PROCESAREA ŞI ANALIZA ÎNTREBĂRILOR ÎNTR-UN SISTEM DE „ÎNTREBARE-RĂSPUNS"

Victoria Lazu, Tatiana Prodan
Universitatea Tehnică a Moldovei
lazu_vica@mail.utm.md, tatiana.prodann@gmail.com

Abstract. *This paper tackles the issue of the question-answering system for Romanian speakers that is being developed within the project "Research in the field of Information Retrieval for question-answering system creation". The document collection provided by the Information Society Development Institute serves as basis for the first stage of system creation and implementation. The questions extracted from these documents have been classified according to the question types defined by the linguists. The two modules of the system enable it to process the questions and retrieve the answers. The created system was subject to primary testing and the results are presented.*

Cuvinte-cheie: *sistem de tip „întrebare-răspuns", clasificarea întrebărilor, procesarea automată a întrebărilor, căutarea şi extragerea informației*

I. Introducere

Societatea modernă acordă o importanță majoră informației, iar obținerea operativă a acesteia a devenit o necesitate zilnică pentru mulți cetățeni. De fapt, documentele expuse pe paginile web nu garantează accesul rapid şi extragerea informației necesare unui utilizator. Regăsirea informației în volume mari de texte reprezintă una din cele mai acute probleme. Prin urmare, o atenție sporită este acordată metodelor de analiză a interogărilor în formă de întrebări şi nu în formă de cuvinte cheie. Astfel, pentru a soluționa această problemă şi a-i oferi utilizatorului posibilitatea de a găsi în timp util informația căutată, sistemele de tip întrebare-răspuns urmăresc scopul de a furniza răspunsuri la întrebările adresate în limbaj natural.

Crearea unui sistem de întrebare-răspuns reprezintă un proces complex care implică multe componente şi resurse importante. Un sistem de întrebare-răspuns bazat pe o colecție de documente, de obicei, este format din trei componente principale [5]:

1) Modul de analiză a întrebării - convertește întrebările din limbaj natural în interogări pentru motorul de achiziție de documente.

2) Modul de achiziție de documente - caută în totalitatea de documente articolele relevante pentru interogarea făcută de utilizator în baza datelor returnate de modulul de analiză a întrebării.

3) Modul de extragere a răspunsului - din colecția de articole returnate de modulul de achiziție de documente, extrage un răspuns succint şi care constituie răspunsul în limbaj natural la întrebarea utilizatorului.

Sistemul care este dezvoltat în cadrul proiectului „Cercetarea în domeniul de Regăsire a Informației în scopul creării sistemului electronic de informare publică" este lansat în cooperare cu Institutul de Dezvoltare a Societății Informaționale (IDSI) şi urmărește scopul de a cerceta problemele de regăsire a informației şi de a crea un sistem automat de Întrebare-Răspuns pentru vorbitorii de limbă română. Sistemul va procesa interogările şi va extrage răspunsurile la acestea din documentele relevante. Prima etapă de implementare a proiectului prevede elaborarea unui sistem de întrebare-răspuns de tip închis din punct de vedere al informației vehiculate în baza unei colecții de documente legislative. Astfel, documentele utilizate la această etapă în procesul de elaborare al

sistemului includ:

1) Codul cu privire la ştiinţă şi inovare al Republicii Moldova nr. 259-XV din 15.07.2004;
2) Legea cu privire la parcurile ştiinţifico-tehnologice şi incubatoarele de inovare nr. 138-XVI din 21.06.2007;
3) Hotărâre cu privire la aprobarea Acordului de parteneriat între Guvern şi Academia de Ştiinţe a Moldovei pentru anii 2009-2012 nr. 27 din 22.01.2009.

Documentele menţionate au fost analizate detaliat şi ca urmare a fost creat manual setul de întrebări corespunzătoare pentru a fi utilizate la elaborarea şi testarea primară a sistemului.

II. Procesarea şi testarea întrebărilor

Sistemul format cuprinde un modul de procesare a întrebării şi un modul de căutare şi extragere a fragmentelor de text care conţin un potenţial răspuns.

Pentru a procesa întrebările adresate de către un utilizator prin intermediul sistemului descris, iniţial este necesar să se determine tipul interogării respective. Definirea tipului întrebării în raport cu o clasificare existentă va permite stabilirea tipului de răspuns ce urmează a fi generat de sistem.

Din punct de vedere semantic, întrebările se clasifică în mai multe tipuri. În procesul de elaborare a sistemului menţionat clasificarea întrebărilor a fost efectuată în baza categoriilor QA@CLEF [7]. În tabelul 1 sunt prezentate categoriile majore care au fost definite pentru întrebările extrase şi exemple de întrebări.

Tabelul 1. Clasificarea întrebărilor după categorii şi exemple de întrebări

Nr.	Categorie	Exemplu
1.	Definiţie (DEF)	Ce este Consiliul Naţional pentru Acreditare şi Atestare?
2.	Nume/persoană (N)	Cine gestionează Complexul patrimonial al Academiei de Ştiinţe?
3.	Timp (TMP)	În ce an a fost adoptat Codul cu privire la ştiinţă şi inovare al Republicii Moldova?
4.	Interval de timp (ITMP)	De câte ori pe an se convoacă Comisia de acreditare a organizaţiilor din sfera ştiinţei şi inovării?
5.	Listă (LST)	Ce funcţii ştiinţifice există în sfera ştiinţei?
6.	Loc (LOC)	În ce articol sunt descrise consecinţele neacreditării?
7.	Măsură (MES)	Câte etape are procesul de acreditare?
8.	Explicaţie/descriere (EXP)	Care este Structura Consiliului Naţional pentru Acreditare şi Atestare?

Modulul de procesare a întrebării în cadrul sistemului creat se efectuează în trei etape consecutive în rezultatul cărora are loc formalizarea interogărilor limbajului natural. Astfel, etapele includ analiza părţilor întrebării, formarea listei de cuvinte cheie şi reformularea interogării în răspuns.

Analiza lingvistică a părţilor întrebării este efectuată prin definirea grupului prepoziţional, grupului verbal şi grupului nominal al acesteia [2]. Pornind de la faptul că orice întrebare adresată de către un utilizator prin intermediul sistemului dat începe cu partea interogativă, elementele prevăzute în acest fragment al întrebării, sau mai bine zis în grupul prepoziţional, includ următoarele părţi de vorbire - prepoziţia, pronumele şi particula. Spre exemplu, în întrebarea „*Cine este titular de drept*?" partea interogativă este formată din pronumele „*Cine*". De asemenea, determinarea tipului interogării în baza categoriilor de întrebări menţionate mai sus are loc prin stabilirea indicatorului corespunzător. În exemplul dat, partea interogativă reprezintă în acelaşi timp şi indicatorul tipului întrebării. Astfel, pronumele „*Cine*" este indicatorul pentru tipul de întrebare

„*Nume/persoană (N)*", sistemul urmând să extragă ca răspuns numele sau descrierea unui agent, persoane, societăți etc.

Grupul nominal al întrebării exemplificate constituie lista cuvintelor cheie - „*titular de drept*". Reformularea întrebării în prima parte a răspunsului posibil are loc, în cazul dat, prin eliminarea părții interogative şi plasarea grupului verbal „*este*" după grupul nominal - „*titular de drept este ...*". Rezultatul căutării formează răspunsul integral la interogarea utilizatorului.

Tabelul 2. Exemplu de întrebare şi reformulare a întrebării de către sistem

Întrebarea în limbaj natural			
Grup prepoziţional interogativ	***Grup verbal***	***Grup nominal***	
Cine	*este*	*titular de drept?*	
Pentru ce	*sunt destinate*	*proiectele ştiinţifice?*	
Întrebarea reformulată			
	Grup nominal	***Grup verbal***	***Presupus răspuns***
	Titular de drept	*este*	*...*
	Proiectele ştiinţifice	*sunt destinate*	*...*

După finalizarea operațiilor din cadrul primului modul are loc căutarea şi extragerea răspunsului propriu-zis din documentele disponibile. Căutarea este efectuată pe fragmente de texte (aliniate) în baza cuvintelor cheie obținute după procesarea întrebării. Dacă sistemul identifică un cuvânt cheie, fragmentul de text în care se găseşte acesta este memorat în lista răspunsurilor posibile. Pentru a determina şi a prezenta utilizatorului cel mai corespunzător răspuns la întrebarea sa, sistemul calculează ponderea (scorul) pentru fiecare fragment memorat cu ajutorul formulei:

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

unde:

a_i - aliniatul numărul i;

$P(a_i)$ - ponderea aliniatului cu numărul i;

j - numărul de ordine a frazei-cheie în lista frazelor-cheie;

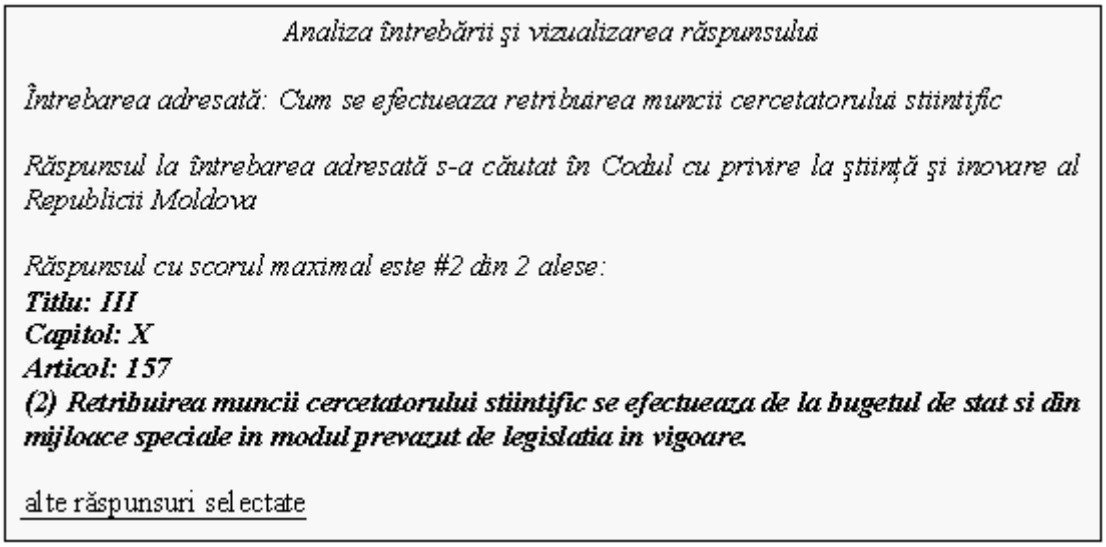
$length(f_1)$ - lungimea fragmentului aliniatului (numărul de caractere) înainte de fraza-cheie găsită de la primul caracter până la fragmentul găsit;

$length(fr_j)$ - lungimea cuvântului cheie găsit (numărul de caractere).

În această relație primul element arată creşterea ponderii la fiecare găsire de fraze-cheie în aliniatul procesat. Al doilea element indică importanța frazei cheie - cu cât numărul ei de ordine în lista frazelor cheie este mai mare cu atît ponderea este mai mică. Al treilea element ia în considerare distanța dintre începutul aliniatului şi fraza cheie. Se cunoaşte că în orice aliniat informația importantă se află la începutul textului, iar la urmă sunt menționate lucruri secundare. Astfel, dacă fraza cheie este enunțată 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.

Prin urmare, sistemul afişează răspunsul care are ponderea maximă, însă în cazul răspunsurilor multiple celelalte rezultate sunt înregistrare într-un fişier disponibil pentru utilizator prin accesarea link-ului apărut pe aceeaşi pagină.

Pentru a verifica corectitudinea funcționării sistemului creat, acesta a fost testat cu o serie de întrebări. Un exemplu de întrebare adresată şi răspuns găsit este prezentat în figura 1. Precum este în figură, sistemul afişează în acelaşi timp titlul, capitolul şi articolul din actul legislativ în care a fost găsit răspunsul. În cazul în care utilizatorul este satisfăcut de răspunsul oferit, sistemul îl memorează în lista întrebărilor şi răspunsurilor care au fost deja căutate de alți utilizatori.


Analiza întrebării şi vizualizarea răspunsului

Întrebarea adresată: Cum se efectueaza retribuirea muncii cercetatorului stiintific

Răspunsul la întrebarea adresată s-a căutat în Codul cu privire la ştiinţă şi inovare al Republicii Moldova

Răspunsul cu scorul maximal este #2 din 2 alese:
Titlu: III
Capitol: X
Articol: 157
(2) Retribuirea muncii cercetatorului stiintific se efectueaza de la bugetul de stat si din mijloace speciale in modul prevazut de legislatia in vigoare.

alte răspunsuri selectate

Fig. 1. Exemplu de întrebare şi răspuns

După efectuarea testării preliminare a sistemului de întrebare-răspuns elaborat în cadrul proiectului menționat, rezultatele obținute au reliefat punctele pozitive şi negative ale sistemului. În primul rând, o mare parte a răspunsurilor la întrebări au fost extrase greşit. De multe ori răspunsul este incomplet, sistemul generând doar numărul titlului, capitolului, precum şi denumirea articolului fără a afişa textul integral din articol ce corespunde răspunsului. Fenomenul dat are loc în special în cazul întrebărilor de tip „Listă". De asemenea, sistemul a identificat incorect cuvintele cheie la unele întrebări mai complicate. În acelaşi timp, sistemul a extras la etapa actuală un număr suficient de răspunsuri corecte. În tabelul 3 pot fi vizualizate datele privind răspunsurile obținute la întrebările testate.

Tabelul 3. Ponderea răspunsurilor obținute

Răspunsuri corecte	**Răspunsuri incorecte**	**Întrebări fără răspuns**	**Răspunsuri multiple**
35,82%	39,55%	24,63%	44,78%

Este necesar de menționat şi faptul că răspunsurile incorecte au inclus şi rezultatele parțial corecte acestea fiind înregistrate în fişierul răspunsurilor multiple. Acestea au constituit 64,15% din totalul întrebărilor incorecte.

O altă remarcă ține de faptul că majoritatea răspunsurilor corecte au fost extrase pentru întrebările de tip „Definiție" şi de tip „Nume/persoană". În tabelul 4 sunt prezentate datele privind răspunsurile corecte obținute în dependență de tipul întrebării.

Tabelul 4. Procentajul rezultatelor de răspunsuri corecte în raport cu tipul întrebării

Tip întrebare	Nume/persoană	Definiție	Explicație	Listă	Interval de timp
Acurate□e	50%	54,55%	24%	13,16%	0

Sistemul urmează a fi testat în continuare pentru toate tipurile de întrebări în scopul identificării soluțiilor la problemele existente şi perfecționării modulelor acestuia.

III. Probleme şi neajunsuri actuale ale procesării

La moment, sistemul creat se confruntă cu unele probleme care urmează a fi soluționate în

următoarele etape de testare şi implementare. Lista problemelor majore include:

1) Răspunsurile la întrebări se caută doar într-un singur document, utilizatorul fiind constrâns să selecteze documentul la care se referă întrebarea sa. În procesul de implementare ulterioară, colecţia de documente va spori considerabil. Prin urmare, sistemul va fi ajustat cerinţelor şi va fi modificat astfel încât utilizatorul să aibă posibilitate să caute informaţiile necesare în mai multe documente simultan.

2) Sistemul procesează corect un număr limitat de tipuri de întrebări. Precum a fost descris mai sus, sistemul extrage o rată mai mare de răspunsuri corecte la întrebările de tip „Definiţie" şi „Nume/persoană". În consecinţă, va fi propusă o metodă complexă de analiză a tuturor tipurilor de întrebări în baza căreia să fie căutate şi selectate răspunsurile corecte.

3) Formele morfologice ale cuvintelor nu sunt considerate. Spre exemplu răspunsul la întrebarea „*Ce reprezintă certificatul de acreditare?*" nu va fi găsit din motiv că în documentul legislativ această sintagmă apare ca „*certificat de acreditare*". În mod corespunzător, este necesar ca sistemul să analizeze toate formele morfologice ale unui cuvânt - aspect ce urmează a fi soluţionat.

4) Întrebările şi răspunsurile sunt scrise fără semne diacritice, însă această problemă ţine de aspectul tehnic al sistemului şi de asemenea, va fi soluţionată în cadrul următoarei etape.

IV. Concluzii

În articolul dat se descrie procesarea şi analiza întrebărilor pentru sistemul de tip întrebare-răspuns ce se crează în baza documentelor IDSI în cadrul proiectului „Cercetarea în domeniul de Regăsire a Informaţiei în scopul creării sistemului electronic de informare publică". Trebuie de menţionat că sistemul este disponibil on-line sub forma unei aplicaţii web. Sistemul a fost testat cu o serie de întrebări, răspunsurile fiind analizate şi structurate. În urma experimentului s-a depistat că sistemul răspunde corect la 35,8% de întrebări. S-a observat că sistemul returnează relatări pozitive la interogări de tip "*Definiţie*", acurateţea fiind de 0.54. Totodată, este necesară revizuirea modulului de analiză morfologică a întrebării pentru corelarea căutării răspunsului în sistem.

V. Referinţe

1. Bobicev V. Preprocesarea textelor în sistemele de tip „întrebare-răspuns". 7th International Conference on Microelectronics and Computer Science, Chişinău, 2011, p.278-282, ISBN 978-9975-45-174-1

2. Botnaru R., Bobicev V. Studiul tipurilor de întrebări din sistemul de întrebare-răspuns. 7th International Conference on Microelectronics and Computer Science, Chişinău, 2011, ISBN 978-9975-45-174-1

3. Carcea L. Abordări în dezvoltarea sistemelor „întrebare-răspuns". 7th International Conference on Microelectronics and Computer Science, Chişinău, 2011, p.207-210, ISBN 978-9975-45-174-1

4. Giampiccolo D., Forner P., Peñas A., Ayache C., Cristea D., Jijkoun V., Osenova P., Rocha P., Sacaleanu B. and Sutcliffe R. Overview of the CLEF 2007 Multilingual Question Answering Track. Online proceedings of CLEF 2007 Working Notes, Budapest, September, 2007, ISBN: 2-912335-31-0

5. Maxim V. Metode utilizate pentru elaborarea unui sistem „întrebare-răspuns". 7th International Conference on Microelectronics and Computer Science, Chişinău, 2011, p.293-298, ISBN 978-9975-45-174-1

6. Tufiş D., Ştefănescu D., Ion R. and Ceauşu A. 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.