

ANALIZA RĂSPUNSURILOR ÎN SISTEMUL DE „ÎNTREBARE-RĂSPUNS”

Victoria LAZU

Universitatea Tehnică a Moldovei
lazu_vica@mail.utm.md

Abstract. This paper deals the issue of the question-answering system for Romanian that is being developed within the project “Research in the field of Information Retrieval for question-answering system creation”. The questions extracted from the documents, provided by the Information Society Development Institute, 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 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 și extragerea informației

I. Introducere

Regăsirea informației în volume mari de texte ce ne asaltă zilnic reprezintă una din cele mai acute probleme a societății moderne. Prin urmare, o atenție sporită este acordată metodelor de analiză a interogărilor în formă de întrebări și nu în formă de cuvinte cheie. Astfel, pentru a soluționa această problemă și 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 și resurse importante. Un sistem de întrebare-răspuns bazat pe o colecție de documente, de obicei, este format din trei componente principale [3]:

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

5) 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.

6) Modul de extragere a răspunsului - din colecția de articole returnate de modulul de achiziție de documente, extrage un răspuns succint și 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) și urmărește scopul de a cerceta problemele de regăsire a informației și de a crea un sistem automat de Întrebare-Răspuns pentru vorbitorii de limbă română. Sistemul va procesa interogările și 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:

▲ Codul cu privire la știință și inovare al Republicii Moldova nr. 259-XV din 15.07.2004;

▲ Legea cu privire la parcurile științifico-tehnologice și incubatoarele de inovare nr. 138-XVI din 21.06.2007;

▲ Hotărâre cu privire la aprobarea Acordului de parteneriat între Guvern și Academia de Științe a Moldovei pentru anii 2009-2012 nr. 27 din 22.01.2009.

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

II. Clasificarea întrebărilor

Sistemul format cuprinde un modul de procesare a întrebării și un modul de căutare și extragere a fragmentelor de text care conțin un potențial răspuns.

Pentru a procesa întrebările adresate de către un utilizator prin intermediul sistemului descris, inițial este necesar să se determine tipul interogării respective. Definierea 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 menționat clasificarea întrebărilor a fost efectuată în baza categoriilor QA@CLEF [4]. În tabelul 1 sunt prezentate categoriile majore care au fost definite pentru întrebările extrase și exemple de întrebări.

Tabelul 1. Clasificarea întrebărilor după categorii și exemple de întrebări

Nr.	Categorie	Exemplu
1.	Definiție (DEF)	Ce este Consiliul Național pentru Acreditare și Atestare?
2.	Nume/persoană (N)	Cine gestionează Complexul patrimonial al Academiei de Științe?
3.	Timp (TMP)	În ce an a fost adoptat Codul cu privire la știință și inovare al Republicii Moldova?
4.	Interval de timp (ITMP)	De câte ori pe an se convoacă Comisia de acreditare a organizațiilor din sfera științei și inovării?
5.	Listă (LST)	Ce funcții științifice există în sfera științei?
6.	Loc (LOC)	În ce articol sunt descrise consecințele neacreditării?
7.	Măsură (MES)	Câte etape are procesul de acreditare?
8.	Explicație/descriere (EXP)	Care este Structura Consiliului Național pentru Acreditare și Atestare?

III. Procesarea întrebărilor

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 părților întrebării, formarea listei de cuvinte cheie și reformularea interogării în răspuns.

Analiza lingvistică a părților întrebării este efectuată prin definirea grupului prepozițional, grupului verbal și 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 prepozițional, includ următoarele părți de vorbire - prepoziția, pronumele și 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 menționate mai sus are loc prin stabilirea indicatorului corespunzător. În exemplul dat, partea interogativă reprezintă în același timp și 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 și plasarea grupului verbal „este” după grupul nominal - „titular de drept este ...”. Rezultatul căutării formează răspunsul integral la interogarea utilizatorului.

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

Întrebarea în limbaj natural			
Grup prepozițional interogativ	Grup verbal	Grup nominal	
<i>Cine</i>	<i>este</i>	<i>titular de drept?</i>	
<i>Pentru ce</i>	<i>sunt destinate</i>	<i>proiectele științifice?</i>	
Întrebarea reformulată			
	Grup nominal	Grup verbal	Presupus răspuns
	<i>Titular de drept</i>	<i>este</i>	...
	<i>Proiectele științifice</i>	<i>sunt destinate</i>	...

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

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

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;

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

$\text{length}(f_r_j)$ - lungimea cuvântului cheie găsit (numărul de caractere).

În această relație primul element arată creșterea ponderii la fiecare găsim 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 atât ponderea este mai mică. Al treilea element ia în considerare distanța dintre începutul aliniatului și fraza cheie. Se cunoaște 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 afișează răspunsul care are ponderea maximă, însă în cazul răspunsurilor multiple celelalte rezultate sunt înregistrate într-un fișier disponibil pentru utilizator prin accesarea link-ului apărut pe aceeași pagină.

IV. Testarea întrebărilor

Pentru a verifica corectitudinea funcționării sistemului creat, acesta a fost testat cu o serie de întrebări. Un exemplu de întrebare adresată și răspuns găsit este prezentat în figura 1. Precum este în figură, sistemul afișează în același timp titlul, capitolul și 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 și răspunsurilor care au fost deja căutate de alți utilizatori.

Analiza întrebării și vizualizarea răspunsului

Întrebarea adresată: Cum se efectueaza retribuirea muncii cercetatorului stiintific

Răspunsul la întrebarea adresată s-a căutat în Codul cu privire la știință și 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 și răspuns

După efectuarea testării preliminare a sistemului de întrebare-răspuns elaborat în cadrul proiectului menționat, rezultatele obținute au reliefat punctele pozitive și negative ale sistemului. În primul rând, o mare parte a răspunsurilor la întrebări au fost extrase greșit. De multe ori răspunsul este incomplet, sistemul generând doar numărul titlului, capitolului, precum și denumirea articolului fără a afișa 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 același 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
39,85%	36,84%	24,03%	46,6%

Este necesar de menționat și faptul că răspunsurile incorecte au inclus și rezultatele parțial corecte acestea fiind înregistrate în fișierul răspunsurilor multiple. Acestea au constituit 60,87% din totalul întrebărilor incorecte.

O altă remarcă ține de faptul că majoritatea răspunsurilor corecte au fost extrase pentru întrebările de tip „Definiție” și 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
Acuratețe	71,4%	57,4%	24%	10,16%	0

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

V. Concluzii

În articolul dat se descrie procesarea și analiza întrebărilor pentru sistemul de tip întrebare-răspuns ce se crează în baza documentelor IDSI în cadrul proiectului „Cercetarea în domeniul de Regăsire a Informației în scopul creării sistemului electronic de informare publică”. Trebuie de menționat că sistemul este disponibil on-line sub forma unei aplicații web. Sistemul a fost testat cu o serie de întrebări, răspunsurile fiind analizate și structurate. În urma experimentului s-a depistat că sistemul răspunde corect la 39,85% de întrebări. S-a observat că sistemul returnează relatări pozitive la interogări de tip „Nume/Persoană” cu acuratețea 0.71 și „Definiție” cu acuratețea de 0.57. Totodată, este necesară revizuirea modulului de analiză morfologică a întrebării pentru corelarea căutării răspunsului în sistem.

VI. Referințe

- Bobicev V. Preprocesarea textelor în sistemele de tip „întrebare-răspuns”. 7th International Conference on Microelectronics and Computer Science, Chișinău, 2011, p.278-282, ISBN 978-9975-45-174-1
- Botnaru R., Bobicev V. Studiul tipurilor de întrebări din sistemul de întrebare-răspuns. 7th International Conference on Microelectronics and Computer Science, Chișinău, 2011, ISBN 978-9975-45-174-1
- 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
- Tufiș D., Ștefănescu D., Ion R. and Ceaușu 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.