacyjna) jest czymś w rodzaju „kolektywnej" przestrzeni semantycznej społeczności posługującej się danym językiem. Ta „kolektywna" przestrzeń wcale nie musi odpowiadać „przestrzeni indywidualnej". Chodzi o to, aby te przestrzenie jak najbardziej zbliżały się do siebie. W dobie blogów można próbować zejść na poziom jednostki i próbować kontrolować wielowymiarowe przestrzenie semantyczne na podstawie tekstów pochodzących od jednego człowieka.

4. Wyszukiwanie informacji

„Wyszukiwanie informacji (IR) jest znajdowaniem materiału (najczęściej dokumentów) w postaci niestrukturalnej (zazwyczaj tekstu) w dużych zbiorach (zazwyczaj przechowywanych komputerowo), które zaspokajają potrzeby

informacyjne." [13]. Metody IR przeciwstawiane są modelowi wyszukiwania strukturalnego, stosowanego najczęściej w bazach danych. Wyszukiwanie w zbiorach informacji strukturalnej wymaga znajomości struktury wykorzystanej do przechowywania danych, przeznaczenia poszczególnych pól oraz powiązań zachodzących pomiędzy elementami rekordu. Proces wyszukiwania polega tu m.in. na wskazaniu pola, którego zawartość ma zostać porównana z zapytaniem, oraz sposobu bądź metody porównawczej, jest więc tylko dostępny dla osób przeszkolonych w wyszukiwaniach tego typu.

Autorzy przytoczonej definicji terminu „wyszukiwanie informacji" trafnie wskazują, że przeszukiwanie pełnotekstowe uniezależnia systemy informacyjno-wyszukiwawcze od danych przechowywanych w postaci strukturalnej. Pozwala to na przechowywanie dokumentów w postaci tekstu, bez tworzenia i wypełniania treścią specjalnych pól jak w systemach bazodanowych. Innym zastosowaniem jest możliwość wyszukiwania łącznego w różnych elementach formalnego opisu dokumentu, na przykład w tytule oraz w treści. IR można również stosować do filtrowania i grupowania dokumentów w zbiorze w zależności od ich zawartości. W tym zakresie znaczące osiągnięcia ma SIGIR[4].

Metody NLP sprawdzają się przede wszystkim w operacjach na pełnych tekstach, takich jak wyszukiwanie informacji, automatyczna klasyfikacja treści czy wskazywanie dokumentów podobnych do siebie treściowo.

Typowe etapy przetwarzania języka naturalnego w systemie informacyjno-wyszukiwawczym [9] to:

– rozpoznawanie mowy (ang. *speech recognition*) – zamiana dźwięku na zapis tekstu, gdy nośnikiem wypowiedzi jest mowa,

– tokenizacja i segmentacja – wydzielenie w tekście podstawowych niepodzielnych jednostek, tzw. tokenów oraz podział tekstu na bloki strukturalne, np. zdania; tokenizacja stanowi szczególny przypadek segmentacji,

– analiza morfosyntaktyczna – formalny opis poszczególnych tokenów pod względem ich własności składniowych, rozpoznanie form wyrazowych jako realizacji poszczególnych leksemów,

– ujednoznacznienie sensu słów (ang. *sense disambiguation*) – rozstrzygnięcie niejednoznaczności w przypisaniu znaczenia leksemu do tokenu,

– analiza składniowa – przypisanie poszczególnym wyrażeniom językowym jednej lub więcej struktur składniowych, na przykład w postaci drzewa rozbioru składniowego,

– analiza semantyczna – przejście od struktury leksykalno-składniowej do pewnej formy reprezentacji znaczenia poszczególnych wyrażeń językowych – przypisanie wyrażeniom językowym wyrażeń pewnego języka formalnego,

– analiza dyskursu – analiza powiązań znaczeniowych pomiędzy poszczególnymi wyrażeniami językowymi, pragmatycznej struktury wypowiedzi, pełnego znaczenia wypowiedzi w relacji do kontekstu itd.

Dziedziną przydatną w nauce o informacji jest lingwistyka informacyjna (infolingwistyka), której zakres jest lokowany na styku nauki o informacji i badań lingwistycznych. Powodem jej uprawiania jest potrzeba dostosowania

[4] SIGIR – Special Interest Group of Information Retrieval. Grupa ta publikuje specjalne raporty z warsztatów organizowanych w różnych miejscach.

sposobów prezentacji informacji do oczekiwań jej odbiorców, niezbędność zmian w interfejsach wyszukiwawczych oraz konieczność rezygnacji z kontroli semantycznej na rzecz słownictwa swobodnego, niekontrolowanego, czyli słów kluczowych, i/lub wyszukiwania pełnotekstowego.

Przetwarzanie języka naturalnego na potrzeby działalności informacyjnej przejawia się w następujących formach:

– indeksowanie dokumentów (specjaliści – indeksatorzy);

– społeczne opisywanie treści dokumentów (tagowanie) czyli wskazywanie słów kluczowych przez odbiorców treści;

– automatyczne tworzenie surogatów dokumentów, zwane również automatycznym indeksowaniem/streszczaniem.

Oczekiwania nauki o informacji od dziedziny przetwarzania języka naturalnego dotyczą posługiwania się językiem naturalnym w jego etnicznych odmianach w kodzie fonicznym, a przynajmniej graficznym, przekładalności (wyszukiwanie w różnych językach etnicznych i przekład maszynowy), docierania do mikroinformacji (wyszukiwanie pełnotekstowe).

5. Perspektywy dalszych badań i prac wdrożeniowych

W przyszłości będzie dominowało wyszukiwanie informacji za pomocą słownictwa języka naturalnego, zapewne w znacznie szerszym zakresie niż obecnie. W niezbyt odległej przyszłości będziemy porozumiewać się z komputerem za pomocą mowy. Do tego jest niezbędne zbudowanie algorytmów rozumienia mowy. Aby swobodnie porozumiewać się z komputerem, trzeba rozwiązać szereg problemów dotyczących komputerowej interpretacji tekstu. Oczekiwania, że słowa kluczowe bądź słowa z tekstu będą wystarczającym automatycznym narzędziem wyszukiwawczym w bazach danych i w zasobach internetowych, spełniają się w coraz większym stopniu. Dostęp do nich będzie efektywniejszy, gdy w tworzeniu charakterystyk wyszukiwawczych w większym zakresie stosowane będą relacje intertekstualne, ponieważ umożliwiają prowadzenie wyszukiwań zarówno szerszych, jak i węższych, dodając konteksty do używanych terminów wyszukiwawczych. Język naturalny w różnych formach będzie stopniowo wypierać inne języki stosowane w wyszukiwaniu informacji.

W obszarze automatycznego indeksowania nie zostały do tej pory rozwiązane wszystkie problemy automatycznej analizy tekstu i ekstrakcji słów kluczowych. Badania te są nadal na etapie wstępnym. Przyśpieszyłoby je opracowanie metod generowania tzw. wiedzy niejawnej (*implicite knowledge*). Głównym wyzwaniem jest opracowanie metod pozwalających na odczytywanie znaczenia na poziomie pełnotekstowym. Pewien postęp, przynajmniej w odniesieniu do rozumienia znaczenia terminów i małych fragmentów dokumentów, stanowi wiązanie technik stosowanych w badaniach nad wyszukiwaniem informacji ze sztuczną inteligencją (sieciami neuronowymi) i lingwistyką komputerową. Cel, jaki stoi przed tymi badaniami, to wypracowanie semantyki pełnego tekstu. Jest on jeszcze daleki do osiągnięcia. Włożony wysiłek oraz koszty formalizacji analizy składniowej i uwzględnienia wartości informacyjnych tak naprawdę są niewspółmierne do uzyskanych efektów, dlatego trzeba zdać się na rozwiązania informatyków.

O zakresie wykorzystania języka naturalnego w wyszukiwaniu informacji decyduje a jednocześnie ogranicza go poziom rozwoju technologii komputerowej. Automatyczne indeksowanie może znacząco wspomóc indeksowanie intelektualne, zwłaszcza że gigantycznie zwiększa się liczba zróżnicowanych strukturalnie, nieustrukturalizowanych, heterogenicznych i nieustannie zmiennych zasobów internetowych.

Szybszy postęp w tym zakresie będzie możliwy dzięki programom komputerowym wykorzystującym efekty formalizacji języka naturalnego w obrębie semantyki i relacji paradygmatycznych, co umożliwi wyszukiwanie na pożądanym poziomie szczegółowości-ogólności, relewancji, dokładności i kompletności informacji.

Przeprowadzone do tej pory badania potwierdziły to, co już wcześniej zauważono, że pomimo znacznego rozwoju metod, narzędzi i technologii komputerowych aplikowanych w przetwarzaniu języka naturalnego na potrzeby wyszukiwania informacji, przynajmniej na razie pełna automatyzacja tego procesu nie jest możliwa, chociaż technologie przetwarzania języka posunęły się znacznie do przodu. Przed badaczami tej problematyki jest więc jeszcze długa droga do pełnej automatyzacji tego typu procesów oraz perspektywa, że na razie człowieka wyeliminować się nie da. Uzyskane wyniki badań pozwalają twierdzić, że proces automatycznego generowania charakterystyk wyszukiwawczych dokumentów może i powinien znacząco wspomagać proces kognitywnego (wykonywanego przez człowieka) opracowania rzeczowego dokumentów i informacji [2].

Podsumowanie

Celem prac badawczych NLP jest zarówno przetwarzanie, jak i generowanie wyrażeń językowych. Jednym z zadań nauki o komputerowym przetwarzaniu języka naturalnego jest usprawnienie tworzenia charakterystyk wyszukiwawczych dokumentów w systemach informacyjno-wyszukiwawczych. Jego praktyczna realizacja sprowadza się m.in. do wyposażenia komputerów w narzędzia pozwalające na „zrozumienie" podawanych przez użytkowników w języku naturalnym (zarówno w formie pisanej, jak i mówionej) kwerend oraz umożliwiające generowanie odpowiedzi zrozumiałych i sensownych dla człowieka.

Informacja naukowa w Polsce wytworzyła już fachowe i teoretyczne zaplecze, a więc posiada wartościowy potencjał, który oprócz osiągnięć dyscyplin pokrewnych może być wykorzystany w pracach nad przetwarzaniem języka naturalnego na potrzeby systemów wyszukiwania informacji. Istnieje zatem pilna potrzeba aktywnego włączenia się specjalistów w zakresie nauki o informacji w nurt prac nad przetwarzaniem języka naturalnego.

Współczesne metody przetwarzania języka naturalnego, próbują integrować się z narzędziami wyszukiwania informacji nie tylko dokumentacyjnych systemów informacyjno-wyszukiwawczych. Ciągle trwają prace nad uniwersalnym wyszukiwaniem, które mogłoby pozwolić na łączenie potencjału ludzkiego i sztucznej inteligencji, tworzyć bogate semantycznie środowiska i oferować łatwiejszy dostęp do informacji elektronicznej, nie tylko dla ekspertów i fascy-

natów, lecz także dla osób, dla których nowoczesne technologie komunikacyjne są jeszcze względnie obce [10, 8]. Rozwiązywanie tak złożonych problemów wymaga integracji dotychczasowych osiągnięć badawczych różnych dyscyplin naukowych, w tym zauważenia potencjału tkwiącego w nauce o informacji.

Bibliografia

1. Babik W.: *Generowanie języków informacyjno-wyszukiwawczych ze słowników terminologicznych*. Kraków 1996.
2. Babik W.: *Słowa kluczowe*. Kraków 2010.
3. Belkin N.J.: *Some(what) Grand Challenges for Information Retrieval*. [online]. [dostęp: 1.05.2013]. Dostępny w World Wide Web <http://www.sigir.org/forum/2008 J/2008j--sigirforum-belkin.pdf>.
4. Bojar B.: *Językoznawstwo dla studentów informacji naukowej*. Warszawa 2005.
5. Chowdhury G.: *Natural language processing*. „Annual Review of Information Science and Technology". 2003 vol. 37, pp. 51- 89.
6. Dobrowolski Z., Nicholas D.: *Informacyjny gracz: nowa koncepcja użytkownika informacji*. „Praktyka i Teoria Informacji Naukowej i Technicznej" 2001 nr ½, s. 4-9.
7. Dura E.: *Natural Language in Information Retrieval*. In: *Computational Linguistics and Inteligent Text Processing. Lectures Notes in Komputer Science*. 2003 vol. 2588, pp. 537-540.
8. Gontar B., Papińska-Kacperek J.: *Semantyczne wyszukiwarki internetowe*. „Acta Universitatis Lodziensis. Folia Oeconomica" 2011 vol. 261, s. 165 -179.
9. Jurafsky D., Martin J. H.: *Speech and language processing. An introduction to Natural Language Processing Computational Linguistics and Speech Recognition*. New Jersey 2000.
10. Kłopotek M. A.: *Inteligentne wyszukiwarki internetowe*. Warszawa 2001.
11. Malak P.: *Indeksowanie treści. Porównanie skuteczności metod tradycyjnych i automatycznych*. Warszawa 2012.
12. Malak P.: *Rozwój badań nad przetwarzaniem języka naturalnego*. „Zagadnienia Informacji Naukowej" 2010 nr 2(96), s. 21- 30.
13. Manning Ch. D., Raghavan P., Schutze H.: *An introduction to information retrieval.* Cambridge University Press 2009. [online]. [dostęp: 1.05.2013]. Dostępny w World Wide Web. <Http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>.
14. Murthy Kavi Narayana: *Natural language processing. An information access perspective.* Bangalore 2005.
15. Nowak P., Nowakowski P.: *Infolingwistyka jako forma integracji językoznawstwa z nauką o informacji.* W: *Studia nad językiem, informacją i komunikacją*. Pod red. W. Krzemińskiej i P. Nowaka. Poznań 2003, s. 193-203.
16. Piasecki M.: *Cele i zadania lingwistyki informatycznej*. W: *Metodologie językoznawstwa. Współczesne tendencje i kontrowersje*. Pod red. P. Stelmaszczyka. Kraków 2008, s. 258-290.
17. *Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych*. Oprac. B. Bojar. Warszawa 2002.
18. *Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu.* Pod red. W. Lubaszewskiego. Kraków 2009.
19. Świdziński M.: *Lingwistyka korpusowa w Polsce – źródła, stan, perspektywy*. „LingVaria" 2006 nr 1 s. 23-24. [online] [dostęp: 1.05.2013]. Dostępny w World Wide Web.<http://www2.polonistyka.uj.edu.pl/LingVaria/arcyhiwa/LV_1_2006_pdf/02_Swidzinski.pdf>.
20. Vetulani Z.: *Human Language Technologies: Tradition and New Challenges*. „Proceedings of Artificial Intelligence". 2005, vol. 2(25), pp. 5-31.
21. Vetulani Z.: *Komunikacja człowieka z maszyną. Komputerowe modelowanie kompetencji językowej.* Warszawa 2004.

Summary

Man uses primarily a natural language in the information space, but not only. There are also information retrieval languages (indexing languages) whose future is rather unsure, mainly because of the current more and more common tendencies to retrieve information in the indexing systems with natural languages. The object of my paper is natural language in information retrieval. Information retrieval is one of the basic functions of a natural language processing. This paper is intended to offer an answer to the question about the status of advancement and prospects of further works on the use of natural languages in the information retrieval process applied in indexing and retrieval systems, including the Internet, based on the existing literature. My answer to that question may become a starting point for further considerations on a broader inclusion of Polish information scientists in the course of general research and application studies.