ROZWÓJ BADAŃ NAD PRZETWARZANIEM JĘZYKA NATURALNEGO

Piotr Malak
Instytut Informacji Naukowej i Bibliologii
Uniwersytet M. Kopernika w Toruniu

Przetwarzanie języka naturalnego, metody statystyczne NLP, gramatyki generatywne, wyszukiwanie informacji

Rosnąca popularność cyfrowych form przechowywania i przekazywania informacji powoduje, że konieczne staje się opracowywanie nowych, wydajnych metod ułatwiających zarządzanie czy też wyszukiwanie takiej informacji. Z informacją cyfrową związane są częściowo lub w całości m.in. nowe formaty opisu dokumentu (np. Dublin Core), biblioteki cyfrowe czy też wyszukiwarki sieciowe.

Ogromną zaletą źródeł elektronicznych w porównaniu do ich tradycyjnych poprzedników jest możliwość swobodnego przeszukiwania tekstu. Użytkownik dokumentu elektronicznego za pomocą specjalnego interfejsu może przeszukać całą treść dokumentu dla wybranych haseł czy słów kluczowych, tworzonych w sposób dowolny, całkowicie swobodnie. Nie jest już ograniczony tylko do haseł klasyfikacji rzeczowej, których wprawne stosowanie wymaga często odpowiedniego szkolenia.

Jednymi z najbardziej zaawansowanych systemów informacyjno-wyszukiwawczych są wyszukiwarki internetowe. Są to specjalistyczne systemy indeksujące treść dokumentów dostępnych online i zwracające użytkownikowi listy dokumentów spełniających zadane przez niego kryteria wyszukiwania. Przy czym w systemach zaawansowanych technologicznie dopasowanie treści dokumentu odbywa się z uwzględnieniem zasad gramatyki danego języka. Analizowane automatycznie są wszystkie prawidłowe formy gramatyczne wyrazów tworzących zapytanie użytkownika.

Od początków stosowania komputerów rozwija się dziedzina badań nad automatycznym przetwarzaniem języka naturalnego (ang. NLP – *Natural Language Processing*). Dziedzina ta dostarcza rozwiązań zarówno teoretycznych, na poziomie lingwistyki, jak i praktycznych zastosowań wykrytych prawidłowości i zależności lingwistycznych. Jako przykłady można podać moduły korekty ortograficznej i gramatycznej, systemy automatycznej translacji czy wspomniane wcześniej wyszukiwarki sieciowe.

Artykuł ma na celu wprowadzenie czytelnika w powstanie i rozwój tej niezwykle dynamicznej i interdyscyplinarnej dziedziny.

Początki NLP

Podstawy teoretyczne dla NLP przygotował m.in. Alan Turing, opracowując w 1936 r. teorię automatu. Teorię tę rozwinął następnie w latach 50. XX w. Stephen C. Kleene, wzbogacając ją o pojęcia automatu skończonego oraz zbiorów regularnych. Kolejnym badaczem, który wniósł duży wkład w rozwój nowego kierunku badań, był Claude E. Shannon. Dostosował on modele Markova do tworzenia modeli lingwistycznych oraz opracował koncepcję entropii w teorii informacji, a także pojęcie ilości informacji. NLP zaadaptowało również teorię języków formalnych i gramatyk Noama Chomskyego z 1956 r. Na podstawie prac tych badaczy wypracowano wiele teorii opisu języka oraz analizy języków naturalnych i sztucznych[1].

Początki przetwarzania języka naturalnego jako dziedziny badań naukowych związane są w oczywisty sposób z początkiem ery komputerów, maszyn dysponujących mocą obliczeniową wystarczającą do automatycznego przeprowadzania operacji na danych. Już w latach 40. XX w. w Stanach Zjednoczonych podjęto próby automatycznego tłumaczenia tekstów. Próby te nie były zbyt skuteczne i szybko z nich zrezygnowano, wskazując jednocześnie inne obiecujące kierunki badań NLP.

W latach 50. XX w. stosowano przetwarzanie danych w postaci wyrażeń języków naturalnych dla celów wyszukiwania, klasyfikacji i selekcji informacji w dużych zbiorach. Do końca lat 80. minionego stulecia w badaniach NLP rozwijały się dwa niezależne trendy: analiza statystyczna oraz gramatyki generatywne. Metody statystyczne wykorzystywane są szeroko do wyszukiwania dokumentów w dużych zbiorach (nurt IR – *Information Retrieval*). W analizie statystycznej poszczególne słowa stanowiące treść analizowanych dokumentów tworzą zbiór wspólny, z zachowaniem informacji o częstości wystąpień danego słowa w zbiorze. W literaturze angielskiej zbiór ten nazywany jest *bag-of-words*, w piśmiennictwie polskim można spotkać powtórzenie nazwy angielskiej lub wyrażenia typu *wielozbiór.* W metodach statystycznych słowa kluczowe dla poszczególnych dokumentów wskazywane są z reguły na podstawie porównania częstości ich wystąpienia w danym dokumencie z liczbą wystąpień w całym zbiorze. Podejście takie cechuje się łatwością implementacyjną, relatywnie niskimi kosztami operacyjnymi, wysoką skutecznością wyszukiwania dokumentów oraz niezależnością od konkretnego języka naturalnego. Alternatywny nurt bazował na teorii automatów A. Turinga oraz pracach N. Chomsky'ego dotyczących gramatyk formalnych i generatywnych. Tworzone w ich wyniku gramatyki wykorzystywane były następnie do odtwarzania struktury zdania w dowolnym wyrażeniu. Do zalet takiego podejścia zalicza się głównie sformalizowanie i pogłębienie wiedzy lingwistycznej. Praktycznym zastosowaniem wyników badań nad gramatykami są np. parsery, czyli analizatory składniowe[2].

Obie metody nie są wolne od wad. Podejście statystyczne pomija znaczenie analizowanej treści, nie są rozpoznawane ani rejestrowane związki pomiędzy

---

[1] Ref. za: D. Jurafsky, J. H. Martin: *Speech and language processing. An introduction to Natural Language Processing, Computational Linguistics and Speech Recognitio.* New Jersey 1999, s. 10-11.

[2] Por. Ibidem, s. 12-15.

wyrazami, dlatego metody takie najlepiej sprawdzają się w wyszukiwaniu informacji na podstawie wystąpienia haseł wyszukiwawczych. Natomiast w przypadku metod bazujących na gramatykach i formalizmach problemem są koszty (obliczeniowe oraz czasowe) przygotowania poprawnej gramatyki oraz analizy treści dokumentu. Rezultatem zastosowania przygotowanych gramatyk jest duży zbiór możliwych interpretacji analizowanych zdań, których ujednoznacznienie często nie jest możliwe na drodze wyłącznie automatycznego przetwarzania.

Od końca lat 80. XX w. coraz większego znaczenia nabierają metody statystyczne NLP. Wiąże się to z dostępnością odpowiednio przygotowanych korpusów tekstów reprezentatywnych dla danego języka, co zwiększa wartość analiz statystycznych. Dzięki wykorzystaniu adnotowanych korpusów w opracowaniu frekwencyjnym tekstu można wykorzystać nie tylko lokalne, typowe dla danego dokumentu czy zboru wartości, ale również uwzględnić związki formalne pomiędzy poszczególnymi wyrazami[3].

Etapy rozwoju badań nad przetwarzaniem języka naturalnego

W rozwoju badań nad przetwarzaniem języka naturalnego wyodrębnić można kilka następujących etapów:

1) do 1957 r. można wskazać następujące osiągnięcia:
   - model obliczeń – na podstawie pracy A. Turinga (Turing, A., *On computable numbers, with an application to the Entscheidungsproblem,* [w:] *Proceedings of the London Mathematical Society,* 2(42), s. 230-165),
   - zastosowanie modeli Markova do analizy języka (w pracy Shannon, C. E., *A mathematical theory of communication*, [w:] *The Bell Systems Technical Journal,* 27, 279-423, 623-656),
   - wprowadzenie automatów skończonych oraz wyrażeń regularnych,
   - zastosowanie automatów do reprezentowania gramatyk (Chomsky, N., *Three models for the description of language* [w:] *IRE Transactions of Information Theory*, 2(3)),
   - wprowadzenie formalnego opisu języka oraz gramatyk bezkontekstowych (głównie N. Chomsky, op. cit., ale też Backus, J. W., *The syntax and the semantics of the proposed international algebraic of the Zürich ACM-GAMM Conference.* [w:] *Proc. of the Inf. Conf. on Information Processing,* Paris 1959; oraz Naur., P, et all., *Report on the algorithmic language algol 60.* [w:] *Communications of the ACM,* 1960),
   - zdefiniowanie entropii jako miary ilości informacji (Shannon, C. E., *Prediction and entropy of printed English* [w:] *The Bell Systems Technical Journal, 30, 1951);*
2) lata 1957-1970 – wtedy wykrystalizowały się dwa, konkurencyjne podejścia do przetwarzania języka naturalnego:
   a) metody symboliczne (formalne), do których można zaliczyć:
      - gramatykę generatywną,
      - parsery syntaktyczne,

[3] O trendach w NLP por. m.in. A. Przepiórkowski: *Powierzchniowe przetwarzanie języka polskiego*. Warszawa 2008, s. 9-11.

- metody sztucznej inteligencji;

b) metody statystyczne, a wśród nich:
- metody Bayes'a,
- optyczne rozpoznawanie liter (ang. *Optical Character Recognition*, OCR),
- identyfikacja autorów tekstów,
- tworzenie korpusów;

3) lata 1970-1983 – wypracowano cztery paradygmaty NLP:
   a) modele statystyczne: rozpoznawanie mowy, synteza mowy; Ukryte Modele Markowa (ang. *Hidden Markov Models*, HMM),
   b) logika formalna (język Prolog; Definite Clasue Grammar, DCG; teoria Lexical-Functional Grammar, LFG),
   c) rozumienie języków naturalnych,
   d) modelowanie dyskursu;
4) lata 1983-1993 – odrodzenie modeli skończonych stanów oraz empiryzmu:
   a) morfologia i fonologia za pomocą modeli skończonych,
   b) modele skończonych stanów składni,
   c) metody stochastyczne wykraczające poza rozpoznawanie mowy (IBM);
5) lata 1993 – obecnie – integracja dotychczasowych osiągnięć i metod:
   a) metody statystyczne w symbolicznych metodach analizy języka na wszystkich poziomach,
   b) systemy ekstrakcji informacji,
   c) komercyjne zastosowanie wyników badań (na komputerach domowych): rozpoznawanie mowy, korektory ortograficzne i gramatyczne[4].

Wybrane kierunki działań przetwarzania języka naturalnego

Zestawienie historyczne zaprezentowane powyżej prezentuje bogaty zasób wypracowanych na potrzeby przetwarzania języka naturalnego metod i narzędzi badawczych oraz ukazuje, że ta młoda dziedzina rozwijała się od samych początków bardzo prężnie. Poniżej zostaną zaprezentowane wybrane, popularne i sprawdzone metody NLP związane z tematyką niniejszej pracy. Metody te należą do nurtu statystycznego, który (przypomnijmy) cechuje się stosunkowo niskimi kosztami operacyjnymi analizy. Jednym z najstarszych zastosowań automatycznego przetwarzania danych językowych jest wyszukiwanie informacji w dokumentach tekstowych, kolejnym, wnoszącym przydatne rozwiązania, jest grupowanie dokumentów.

Wyszukiwanie informacji w dokumentach

W pracy *An introduction to information retrieval* jej autorzy w następujący sposób definiują pojęcie w y s z u k i w a n i e i n f o r m a c j i :

[4] Dane do zestawienia za: A. Mykowiecka: *Inżynieria lingwistyczna. Komputerowe przetwarzanie tekstów w języku naturalnym*. Warszawa 2007, s. 17-18, 329-343; por. również D. Jurafsky, J. H. Maritn: *Speech…*, op. cit., s. 10-15.

Wyszukiwanie informacji (IR) jest znajdowaniem materiału (najczęściej dokumentów) w postaci niestrukturalnej (przeważnie tekstu) w dużych zbiorach (zazwyczaj przechowywanych komputerowo), które zaspokajają potrzeby informacyjne[5].

Ch. Manning i pozostali autorzy zdefiniowane tak wyszukiwanie informacji przeciwstawiają modelowi wyszukiwania strukturalnego, stosowanego najczęściej w bazach danych (m.in. relacyjnych) lub w zautomatyzowanych katalogach bibliotecznych. Wyszukiwanie w zbiorach informacji strukturalnej wymaga znajomości struktury wykorzystanej do przechowywania danych, przeznaczenia poszczególnych pól oraz powiązań zachodzących pomiędzy różnymi elementami rekordu. Proces wyszukiwania polega między innymi na wskazaniu pola, którego zawartość ma zostać porównana do zapytania oraz sposobu lub metody porównawczej, jest więc dostępny dla osób przeszkolonych w wyszukiwaniach tego typu. Metodologia IR zakłada w pełni swobodne przeszukiwanie pełnego tekstu oraz ewentualnych pól metainformacyjnych dokumentu. Zapytania budowane są zazwyczaj w postaci listy słów bądź wyrażeń kluczowych opisujących informacje, na których zależy użytkownikowi. Wynikiem takiego wyszukiwania jest lista dokumentów zawierających wskazane w zapytaniu wyrażenia. W przypadku wyszukiwania swobodnego zbiór dokumentów do przeszukania nie musi być wstępnie opracowany, a przeglądanie i porównywanie zawartości dokumentów tekstowych jest procesem w pełni zautomatyzowanym. Pozwala to na obniżenie kosztów samego procesu przetwarzania dokumentów poprzez pominięcie etapu opracowania rzeczowego i formalnego dokumentu. Założenia wyszukiwania informacji (IR) zostały w praktyce zaimplementowane w wyszukiwarkach sieciowych. Intuicyjne interfejsy użytkownika wyszukiwarek oraz możliwość wskazania dokumentów jedynie na podstawie słów kluczowych występujących w treści pozwoliły na swobodne prowadzenie wyszukiwania informacji przez miliony użytkowników Internetu[6].

Kolejna oszczędność kosztów operacyjnych, a jednocześnie racjonalizacja procesu wyszukiwania pełnotekstowego została osiągnięta po wprowadzeniu plików indeksów odwróconych. W plikach tych przechowywana jest informacja o lokalizacji każdego wystąpienia każdego tokenu lub słowa we wszystkich dokumentach kolekcji. Proces odwzorowania lokalizacji poszczególnych słów i tokenów nazywany jest procesem indeksowania. Przeprowadzany jest w momencie akwizycji dokumentu do zbioru. System wyszukiwawczy może odwołać się bezpośrednio do wskazanego przez użytkownika słowa lub wyrażenia kluczowego i w krótkim czasie wskazać wszystkie dokumenty zawierające dane wyrażenie, bez konieczności każdorazowego analizowania treści dokumentów dla poszczególnych zapytań od użytkowników.

Łatwość dostępu oraz zastosowania tej formy wyszukiwania informacji nie jest jedyną zaletą IR. Autorzy pracy *An introduction...* wskazują, że kolejnym

[5] Tłumaczenie własne na podstawie definicji: *Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).* Za Ch. D. Manning, P. Raghavan, H. Schütze: *An introduction to Information Retrieval*. Cambridge 2009 s. 1. [on-line]. [dostęp: 17. 08.2009]. Dostępny w World Wide Web: <http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>.

[6] Ibidem, s. 2-3.

udogodnieniem jest wyszukiwanie pełnotekstowe, które 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 się to dzieje w systemach bazodanowych. Innym wskazywanym zastosowaniem jest możliwość wyszukiwania łącznego informacji w różnych elementach dokumentu, na przykład w tytule oraz w treści. IR można również stosować do filtrowania i grupowania dokumentów w zbiorze w zależności od ich zawartości[7].

Systemy wyszukiwawcze

W tej samej pracy został zaproponowany podział systemów wyszukiwawczych ze względu na skalę operacyjną, sprowadzającą się w tym przypadku do ilości dokumentów w kolekcji. Jedną z wymienionych kategorii są wyszukiwarki internetowe, indeksujące miliony dokumentów. Problemy występujące w wyszukiwaniu na tak ogromną skalę to między innymi pozyskiwanie dokumentów do indeksowania zawartości, przechowywanie zaindeksowanej treści oraz zapewnienie efektywnego przetwarzania uzyskanych w ten sposób danych. Ponadto należy rozważyć specyfikę dokumentów webowych, m.in. hiperłącza, czy też próby zafałszowania treści witryn w celu zwiększenia pozycji na listach rankingowych wyszukiwarek. Kolejną wskazaną kategorią są wyszukiwarki indywidualne, instalowane na komputerach osobistych. Programy tego typu domyślnie w sposób ciągły indeksują zawartość plików na komputerach użytkowników i odwzorowują ją również w plikach indeksów odwróconych. Oprócz aplikacji systemowych do kategorii tej można zaliczyć również programy klienckie poczty elektronicznej. Aplikacje te oprócz wyszukiwania tekstu w treściach wiadomości oferują również funkcje klasyfikacji tekstu (ang. *clustering*) wskazując spam, który dociera na konta poczty elektronicznej. W tym przypadku klasyfikacja odbywa się na podstawie indeksowanej treści wiadomości i wykrycia typowych dla spamu słów kluczowych. Ponadto użytkownik programów pocztowych ma możliwość klasyfikowania pozostałych wiadomości według różnych kryteriów jak nazwa nadawcy, wątki tematyczne, itp. Problemy tego typu wyszukiwania to duża liczba formatów plików, w jakich użytkownicy przechowują swoje dane, zapewnienie ciągłej pracy systemu wyszukiwawczego bez zbytniego obciążania systemu operacyjnego, który powinien cały czas pracować w sposób wydajny i wygodny dla użytkownika oraz efektywne zarządzanie zajmowanym przez indeksy miejscem na dysku twardym. Ostatnią z proponowanych kategorii, mieszczącą się pomiędzy poprzednimi dwiema, są wyszukiwarki korporacyjne. Wyszukiwania korporacyjne zasięgiem wykraczają poza komputery indywidualne, ale nie obejmują całej sieci Internet. Ograniczone są do sieci korporacyjnych, intranetów, bądź domen webowych. W systemach tego typu zbiorem danych są zazwyczaj dokumenty wewnętrzne organizacji, przechowywane w scentralizowanym systemie plików, np. na wydzielonym serwerze plików lub baz danych, ponieważ ten typ wyszukiwania obejmuje również indeksowanie zasobów wewnętrznych baz danych. W przypadku

[7] Ibidem, s. 2.

wyszukiwania korporacyjnego również można spotkać wiele formatów plików i różne źródła danych. Konieczne jest wdrożenie przetwarzania rozproszonego oraz implementacja takiego sposobu indeksowania treści zasobów, który nie będzie blokował dostępu do zasobów innym użytkownikom sieci korporacyjnej[8].

Modele wyszukiwania informacji

Można wyróżnić dwa podstawowe podejścia do wyszukiwania informacji: model logiki Boola (ang. *Boolean Logic Model*, BLM) oraz model rankingowy (ang. *ranked-output model*). W przypadku modelu Boolowskiego zapytanie buduje się ze słów lub fraz połączonych operatorami logicznymi. Metoda ta pozwala wyłonić ze zbioru dokumentów te, których treść spełnia zadany warunek. Model rankingowy pozwala ocenić podobieństwo treści dokumentów z treścią zapytania i utworzyć na tej podstawie listę rankingową dokumentów trafnych. Przy tworzeniu rankingów wykorzystywane są najczęściej następujące modele oceny podobieństwa[9]:

1) model wektorowy (ang. *Vector Space Model*, VSM),
2) model probabilistyczny (ang. *Probabilistic Model*, PM).

Obie metody można połączyć, wyszukując za pomocą algebry Boola dokumenty zgodne z zapytaniem, a następnie, oceniając stopień zgodności, przedstawić je użytkownikowi w postaci listy rankingowej.

Grupowanie dokumentów (*Clustering*)

Celem klasteryzacji dokumentów jest podział analizowanego zbioru na grupy (zwane też klastrami) jednorodne tematycznie, gdzie podstawą podziału jest treść dokumentu. W wyniku tego procesu otrzymuje się podzbiory dokumentów podobnych do siebie, a różniących się od dokumentów w pozostałych podzbiorach.

Autorzy pracy *An introduction to information retrieval* opisują grupowanie jako najpowszechniejszą formę uczenia się nienadzorowanego – zdobywanie wiedzy przez systemy komputerowe bez udziału człowieka. Przypisanie dokumentu do wybranej klasy odbywa się automatycznie, podobnie zresztą jak wskazanie klas tematycznych dla analizowanego zboru dokumentów. Grupowanie przeciwstawiane jest klasyfikacji, gdzie dokumenty przypisywane są do ustalonych *a priori* klas, w związku z czym klasyfikacja nazywana jest uczeniem się nadzorowanym[10].

Podstawą wskazywania klas i przypisywania do nich dokumentów jest wyznaczenie odległości między dokumentami w przestrzeni dwuwymiarowej. Badacze zagadnienia wskazują dwa poziomy grupowania: g r u p o w a n i e

[8] Ibidem, s. 2.

[9] Por. A. Kempa: *Zastosowanie rozszerzonej metodologii wnioskowania na podstawie przypadków – Textual CBR w pracy z dokumentami tekstowymi*. [online]. [dostęp: 10.06. 2009]. Dostępny w World Wide Web: <http://www.swo.ae.katowice.pl/_pdf/221.pdf>.

[10] Ch. D. Manning, P. Raghavan, H. Schütze: *An introduction...*, op. cit., s. 349.

płaskie oraz grupowanie hierarchiczne. Pierwszy rodzaj tworzy klasy niepowiązane ze sobą żadnymi relacjami, natomiast drugi dostarcza układ hierarchiczny klas, ze wskazanymi relacjami zależności między poszczególnymi klasami. Metody grupowania hierarchicznego można podzielić na nieostre, gromadzące (ang. *hard*) oraz ostre, wyróżniające (ang. *soft*). W przypadku ostrych algorytmów grupowania każdy dokument przypisywany jest tylko i wyłącznie do jednej klasy, natomiast algorytmy nieostre mogą przypisać dokument do kilku klas. Klasy powstałe w wyniku grupowania hierarchicznego można przedstawić w postaci dendrogramu (drzewa zależności)[11].

Ze względu na różne dostępne metody przyporządkowywania dokumentów do klas wyróżnia się kilka rodzajów klasyfikacji. Charakterystyki wyszukiwawcze dokumentów przedstawione w postaci wektorów pozwalają przyporządkować konkretne dokumenty do zdefiniowanych, stałych klas. Możliwe jest również automatyczne generowanie klas tematycznych na podstawie dodatkowej analizy podobieństwa pomiędzy dokumentami relewantnymi do zapytania.

Stałe, zdefiniowane wcześniej klasy wykorzystywane są z kolei w procesie klasyfikacji, który z poziomu mechanizmów porównujących ma wiele wspólnego z grupowaniem. W procesie klasyfikacji można wskazać dwa główne nurty: klasyfikację opartą o wzorce oraz bezwzorcową.

Klasyfikacja oparta o wzorce

Jest to najprostszy sposób klasyfikowania dokumentów, polegający na przypisaniu do jednej z ustalonych klas tematycznych. Dla każdej określonej klasy dokumentów należy wskazać wzorce pozwalające przypisać treść do klasy. Metoda ta świetnie sprawdza się w katalogach internetowych, które operują właśnie na zbiorach ustalonych klas. Klasy tematyczne należy określać na tyle elastycznie, żeby można było przypisać do nich dokumenty nie w pełni klasyfikujące się do danej klasy. Ewentualnie można utworzyć klasę INNE, ale nie jest to rozwiązanie eleganckie i w pełni profesjonalne.

Klasyfikacja bezwzorcowa

Pewną alternatywą jest automatyczne tworzenie klas dostosowanych do posiadanej kolekcji dokumentów. Metoda ta nadaje się dobrze do zastosowania w wyszukiwarkach internetowych, ponieważ opiera się na zbiorze niesklasyfikowanych dokumentów. Dopiero na podstawie charakterystyk treściowych oraz rozkładzie częstotliwości podobnych reprezentacji system generuje klasy i przypisuje do nich poszczególne dokumenty.

W przypadku zamkniętych, kontrolowanych systemów informacyjno-wyszukiwawczych dostępny zestaw słów kluczowych jednoznacznie lokalizuje zakres treściowy dokumentu, ułatwiając użytkownikowi wybór najbardziej relewantnego. Wyszukiwarki internetowe pracują w środowisku otwartym, bez obowiązu-

[11] O grupowaniu dokumentów por. Ibidem, s. 349 oraz D. Jurafsky, J. H. Martin: *Speech...*, op. cit., s. 679.

jących powszechnie reguł tworzenia charakterystyk wyszukiwawczych, dlatego w związku z potrzebą standaryzowania i ujednolicenia sposobu komunikacji wyszukiwarki z użytkownikiem (w zakresie prezentacji listy wyników) stosowano różne metody prezentacji treści dokumentów zgodnych z zapytaniem. Ze względów praktycznych (poziomu akceptacji przez użytkowników) najpopularniejszą metodą prezentowania tematyki dokumentu użytkownikowi jest wyświetlenie kilku pierwszych zdań lub kilku zdań sąsiadujących z miejscem zlokalizowania w treści słowa kluczowego z pozycji zwróconych w odpowiedzi na kwerendę.

Podsumowanie

Artykuł miał na celu zapoznanie czytelnika z rozwojem badań i możliwościami zastosowania badań nad przetwarzaniem języka naturalnego. Wyniki tych badań znajdują szerokie zastosowanie we współczesnym przetwarzaniu i zarządzaniu informacją. Aplikacje stosujące wypracowane przez badaczy NLP prawidłowości lingwistyczne są wykorzystywane w świecie biznesu, ale również w coraz większym zakresie w domenie publicznej, na przykład w bibliotekach cyfrowych. Znajomość praw i zależności lingwistycznych może przyczynić się do bardziej precyzyjnego dostarczania użytkownikowi potrzebnej mu informacji, ale też do obniżenia kosztów technicznych, finansowych i czasowych przetwarzania i zarządzania informacją.

Bibliografia

1. Jackson P., Moulinier I.: *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization.* Amsterdam/Philadelphia 2002.
2. Jurafsky D., Martin J. H.: *Speech and language processing. An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition.* New Jersey 1999.
3. Manning Ch. D., Schütze H.: *Foundations of Statistical Natural Language Processin.* Cambridge 1999.
4. Mykowiecka A.: *Przegląd systemów automatycznej generacji tekstów w języku naturalnym.* Warszawa 1987. Prace IPI PAN nr 614.
5. Mykowiecka A.: *Generacja zdań w języku polskim na podstawie reprezentacji ich semantyki.* Warszawa 1988. Prace IPI PAN nr 644.
6. Mykowiecka A.: *Text planning.* Warszawa 1989. Prace IPI PAN nr 665.
7. Mykowiecka A.: *Planowanie struktury tekstu przy wykorzystaniu RTS.* Warszawa 1994. Prace IPI PAN nr 756.
8. Mykowiecka A.: *Inżynieria lingwistyczna. Komputerowe przetwarzanie tekstów w języku naturalnym.* Warszawa 2007.
9. Piasecki M.: *Cele i zadania lingwistyki informatycznej.* W: *Metodologie językoznawstwa. Współczesne tendencje i kontrowersje.* Pod red. P. Stalmaszczyka. Kraków 2007.
10. Przepiórkowski A.: *Powierzchniowe przetwarzanie języka polskiego.* Warszawa 2008.
11. *Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych.* Oprac. B. Bojar. Warszawa 2002.

Summary

The article presents the development of researches in natural languages' processing. It discusses beginnings of these studies, as well as changes in either the research methods, or their range and scope during last 60 years. The author describes two the most popular research methods: statistical analysis and generative grammars. He also evaluates briefly advantages and disadvantages of them both. Additionally, the article presents selected modern trends of NLP activities. Among the most popular current research trends in natural language processing, one can mention information retrieval or documents' grouping.