ETAPELE DE DEZVOLTARE A SISTEMULUI DE TIP „ÎNTREBARE-RĂSPUNS”

Victoria BOBICEV

Universitatea Tehnică a Moldovei

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, căutatrea informației, extragerea informației.*

1. Introducere

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

Cele mai eficiente metode de găsire şi de acumulare a informaţiei o reprezintă, la ora actuală, motoarele de căutara, scopul cărora este de a oferi utilizatorului un set de articole sau pagini web în care acesta să poată găsi informaţia 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 informaţia necesară.

Etapa următoare în domeniul achiziţiei informaţiei constă în dezvoltarea sistemelor capabile să răspundă la întrebările formulate de utilizator în limbajul natural. Un sistem de răspunsuri la întrebări necesită o procesare a limbajului natural mult mai complexă decât sistemele de achiziţie de documente [3][5].

2. 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 colecţii de documente, extrag răspunsul sau un set de răspunsuri găsite în documentele date. Astfel de sisteme sunt considerate ca fiind următorul pas în evoluţia motoarelor de căutare a informaţiei în surse textuale [6].

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 colecţie de documente, de obicei, este format din trei componente principale [5]:

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

2) Modul de achiziţie 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 colecţia de articole returnate de modulul de achiziţie de documente, extrage un răspuns succint şi care constituie răspunsul în limbaj natural la întrebarea utilizatorului.

Sisitemul ce se crează în cadrul proiectului „Cercetarea în domeniul de Regăsire a Informaţiei în scopul creării sistemului electronic de informare publică” la prima etapă va funcţiona în baza documentelor prestate de IDSI (Istitutul de Dezvoltare a Societăţii Informaţionale) ca, de exemplu, „Codul cu privire la ştiinţă şi inovare al Republicii Moldova”, „Acordul de parteneriat între Guvern şi Academia de Ştiinţe a Moldovei pentru anii 2009-2012”, „Lege cu privire la parcurile ştiinţifico-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.

3. Analiza întrebărilor

Metodologia clasică de analiza semantică a întrebărilor propusă în [1] presupune împărţirea propoziţiei interogative în următoarele elemente: pronume interogativ (de exemplu, cine, care, unde, cînd) sau grup prepoziţional interogativ (de exemplu, în ce, pe cine, de unde, la care) şi propoziţia inversată. Propoziţia inversată începe cu grupul verbal dupa care urmează grupul nominal. La sfîrşitul propoziţiei se presupune prezenţa 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 propoziţiilor în limba română. Exemple de împărţire a propoziţiilor engleze sunt prezentate în figura 1. Tabelul 1 conţine exemplele propoziţiilor în limba engleza şi română.

Grupul nominal în cazul dat este considerat fraza-cheie pentru căutarea răspunsului. Respectiv, întebarea se reformulează pentru obţinerea răspunsului. Grupul verbal şi cel nominal în propoziţia inversată se schimbă cu locul şi dacă întrebarea începe cu o prepoziţie aceasta se ataşează la sfîrşit. Locul liber pentru răspuns urmează dupa elementele acestea.

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].

4. Problemele rezolvate şi nerezolvate la etapa dată

- A fost rezolvată problema semnelor diacritice. Cu scopul procesării corecte a textului român cu litere speciale cu semne diacritice toate documentele sistemului au fost recodificate în UTF-8. Codificarea dată păstrează literele speciale şi asigură vizualizarea lor corectă în diferite medii cum ar fi editoare de text sau browsere.

Documentele html se codifică în UTF-8 cu ajutorul editoarelor în care sunt create. Ca regulă, editoarele de texte au opţiunea de a schimba codificarea textului în cadrul procesului de salvare. De exemplu, Notepad oferă opţiunea dată în fereastra de salvare, ultima casetă de selectare Encoding oferă o listă de codificări posibile ce include şi opţiunea UTF-8. Fragmentul de fereastră este prezentat în figura 1.

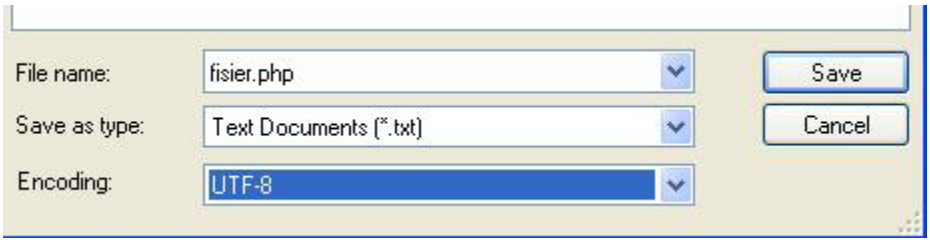

Figura 1. Fragmentul de fereastră de salvare a fişierelor în Notepad cu opţiunea de codificare a documentului.

În Notepad++ codificarea se indică cu ajutorul punctului meniului Encoding şi apoi se salveză în mod normal. Pentru vizualizarea corectă a documentelor html în browsere este necesară adăugarea tag-ului meta în secţiunea head al documentului:

<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />

În cazul editării documentelor în Dreamweaver se selectează din Modify->Page Properties->Title/Encoding->Encoding->Unicode(UTF-8), iar la Unicode Normalization Form se selecează None (secvenţa meta din head va fi automat scrisă).

Transmiterea dateor întroduse de utilizator în formular se efectuiază prin CGI – Common Gateway Protocol care ca regulă recodifică textul în mod specific. Codificarea dată depinde de proprietăţile documentului html, de atributele tag-ului form care conţine elementele de întroducere a textului, de setările serverului şi de alte condiţii. De exemplu, CGI permite transmiterea directă fără scimbări doar literelor alfabetului englez şi unui număr limitat de caratere (. , : ; ! ? @ # $ şi altelor). Restul caracterelor se codifică cu aşa-numit percent-encoding. De exemplu, litera „ţ" codificată în UTF-8 se va transmite recodificată în

%C4%83 unde semnele procentelor precedă fiecare din cele două octeți care formează litera dată în UTF-8. În afară de aceasta caractere de spații se înlocuiesc cu semnele „+”. Astfel programul care va analiza întrebarea ce a fost introdusă de utilizator în primul rînd trebuie să decodifice textul obținut.

Programul de anaiza a întrebărilor a fost scris în Perl care la fel are probleme cu codificarea UTF-8 însă problemele aceste pot fi rezolvate. În primul rînd textul programului la fel a fost codificat în UTF-8. ‘use utf8;’ la începutul programului indică codificarea şi permite utilizarea caraterelor cu diacritice în textul programului. use URI::Escape; indică că va fi utilizat modulul dat pentru recodificarea datelor obținute. use Encode 'decode_utf8'; introduce modulul pentru recodificarea datelor în UTF-8. şi în final binmode(STDOUT, ':utf8'); permite output programului în UTF-8.

După obținerea datelor din formular textul este recodificat respectiv cu:

$textul = uri_unescape ($textul); şi $textul = decode_utf8 ($textul); după ce textul arată absolut corect, este codificat în UTF-8 şi poate fi procesat de program în orice mod necesar.

Trebuie de menționa că toate fişierele textuale cu care lucrează programul sunt la fel salvate cu codificarea UTF-8 şi respectiv în program se indică faptul acesta la deschiderea lor: open(FISIER,'<:encoding(UTF8)', "$filename"); pentru citirea datelor şi respectiv open(RASPUNSURI, '>:encoding(UTF8)', "raspunsuri.html"); pentru salvarea datelor.

Toate procedeele date permit procesarea corectă a textului român scris corect cu semnele diacritice necesare.

- Există două tipuri de litere „ş” şi „ţ” cu coduri diferite: „ş”, „ş”, „ţ”, „ţ”. Verificînd toate variantele posibile am observat şi perechea de litere „ş”, „ţ” ce arată similar care am adăugat în lista literelor posibile. Respectiv cu scopul procesării corecte tuturor literelor indiferent de codul lor am întrodus modulul care înlocuieşte în întrebarea întrodusă literele „ş”, „ţ” şi „ş”, „ţ” cu „ş”, „ţ” care sunt utilizate în dicţionarul şi în documentele sistemului nostru. Astfel sistemul nu întîmpină dificultăți la căutarea cuvintelor-cheie extrase din întrebarea pusă în dicționar şi în textele utilizate pentru căutarea răspunsului.

În cazul dacă utilizatorul intenționat introduce întrebarea fără semne diacritice, înlocuind „ă”, „î”, „â”, „ş”, „ţ” cu „a”, „i”, „a”, „s”, „t” respectiv textul întrebării nu va fi procesat corect. Sistemul presupune că întrebarea este scrisă cu semne diacritice şi la compararea cuvintelor întrebării cu textul nu se va găsi coincidența. De exemplu, pentru întrebarea Ce este ştiința? întrodusă fără semne diacritice ca Ce este stiinta? se va selecta cuvîntul cheie stiinta care nu va fi găsit în text. Soluția posibila este de a cerut utilizatorul de a indica faptul că semnele diacritice sunt intenționat omise. În acest caz răspunsul se va căuta în texte cu semnele diacritice eliminate la fel ca şi în întrebare ce va permite coincidența cuvintelor şi găsirea răspunsului.

Altă problemă poate fi diferența între modurile de scriere textului: cu litere „î” sau cu litere „â” la mijlocului cuvîntului. Luînd în considerație faptele că:

- în limba română sunt acceptabile ambele metode de scriere;
- scrierea cu „â” la mijlocul cuvîntului este considerată mai modernă în România;
- limba de stat în Moldova se scrie cu „î” la mijlocul cuvîntului;
- ambele tipuri de scris sunt utilizate intens inclusiv în Moldova;

putem presupune că întebările pot fi introduse utilizînd ambele tipuri de scriere. În acest caz putem utiliza modulul creat în cadrul cercetărilor precedente dedicate transformării modurilor de scriere [7].

- În sistemul creat a fost adăugat dicționarul morfologic ce permite căutarea şi analiza tuturor formelor morfologice ale cuvintelor. De exemplu, dacă este întrodusă întrebarea „Ce este cercetarea?”, sistemul 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 dicționar morfologic a limbii române cu aproximativ 90000 de forme morfologice a cuvintelor care poate fi utilizat pentru obținerea 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 intervenția lingviştilor.

Problemele nerezolvate:

- La momentul dat sistemul procesează corect întrebările scrise cu semnele diacritice, însă dacă utilizatorul întroduce întrebarea fără semne diacritice sistemul va avea probleme analizînd astfel de text şi cel mai probabil nu va găsi răspunsul potrivit.

- Trebuie de evaluat lucrul sistemului în browsere diferite; ce puțin cele care sunt acum utilizate intens: Internet Explorer, FireFox, Google Chrome, Opera. De exemplu, parea AJAX care este încadrată în pagina cu răspunsuri nu funcționează în Internet Explorer din cauza problemei detectate în browser-ul dat: acesta nu execută AJAX corect în cazul indicării codificărilor locale cum ar fi Unicode (UTF-8). Noi am codificat situl nostru în UTF-8 cu scopul de a procesa corect textul cu semne diacritice, însă aceasta a dus la probleme cu AJAX.
- Dicționarul utilizat are unele neajunsuri: număr de cuvinte este destul de mare, ce duce la mărirea timpului de procesare; însă nu toate cuvinte care apar în text se găsesc în dicționar ce duce la probleme în căutarea răspunsului corect.
- Se procesează un număr limitat de tipuri de întrebări. La momentul de față sistemul procesează întrebările de tip definiție. 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.

5. Concluzii

În lucrarea dată sunt discutate problemele sistemul de tip întrebare-răspuns creat în cadrul proiectului „Cercetarea în domeniul de Regăsire a Informației în scopul creării sistemului electronic de informare publică”. Modulul de analiză a întrebărilor şi modulul de căutare şi selectare a răspunsului au nevoie de rezolvare a problemelor codificării literelor cu semne diacritice. La prima etapă sistemul a funcționat fără semne diacritice; la momentul dat sistemul este codificat în UTF-8 cu scopul procesării corecte a literelor cu semnele diacritice. Altă modificare efectuată este adăugarea dicționarului morfologic şi utilizarea tuturor formelor ale cuvintelor în procesul de căutare.

Bibliografie

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. Conferința ştiințifică a colaboratorilor doctoranzilor şi masteranzilor UTM, 2011.
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.
7. Bobicev V., Angheluş, V., Corectarea automată a textului român. ICTEI 2010, Chişinău, Moldova.