DEZVOLTAREA SISTEMULUI DE TIP „ÎNTREBARE-RĂSPUNS"

Victoria Bobicev, Liviu Carcea
Universitatea Tehnică a Moldovei
victoria_bobicev@rol.md, carcea@mail.utm.md

Abstract. *The paper presents a question-answering system created in the project „Research in the field of Information Retrieval for question answering system creation". The system consist of question analysis module and answer retrieval and extraction module. At the first stage the system is working on the base of the documents provided by IDSI (Institutul de Dezvoltare a Societății Informaționale - Information Society Development Institute). At prezent the system is installed on-line with the aim to collect questions from different users and its improvement using collected information.*

Cuvinte-cheie: *procesarea automată a textului, sistem întrebare-răspuns, analiza automată a întrebărilor, căutarea informației, extragerea informației.*

I. Introducere

Sistemele de extragere a răspunsurilor la întrebări în limbajul natural uman se înscriu în categoria sistemelor de achiziție a informației.

Procesul de cunoaştere are loc în modul următor: se uneşte un sistem informațional cu un ambient. Drept sistem informațional poate servi un individ care interacționează în mod direct şi conştient cu ambientul şi care îşi construieşte prin observație universul de cunoaştere, dar şi un sistem tehnic care primeşte informația de la diverşi agenți umani şi o procesează sau o pune, în continuare, la dispoziția altor persoane. Internetul reprezintă un sistem informațional care, acționând reciproc cu mediul agenților umani, îşi construieşte o bază proprie de cunoaştere.

Deseori, cunostințele căpătate prin intermediul internetului sunt în forma în care agenții umani o prezintă sistemului, deci în limbaj natural uman.

În prezent, internetul este cea mai mare sursă de cunoştințe ce se extinde şi se actualizează incontinuu. Internetul mai este şi una din cele mai accesibile locații în care aceste cunoştințe pot fi consultate. Însă dezvoltarea rapidă a internetului nu are loc fără aspecte negative, şi anume: din cauza volumului mare de informații disponibile, găsirea informației necesare poate fi, uneori, dificilă sau nesigură.

Cele mai eficiente metode de găsire şi de acumulare a informației o reprezintă, la ora actuală, motoarele de căutarea, scopul cărora este de a oferi utilizatorului un set de articole sau pagini web în care acesta să poată găsi informația necesară. Deseori articolele propuse de motoarele de căutare nu îndeplinesc dezideratul utilizatorului de a căpăta un răspuns satisfăcător. În plus, ele nu oferă răspunsul concret la întrebarea utilizatorului, dar numai un set de pagini web, şi utilizatorul e nevoit să-şi extragă informația necesară.

Etapa următoare în domeniul achiziției informației constă în dezvoltarea sistemelor capabile să răspundă la întrebările formulate de utilizator în limbajul natural. Scopul de bază al unui astfel de sistem constă în a asigura un răspuns la întrebarea utilizatorului care ar îndeplini următoarele trei condiții:

- să fie corect
- să fie formulat tot în limbaj natural
- să fi suficient de succint

Un sistem de răspunsuri la întrebări necesită o procesare a limbajului natural mult mai

complexă decât sistemele de achiziție de documente [3][5].

II. Sistemul de întrebare-răspuns

Sistemele de întrebare - răspuns (în engleză "question answering systems", sau sisteme QA) sunt caracterizate prin faptul că primesc întrebări formulate în limbaj natural şi, în baza unei colecții de documente, extrag răspunsul sau un set de răspunsuri găsite în documentele date. Astfel de sisteme sunt considerate ca fiind următorul pas în evoluția motoarelor de căutare a informației în surse textuale [6].

Ca regulă un astfel de sistem constă din următoarele module:

1. Modulul de analiză a întrebării – transformă întrebările formulate în limbaj natural uman în interogări pentru motorul de achiziție de documente;
2. Modulul de achiziție de articole – caută în colecția de articole articolele relevante pentru întrebarea formulata de utilizator, în baza datelor primite de la modulul de analiză a întrebării;
3. Modulul de extragere a răspunsului – din colecția de articole returnate de modulul de achiziție de articole, extrage un răspuns succint care constituie răspunsul în limbaj natural uman la întrebarea utilizatorului.

Sistemul ce se crează în cadrul proiectului „Cercetarea în domeniul de Regăsire a Informației în scopul creării sistemului electronic de informare publică" la prima etapă va funcționa în baza documentelor prestate de IDSI (Institutul de Dezvoltare a Societății Informaționale) ca, de exemplu, „Codul cu privire la știință şi inovare al Republicii Moldova", „Acordul de parteneriat între Guvern şi Academia de Ştiințe a Moldovei pentru anii 2009-2012", „Lege cu privire la parcurile ştiințifico-tehnologice şi incubatoarele de inovare", şi altele de acelaşi tip.

La momentul dat răspunsul la întrebarea pusă de utilizator se caută doar într-un document, deci, utilizatorul trebuie să aleagă documentul în care se va căuta răspunsul la întrebarea pusă din lista documentelor prestate de IDSI.

Tabelul 1. Exemple de propoziții în limba engleza şi română.

pronume/grup prepozițional interogativ	**Propoziția inversată**		**loc liber pentru răspuns**
	grup verbal	**grup nominal**	
What	is	inquiry ?	
In which documents	are	the explanations ?	
Ce	este	cercetarea aplicată ?	
De care acte	este reglamentată	activitatea în sfera ştiinței şi inovării ?	

III. Analiza întrebărilor

Metodologia clasică de analiza semantică a întrebărilor propusă în [1] presupune împărțirea propoziției interogative în următoarele elemente: pronume interogativ (de exemplu, *cine, care, unde, cînd*) sau grup prepozițional interogativ (de exemplu, *în ce, pe cine, de unde, la care*) şi propoziția inversată. Propoziția inversată începe cu grupul verbal după care urmează grupul nominal. La sfîrşitul propoziției se presupune prezența aşa numitului „gap" – loc liber prevăzut pentru răspunsul care va fi găsit. Astfel de metodologie este dezvoltată pentru întrebările formulate în limba engleză însă ea poate fi adaptată la analiza propozițiilor în limba română. Exemple de împărțire a propozițiilor engleze sunt prezentate în figura 1. Tabelul 1 conține exemplele propozițiilor în limba engleza şi română.

Grupul nominal în cazul dat este considerat fraza-cheie pentru căutarea răspunsului. Respectiv, întrebarea se reformulează pentru obținerea răspunsului. Grupul verbal şi cel nominal în propoziția inversată se schimbă cu locul şi dacă întrebarea începe cu o prepoziție aceasta se ataşează la sfîrşit. Locul liber pentru răspuns urmează dupa elementele acestea. Exemple de inversare a propozițiilor în limba engleza şi română sunt prezentate în Tabelul 2.

Tabelul 2. Exemple de inversare a propozițiilor în limba engleza şi română.

pronume/grup prepoziţional interogativ	**Propoziția inversată**		**loc liber pentru răspuns**
	grup verbal	**grup nominal**	
What	is	inquiry ?	. . .
întrebarea reformulată în răspuns			
	inquiry	is	. . .
In which documents	are	the explanations ?	
întrebarea reformulată în răspuns			
	the explanations	are	in . . .
Ce	este	cercetarea alicată ?	
întrebarea reformulată în răspuns			
	cercetarea alicată	este	. . .
De care acte	este reglamentată	activitatea în sfera ştiinței şi inovării ?	
întrebarea reformulată în răspuns			
	activitatea în sfera ştiinței şi inovării	este reglamentată	de . . .

Pronumele interogativ are şi el rolul său specific, el defineşte ***tipul*** răspunsului: se întreabă despre o persoană (*cine*), despre un obiect (*ce*), despre localitate (*unde*), despre timp (cînd). Sunt posibile şi întrebări mai complicate ce necesită răspunsuri desfăşurate, de exemplu: *Care sunt argumentele pro şi contra ...* [2]. Răspunsuri la astfel de întrebări necesită crearea rezumatului din documente multiple [4].

IV. Căutarea şi extragerea răspunsului

Rezultatul etapei precedente de analiză a intrebării este o listă de cuvinte şi fraze cheie ce sunt utilizate în procesul de căutare a răspunsului. Lista aceasta este ordonată în sens că frazele mai lungi sunt plasate la începutul listei şi cele mai scurte la urmă. De exemplu, pentru întrebarea „De care acte este reglamentată activitatea în sfera ştiinței şi inovării ?" lista frazelor-cheie care vor fi căutate este *(„activitatea în sfera ştiinței şi inovării este reglamentată de" „activitatea în sfera ştiinței şi inovării este reglamentată" „activitatea în sfera ştiinței şi inovării")*. Respectiv, toate elementele listei date sunt căutate în documentul ales.

Cum deja a fost menționat, răspunsul se caută doar într-un document selectat de utilizator. Documentele acestea sunt hotărâri sau legi şi au structura corespunzătoare: sunt împărțite în capitole, articole şi aliniate. Căutarea frazelor-cheie se efectuează pe aliniate. În fiecare aliniat din documentul dat se caută frazele-cheie obținute din întrebarea pusă.

Dacă într-un aliniat se găseşte o frază-cheie din lista de căutare acest aliniat se memorizează în lista răspunsurilor posibile. Pentru fiecare aliniat din lista dată se calculează scorul (ponderea).

$$\mathbf{P(a_i) = P(a_i) + 1/j - length(f_1)/(length(f_1)+length(fr_j))} \quad (1)$$

unde $\mathbf{a_i}$ – aliniatul numărul i;
$\mathbf{P(a_i)}$ – ponderea aliniatului cu numărul i;
j – numărul de ordine a frazei-cheie în lista frazelor-cheie;
$\mathbf{length(f_1)}$ – lungimea fragmentului aliniatului (numărul de caractere) înainte de fraza-cheie găsită de la primul caracter pînă la fragmentul găsit;
length($cuv) – lungimea cuvîntului-cheie găsit (numărul de caractere);

În formula dată primul element indică că ponderea se măreşte de fiecare dată cînd în aliniatul dat se găseşte oricare frază-cheie; al doilea element indică importanța frazei-cheie găsite: cu cît numărul ei de ordine în lista frazelor-cheie este mai mare cu atît ponderea este mai mică. Al treilea element ia în considerație cît de aproape de începutul aliniatului se află fraza-cheie găsită. Se consideră că în fiecare aliniat informația importantă se află la începutului textului iar la urmă sunt menționate lucruri secundare. Astfel, dacă fraza-cheie este menționată la începutul paragrafului ea este importantă în textul dat, iar dacă apare undeva mai departe de început, fraza-cheie în paragraful dat nu are o importanță mare.

La sfîrşitul etapei de căutare obținem lista aliniatelor care probabil conțin răspunsul corect la întrebarea pusă de utilizator. Toate aliniatele acestea sunt memorizate într-un fişier html. În pagina de răspuns care este vizualizată pentru utilizator este afişat doar răspunsul cu ponderea maximă.

Analiza întrebarii si vizualizarea raspunsului

Intrebarea pusă: Ce este cercetare

Răspusul la întrebarea pusă s-a căutat în Codul cu privire la ştiință şi inovare al Republicii Moldova

Răspunsul cu scorul maxim este #7 din 22 alese:

Titlu: I Capitol: II Articol: 6

Cercetare fundamentala - activitate orientata spre dobindirea de noi cunostinte stiintifice, spre formularea si verificarea de noi ipoteze si teorii.

alte răspunsuri selectate
la documentul în care s-a căutat răspunsul
la pagina precedenta

***

Fig. 1. Pagina web cu răspunsul şi opțiunile oferite utilizatorului

Astfel, utilizatorului îi este oferit doar un aliniat de text în calitate de răspuns la întrebarea pusă. În caz dacă răspunsul acesta este satisfăcător utilizatorul are cîteva opțiuni propuse de sistem:

- de a vizualiza documentul în locul unde se află aliniatul afişat şi să verifice dacă răspunsul relevant se află în aliniatele apropiate de aliniatul oferit; în unele cazuri fraza-cheie se găseşte intr-un aliniat iar informația importantă despre conceptele date este repartizată prin aliniatele apropiate;
- de a vizualiza toate răspunsurile găsite de sistem şi memorizate într-un fişier html; este posibil că aliniatul care conține răspunsul corect a fost găsit de sistem însă nu a obținut ponderea maximă.

Pagina cu răspunsul şi opțiunile oferite utilizatorului este prezentată în figura 1.

V. Problemele nerezolvate

- Întrebările şi răspunsurile se scriu fără semne diacritice. De fapt, problema semnelor diacritice este problema comună pentru site-uri scrise în limbi diferite de limba engleză. Introducerea codificării Unicode a rezolvat problema dată pentru paginile web scrise în html. Însă datele introduse în formular şi transferate pe server sunt recodificate în procesul transmiterii. Rezultatul transmis pe server depinde de factori multipli, cum ar fi: sistemul utilizatorului, proprietăţile nodurilor prin care se transferă datele, şi altele. Problema aceasta este pur tehnică, însă la momentul dat nu este rezolvată.
- Nu se iau în consideraţie formele morfologice ale cuvintelor. De exemplu, dacă este întrodusă întrebarea „Ce este cercetarea?", sistemul nu va găsi răspunsul din cauza că în document este scris „cercetare". Motoarele de căutare moderne deja lucrează asupra problemei formelor morfologice a cuvintelor. Noi dispunem de un dicţionar morfologic a limbii române cu aproximativ 90000 de forme morfologice a cuvintelor care poate fi utilizat pentru obţinerea tuturor formelor a cuvintelor-cheie. Problema constă în faptul că în majoritatea cazurilor sunt căutate nu dor cuvinte dar fraze-cheie. De exemplu, dacă avem fraza „*Cercetare fundamentală*", formele ei sunt: „*Cercetarării fundamentale*", „*Cercetarea fundamentală*", „*Cercetărilor fundamentale*", şi altele. Aceasta înseamnă că trebuie de schimbat forma tuturor cuvintelor în frază respectînd acordul morfologic între ele. Problema aceasta poate fi rezolvată însă necesită intervenţia lingviştilor.
- Se procesează un număr limitat de tipuri de întrebări. La momentul de faţă sistemul procesează întrebările de tip definiţie. Sistemul este instalat online cu scopul completării listei de întrebări de la utilizatori şi memorizează toate adresările utilizatorilor, întrebările lor şi răspunsurile găsite. Pe parcursul completării listei date modulul de analiză a întrebărilor va fi perfectat ca să fie capabil să proceseze toate tipurile întrebărilor de la utilizatori.
- Răspunsul se caută numai într-un document. Versiunea curentă a sistemului propune utilizatorului să aleagă documentul în care se va căuta răspunsul. IDSI a prestat un număr redus de documente şi nu este problematic de încercat toate documentele pe rînd. Pe parcursul utilizării sistemului numărul de documente va fi mărit şi sistemul va fi modificat în aşa mod că utilizatorul să poată căuta în cîteva documente simultan sau în toate documente.

VI. Concluzii

În lucrarea dată este prezentat sistemul de tip întrebare-răspuns creat în cadrul proiectului „Cercetarea în domeniul de Regăsire a Informaţiei în scopul creării sistemului electronic de informare publică". Sistemul constă din modulul de analiză a întrebărilor şi modulul de căutare şi selectare a răspunsului. La prima etapă sistemul va funcţiona în baza documentelor prestate de IDSI (Institutul de Dezvoltare a Societăţii Informaţionale). La momentul dat sistemul este instalat on-line cu scopul culegerii întrebărilor de la utilizatori şi perfectării lui în baza informaţiei colectate.

VII. Referinţe

1. R. Baeza-Yates, B. Ribeiro-Nieto, Modern Information Retrieval, Addison Wesley, 2000.

2. R. Botnaru, V. Bobicev Studiul tipurilor de întrebări din sistemul de întrebare-răspuns. ICMCS, Chişinău 2011.

3. L. Carcea. Abordări în dezvoltarea sistemelor „întrebare-răspuns". ICMCS, Chişinău 2011.

4. V. Lazu., T. Prodan. Metodologia reprezentării sensului întrebării în forma logică. Conferinţa ştiinţifică a colaboratorilor doctoranzilor şi masteranzilor UTM.

5. V. Maxim. Metode utilizate pentru elaborarea unui sistem „întrebare-răspuns". ICMCS, Chişinău 2011.

6. Dan Tufiş, Dan Ştefănescu, Radu Ion, and Alexandru Ceauşu. RACAI's Question Answering System at QA@CLEF 2007. In Alessandro Nardi and Carol Peters, editors, Working Notes for the CLEF 2007 Workshop, pages 15–21, 2007.