Systemy Question Answering

Idea, przegląd istniejących rozwiązań, próba porównania.

Yuliya Kandratovich

Paweł Kupiński

- Systemy QA mają być przyszłością dzisiejszych wyszukiwarek
- Pytanie formułowane jest w języku naturalnym
- Odpowiedź jest automatycznie formułowana lub jest to zdanie wybrane z odpowiedniej strony
- Wiele takich systemów można przetestować w sieci

UWAGI HISTORYCZNE

Po raz pierwszy zagadnienia związane z zadawaniem pytań i udzielaniem odpowiedzi ujawniły się w środowisku logiczno-językoznawczym.

Z wczesnych autorów polskich warto odnotować prace Kubińskiego: "Wstęp do logicznej teorii pytań".

W piśmiennictwie światowym najczęściej cytowana jest monografia "The logic of Questions and Answers" autorstwa Belnapa i Steel'a.

UWAGI HISTORYCZNE

Z kolei pierwsze realizacje prototypów systemów umożliwiających użytkownikowi zadawanie pytań i uzyskiwanie odpowiedzi pojawiły się stosunkowo wcześnie w latach sześćdziesiątych równolegle do prac nad tłumaczeniem maszynowym.
Oto kilka przykładów:

- BASEBALL (Green, Wolf, C.Chomsky, Laughery, Univ. California, 1961) - jeden z pierwszych systemów odpowiadających na pytania (reprezentacja wiedzy typu ramowego, analiza gramatyczna w oparciu o prace Harrisa),

UWAGI HISTORYCZNE

- LUNAR (Woods, BBN, 1972) - system konsultowania bazy danych na temat próbek gruntu księżycowego pobranych przez Apollo 11 (Augmented Transition Networks, semantyka proceduralna),

- LADDER (Sacredotti, Sagalowicz, Slocum, SRI, 1977) – system dialogowego dostępu do rozproszonych baz danych (gramatyki semantyczne),

- HAM-ANS (1981-86) i WISBER (1986-89) (Hahn, Hoeppner, Morik, Marburger i inni, Hamburg) – dialog na temat rezerwacji hotelowej w j. niemieckim, dwupoziomowa reprezentacja wiedzy /konceptualna i referencyjna/),

UWAGI HISTORYCZNE

- ORBIS (ok. r. 1983) (Colmerauer, Kittredge) - pytania do prologowej bazy danych na temat planet układu słonecznego, dostęp dwujęzyczny (angielski francuski), moduł polski (Z. Vetulani) z roku 1985.

- Zarówno system ORBIS, jak i rozwijany do dzisiaj na bazie jego polskiego modułu system POLINT, mogą być uważane za dobrze ilustrujące unifikacyjną teorię pytań i odpowiedzi. Wszystkie powyższe przykłady są próbami modelowania rozumienia języka naturalnego (ang. machine understanding, MU).

Wyszukiwanie i ekstrakcja informacji

Rozwój technologii wysunął na bliższy plan zagadnienia uzyskiwania informacji z zasobów zgromadzonych do tej pory głównie w postaci surowej, np. tekstów, obrazów, nagrań, itp. W związku z tym powstały dwa nowe obszary badań o precyzyjnie określonych celach i cechach. Są to: wyszukiwanie informacji (information retrieval; IR) oraz ekstrakcja informacji (information extraction; ER). Te dwie dyscypliny w połączeniu z bardziej tradycyjnymi technikami modelowania rozumienia składają się na nowe wyzwanie będące aktualnie w fazie samo-definiowania się pod nazwą Question&Answering (w skrócie Q&A).

Wyszukiwanie i ekstrakcja informacji

Wyszukiwanie informacji (IR) koncentruje się na problemie znalezienia w dużej populacji dokumentów tych spośród nich, w których znajdują się informacje będące odpowiedzią na kwerendę użytkownika.

Technologia IR jest w pierwszym rzędzie oparta na metodach statystycznych mających ustalić stopień podobieństwa pomiędzy dokumentem a zapytaniem użytkownika. Okazuje się przy tym, że próby poprawienia efektywności przez stosowanie bardziej wyszukanych metod przetwarzania. Języka naturalnego dały, jak na razie, nikłe rezultaty. Obecne techniki powstały jako efekt prób automatyzacji tradycyjnych systemów bibliotecznych stosowanych do tworzenia Referencji bibliograficznych na potrzeby Wyszukiwania książek i czasopism.

Wyszukiwanie i ekstrakcja informacji

Podstawowymi metodami są indeksowanie, czyli wyszukiwanie w tekście terminów, które go reprezentują, oraz uzgadnianie (matching), polegające na wykrywaniu stopnia podobieństwa pomiędzy reprezentacją tekstu a kwerendą. Indeksowanie, w przypadku języka angielskiego, wykorzystuje techniki tokenizacji, wyszukiwania słów gramatycznych i lematyzacji. Zapytania w systemach IR mogą przyjmować różne formy.

Wyszukiwanie i ekstrakcja informacji

Stosowana jest przede wszystkim technika kombinacji boolowskiej terminów, a także model wektorowy, gdzie kwerenda i dokument reprezentowane są jako wektory w przestrzeni wielowymiarowej z metryką (bada się odległość pomiędzy kwerendą i dokumentem).

Inne techniki zapytań oparte są np. na prawdopodobieństwie, na bliskości terminów, na wagach obliczonych na podstawie cech (takich jak ilość wystąpień terminu w tekście). Jeszcze inne, bardziej wyrafinowane, odwołują się do pogłębienia kwerendy przez wygenerowanie nowej na podstawie pierwszych dokumentów uzyskanych w oparciu o kwerendę pierwotną, czy też do wyszukiwania terminów złożonych o podwyższonej częstości występowania w dokumencie. Na wyjściu uzyskuje się referencje do dokumentów uporządkowane pod względem szacowanego stopnia adekwatności w stosunku do kwerendy. Badania z zakresie IR stymulowane są od roku 1992 przez serię dorocznych konferencji roboczych znanych pod nazwą TREC Workshops10. Ich celem jest sukcesywna ocena postępu prac i ich stymulacja przez wyznaczanie nowych zadań.

NOWE WYZWANIE: Q&A

Połączenie technik IE oraz IR
z metodami sztucznej inteligencji
(w zakresie rozumienia tekstu)
pozwala na zrobienie kolejnego
kroku w kierunku systemów, gdzie
w odpowiedzi na pytanie postawione
w języku naturalnym użytkownik
otrzymuje udokumentowaną
(wzgl. możliwą do
udokumentowania) odpowiedź,
też w języku naturalnym.

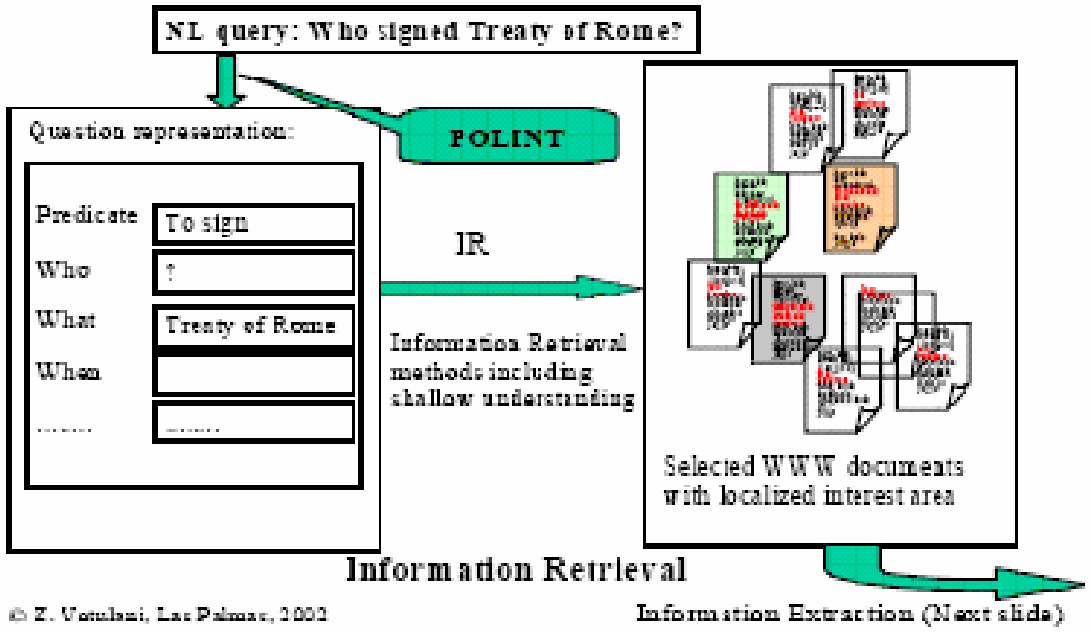
NOWE WYZWANIE: Q&A

Schemat przewiduje następujące operacje:

1. **Zrozumienie pytania (Rys. 1).** Zrozumienie pytania prowadzi do uzyskania szablonu sytuacji będącej przedmiotem pytania, co wymaga pełnej analizy obejmującej wszystkie poziomy opisu (morfologię, składnię, semantykę i pragmatykę). Można to uzyskać wykorzystując systemy MU typu POLINT.

2. **Wyszukanie dokumentów (Rys. 1).** Szablon służy do wygenerowania kwerendy w celu uzyskania dokumentów zawierających odpowiedź oraz, o ile to możliwe, do zlokalizowania w tych dokumentach obszarów, w których spodziewamy się znaleźć elementy odpowiedzi (stosując techniki IR).

NOWE WYZWANIE: Q&A

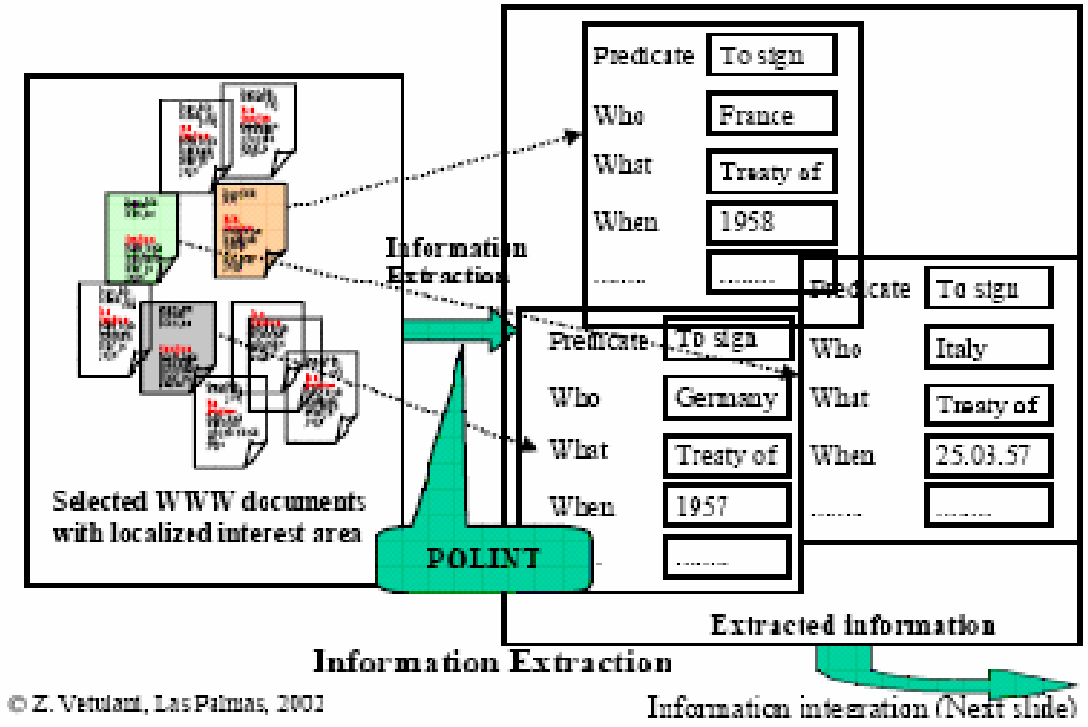


Rys. 1. Zrozumienie pytania i wyszukanie dokumentów

NOWE WYZWANIE: Q&A

3. Ekstrakcja informacji (Rys. 2). Kolejną fazą jest faza ekstrakcji informacji (IE) z dokumentów lub ich fragmentów w sposób sterowany szablonem pytajnym uzyskanym w pierwszej fazie przetwarzania pytania. Na wyjściu należy spodziewać się zbioru szablonów częściowo wypełnionych informacją, z których każdy może być uważany za częściową odpowiedź na pytanie. Podobnie jak w fazie pierwszej wykorzystuje się tu technologię MU. Wypełnione (częściowo) szablony będą wymagały dalszego przetwarzania, które określamy mianem integracji informacji.

NOWE WYZWANIE: Q&A

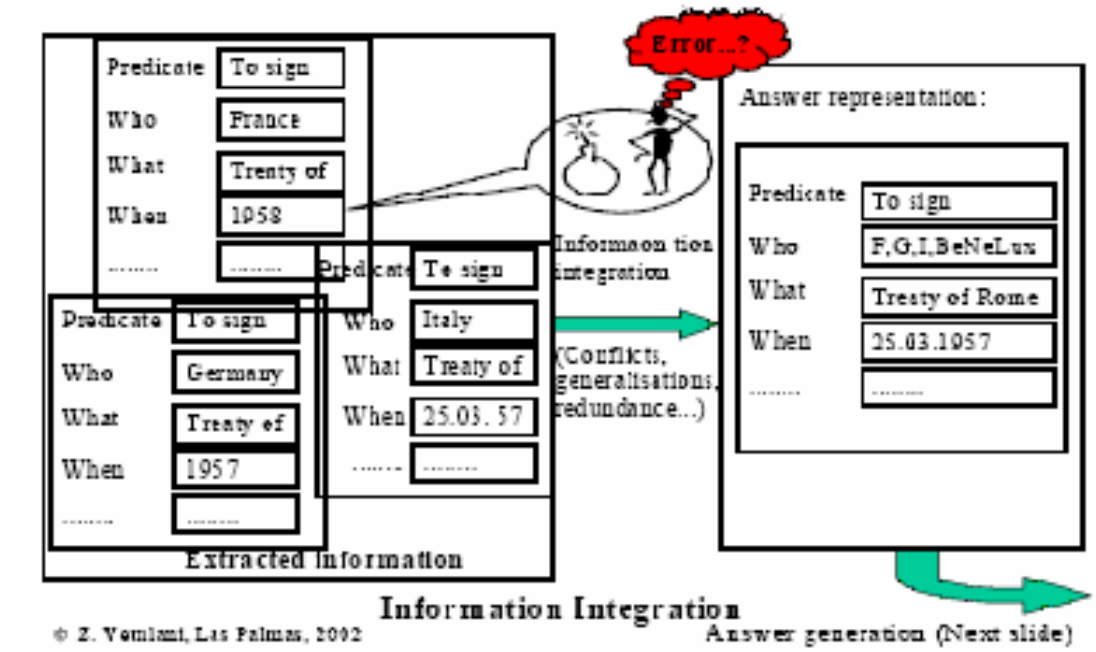


Rys. 2. Ekstrakcja informacji

NOWE WYZWANIE: Q&A

4. Integracja informacji (Rys. 3). Integracja informacji jest niezbędna, gdyż szablony mogą zawierać informację niezupełną (w stosunku do oczekiwań) lub sprzeczną. Jednym z głównych zadań tej fazy jest właściwa identyfikacja bytów co jest niezbędne wobec częstego oznaczania tego samego obiektu różnymi nazwami. Celem integracji dokumentów jest stworzenie jednego wypełnionego szablonu wraz ze wskazaniem na źródła i protokółem rozstrzygnięcia sprzeczności. Faza ta może wymagać dodatkowych kwerend generowanych automatycznie przez system.

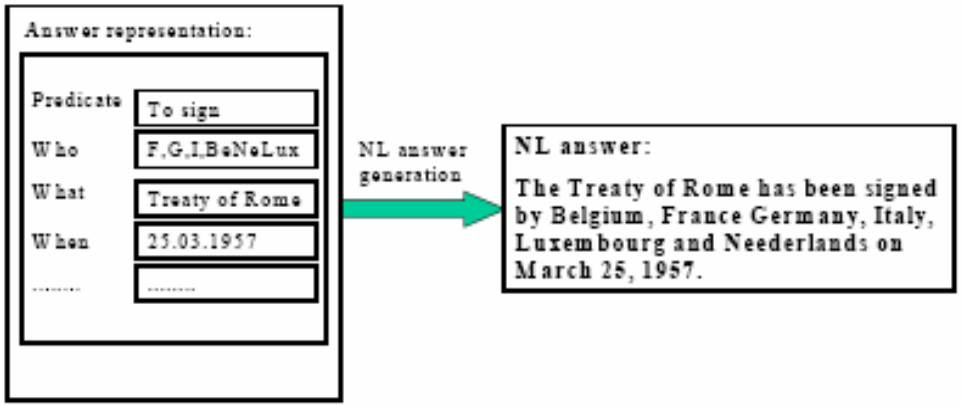
NOWE WYZWANIE: Q&A



Rys. 3. Integracja informacji

NOWE WYZWANIE: Q&A

5. Generowanie odpowiedzi (Rys. 4). Dopiero na podstawie zintegrowanego szablonu nastąpi generowanie szablonu w języku naturalnym.



Rys. 4. Generowanie odpowiedzi

NOWE WYZWANIE: Q&A

W środowisku związanym z konferencjami TREC powstał program stymulowania prac zmierzających do opracowania technologii Q&A jako naturalnego rozszerzenia technologii MU, IE oraz IR. Przy tworzeniu programu jego autorzy (tzw. Roadmap Committee) określili szereg parametrów wyznaczających jakość systemu Q&A. Są to:

• czas rzeczywisty (timeliness): oznacza to imperatyw czasu rzeczywistego (kilka sekund przetwarzania) bez względu na ilość danych, ilość użytkowników czy stopień aktualności wiedzy (system ma brać pod uwagę także najnowsze zdarzenia i stany rzeczy),

NOWE WYZWANIE: Q&A

• dokładność (accuracy): "zła odpowiedź jest gorsza niż żadna", co implikuje konieczność rozwiązywania konfliktów i sprzeczności w danych źródłowych, brania pod uwagę wiedzy ogólnej o świecie i modelu wiedzy użytkownika,

• użytkowość (usability): systemy Q&A mają respektować specyficzne żądania użytkownika, łatwo poddawać się kustomizacji do różnych dziedzin, dopuszczać heterogeniczne źródła informacji (jako, że mimo przewagi dokumentów tekstowych, wiele z nich ma charakter multimedialny). Oznacza to konieczność "wydobywania danych" niezależnie od ich natury i formułowania odpowiedzi w formacie (języku) użytkownika,

NOWE WYZWANIE: Q&A

• zupełność (completeness): pożądana jest kompletność odpowiedzi z punktu widzenia potrzeb użytkownika. Oznacza to konieczność poszukiwania odpowiedzi w wielu źródłach, umiejętność oceny stopnia kompletności przez system Q&A i uwzględnienia wiedzy o użytkowniku.

• istotność (relevance): odpowiedzi systemu Q&A powinny zawierać wszystkie elementy istotne dla użytkownika w określonym kontekście. Oznacza to konieczność rozpoznania tego kontekstu, a przede wszystkim potrzeb i intencji użytkownika.

• Interaktywność systemu może być niezbędna.

Wyszukiwarki – Q&A

Prawdziwy przełom może więc nastąpić dopiero przy zmianie filozofii patrzenia na to, jak powinny działać wyszukiwarki – odejście od wyszukiwania dokumentów w sieci, a skupienie się na poszukiwaniu informacji.

Wynik wydania zapytania do typowej wyszukiwarki, to zazwyczaj lista posortowanych według „adekwatności” odnośników do zasobów w Internecie. Zazwyczaj jest to zalew informacji, który mimo obfitości, jest w praktyce zupełnie nieprzydatny.

Wyszukiwarki – Q&A

Natomiast o wiele ciekawszym wyjściem jest grupowanie tematyczne otrzymanych wyników. Najlepszym obecnie serwisem wykorzystującym tą technologie jest Vivisimo.

http://vivisimo.com

Podobny system, profilowany w szczególności dla języka polskiego, został stworzony na Politechnice Poznańskiej i nazywa się Marchewa(Carrot).

http://www.cs.put.poznan.pl/dweiss/site/research/carrot/index.html

Wyszukiwarki – Q&A

Prowadzone są intensywne badania mające na celu wizualizacje wyników zapytań do wyszukiwarek. Bardzo ładnym, wręcz komiksowym interfejsem odznacza się na tym polu serwis Kartoo. Prezentuje on wyniki w postaci grafu dokumentów, powiązanych pewnymi semantycznymi związkami.

http://www.kartoo.com/

Wyszukiwarki – Q&A

Jeszcze ciekawszym pomysłem jest utworzenie analogii pomiędzy Internetem a geografią i stworzenie „mapy obszaru” wyników.

http://maps.map.net

Wyszukiwarki – Q&A

http://www.brainboost.com/

- Bazuje się na wyszukiwarce google
- Przegląda setki stron i formułuje poprawną odpowiedź na pytanie użytkownika
- Algorytm:

  Użytkownik wpisuje pytanie: „Why is mars red?”

  Pytanie jest tłumaczone na kilka pytań, co zwiększa prawdopodobieństwo znalezienia prawidłowej odpowiedzi na postawione pytanie

  Przy pomocy google znajduje strony odpowiadające słowom kluczowym: red i mars.

  Przegląda pierwszą setkę stron na liście

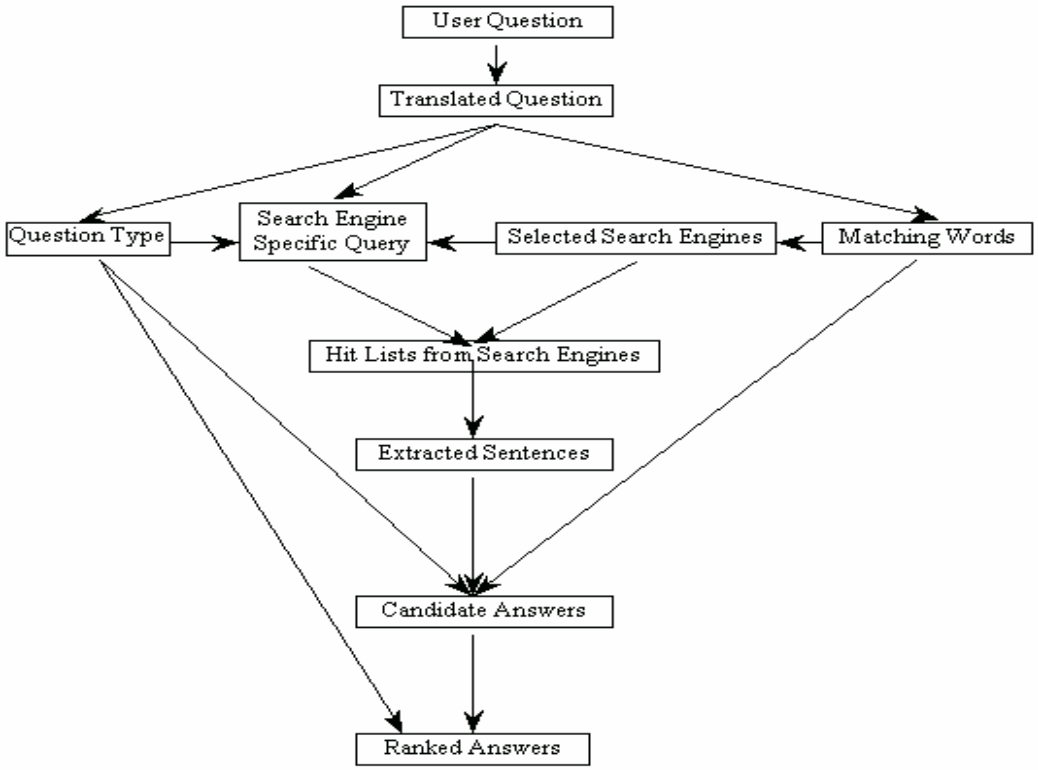
  Znajduje odpowiedzi na pytanie i ustawia ich w rankingu na podstawie technologii AnswerRank. AnswerRank oszacowuje jaka odpowiedź jest najbardziej prawdopodobna. W naszym przypadku to: „Mars is red because of the iron in the soil"

  Odpowiedź jest wyświetlona użytkownikowi.

Wyszukiwarki – Q&A

http://www.answerbus.com/about/

Bazuje się na wyszykiwarkach Google, Yahoo, WiseNut, AltaVista, and Yahoo News)



Wyszukiwarki – Q&A

http://start.csail.mit.edu/

START (SynTactic Analysis using Reversible Transformations)

Start parsuje wchodzące pytania, dopasowuje pytania utworzone z drzewa parsingu do swojej wiedzy w bazie i przezentuje odpowiadające informacje użytkownikowi.

Wyszukiwarki – Q&A

http://www.ask.com/ - Ask Jeeves

http://asked.jp/

http://qa.wpcarey.asu.edu/

http://asked.jp/edw/pc/index.html

Wyszukiwarki – Q&A

Dla Polaków wyszukiwarki anglosaskie mają średnie zastosowanie. Większość algorytmów analizy języka naturalnego, a także słowniki do nich niezbędne, są po prostu nieprzydatne do zapytań wydanych w języku polskim. Zresztą, nie jest to jedyne zmartwienie. Język polski posiada cechę, której większości wyszukiwarek w ogóle nie bierze pod uwagę: odmianę. Wyszukiwarka działająca w oparciu o słowa kluczowe, szukając słowa „burak”, zwróci tylko te strony, na których występuje ono dokładnie w tym brzmieniu. Słowa „buraki” albo „buraczany”, już nie będą uwzględnione. Chyba jedyne w Sieci serwisy wyszukując uwzględniające polską flekcję, jak również fakt, że coraz częściej pojawiają się w Internecie dokumenty pisane bez polskich znaków diakrytycznych („ogonków”), to ONET i NetSprint. Ta ostatnia, mimo użycia Googlo-podobnych czasami rozczarowuje, jeśli chodzi o jakość zwracanych wyników.

Linki:

- *http://www.zsi.pwr.wroc.pl/zsi/missi2002/pdf/p01.pdf* -Zygmunt Vetulani „Automatyczna interpretacja pytań i udzielanie odpowiedzi jako technologia multimedialna
- *http://www.cs.put.poznan.pl/dweiss/site/publications/download/zdjacgoogle.pdf* - Dawid Weiss „Zdjąć na chwile Gogle”
- *http://en.wikipedia.org/wiki/Question_answering* - wikipedia
- *http://trec.nist.gov/* - text retrieval conference