Perfecționarea funcționalităților sistemelor de tip întrebare-răspuns

Liviu CARCEA, Victoria BOBICEV, Tatiana BRAGARENCO
Universitatea Tehnică a Moldovei
carcea@mail.utm.md

***Rezumat* — Această lucrare reprezintă rezultatele unui studiu de analiză, cercetare, elaborare şi perfecționare a sistemelor de tip întrebare-raspuns - IR, in engleza QA-Question Answering. Este vorba de arhitectura şi principiile de funcționare a acestor sisteme: arhitectura generală a unui şistem IR cît şi arhitectura specifică avînd în vedere modul lor de funcționare. Scopul cercetarilor a fost perfectionarea functionalitatii sistemelor la toate nivelele posibile: cercetare, elaborare si implementare. Au fost introduse schimbari ce țin de perfecționarea funcționalităților - analiza morfologică a întorogărilor, algoritmii de căutare a răspunsurilor, modul de calculare a scorurilor.**

***Cuvinte-cheie* — interogare, răspuns, cuvinte cheie, sisteme de tip întrebare-răspuns.**

I. INTRODUCERE

În societatea modernă, care este una informațională, informației îi revine un rol tot mai important. Obținerea operativă a informației a devenit o necesitate zilnică şi pentru mulți cetățeni ai Republicii Moldova.

Una din caracteristicile de bază a societății informaționale este posibilitatea accesului rapid la informație pentru fiecare cetățean. Din păcate deciziile, hotărîrile, actele legislative, şi alte documente similare expuse pe Internet nu garantează această posibilitate. Deseori persoanele care au nevoie de acces la astfel de informații devin dezorientate de volumul imens de date referitoare la problema lor. Sistemul elaborat este un sistem de informare, care prelucrează întrebările formulate de către utilizator în limba româna (propoziții interogative) şi extrage răspunsul din textele relevante. [1]

Sistemele automatizate de informare a populației pot fi utilizate pe larg de orice instituție guvernamentală şi non-guvernamentală, în departamentele de relații cu publicul în scopul maximizării volumului informației obținute de populație şi minimizarea cheltuielilor de timp. În acest mod, persoanele interesate se vor putea documenta în domeniul solicitat, vor putea primi răspuns la întrebările apărute în procesul întocmirii documentelor necesare. [2] De exemplu, care acte sunt necesare pentru înregistrarea întreprinderilor mici, privatizarea locuințelor, tranzacțiilor comerciale legate de imobil, sau intrebări ce țin de condițiile necesare concediilor pentru îgrijirea copiilor etc.

Odată cu creşterea volumului tot mai mare de informații pe web, motoarele tradiționale de căutare, care returnează sute-uri şi chiar mii de documente la o interogare, cer tot mai multă răbdare din partea utilizatorului pentru ai satisface necesitățile sale informaționale. Sistemele QA pe domenii deschise, reprezintă o cercetare de top şi totodată o temă de dezvoltare în cadrul tehnologiilor lingvistice actuale. Spre deosebire de motoarele de căutare standard, care se bazează pe metode tradiționale de achiziții de informații, de la sistemele QA pe domenii deschise, se aşteaptă, nu livrarea unei liste de documente care pot fi relevante pentru cererea utilizatorului, dar a unei propoziții ori alineat care ar răspunde la întrebarea formulată în limbaj natural. [3]

Pentru construirea unui sistem de tip întrebare-răspuns se folosesc două metode:

- Abordare de tip *shallow*, bazată pe cuvinte cheie. În această metodă se folosesc cuvinte cheie pentru a găsi pasaje şi propoziții în text care ar putea reprezenta răspunsuri valide la întrebări.
- Abordarea de tip *deep*, ce implică o analiză mai sofisticată, o procesare sintactică, semantică şi contextuală.

Alegerea unuia dintre cele două modele depinde de complexitatea întrebărilor ce vor fi formulate şi de gradul de performanță dorit de la sistem. Este clar că sistemele din cea de-a doua categorie sunt superioare primelor.

II. CERINȚE ÎNAINTATE LA REALIZAREA SISTEMELOR **IR** ŞI PROBLEME ÎNTÎLNITE.

Sistemele IR trebuie să posede parametri, care corespund următoarelor cerințe:

- Timpul de răspuns. Răspunsul la o întrebare trebuie să fie furnizat în timp real, chiar dacă sistemul este accesat de mii de utilizatori. Noile surse de date trebuie să fie incorporate în sistem, de îndată ce acestea devin disponibile, oferindu-i uitlizatorului un răspuns pentru întrebarile ce se referă la un evenimet ori fapt foarte recent.
- Precizia. Precizia sistemului IR este extrem de importantă – încît oferirea răspunsurilor incorecte este un rău mai mare decît nici un răspuns. Cercetarea în domeniul IR trebuie să se concentreze asupra modalităților de evaluare a corectitudinii răspunsurilor furnizate, care cuprinde, de asemenea, metodele de detectare exactă a cazurilor în care datele disponibile nu conțin răspunsul.
- Relevanță. Răspunsul la întrebarea unui utilizator trebuie să fie relevantă într-un context specific. Complexitatea întrebării şi taxonomia legată de întrebări nu poate fi studiată fără a lua în considerare reprezentarea contextului.
- Caracterul complet. Se doreşte primirea răspunsurilor complete la întrebarea unui utilizator. Uneori, răspunsurile sunt distribuite de-a lungul unui

document sau chiar a mai multor documente în sursele de date. Se cere fuzionarea răspunsurilor într-o formă coerentă. Generarea răspunsului complet trebuie să se bazeze pe implicare, ca urmare a modului economic în care oamenii se exprimă şi dătorită datelor din care au fost extrase şi formulate răspunsurile.

III TIPURI DE INTREBARI

Toate sistemele au la bază un modul de clasificare a întrebării. Fiecare din sistemele elaborate pînă în prezent, înainte de căutarea informației conțin și etapa de clasificare a tipului de întrebare. Acest moment este necesar în scopul de a răspunde în mod corect la o întrebare. Pentru a atinge acest scop este necesar să se înțeleagă ce tip de informații cere această întrebare, deoarece cunoaşterea tipului întrebării poate oferi constrîngeri asupra zonei ce constituie date relevante (răspunsul), care ajută alte module pentru a localiza şi de a verifica în mod corect un răspuns.

Componenta de clasificare a tipului întrebării este, prin urmare, una utilă, dacă nu esențială, componenta care oferă decizii importante într-un sistem QA, cu privire la natura răspunsului solicitat. Prin urmare, întrebarea este clasificată în primul rînd după tipul: **ce, de ce, cine, cum, cînd, unde**, etc.

Deci, în arhitectura unui QA sistem, modulul de procesare a întrebării este un modul important. Avînd o întrebare în limbaj natural drept dată de intrare, funcția de bază a modulului de procesare a întrebării este de a analiza și a procesa întrebarea prin crearea unor reprezentări a informațiilor solicitate. Prin urmare, modulul de prelucrare a întrebării are menirea de a:

- analiza întrebarea, în scopul de a reprezenta principalele informații care sunt necesare pentru a răspunde la întrebarea utilizatorului.
- clasifica tipul de întrebare, de obicei, bazat pe taxonomia posibilelor întrebări deja codificate în sistem, care, la rîndul său, duce la tipul de răspuns aşteptat, prin procesarea semantică superficială a problemei.
- reformula întrebarea, în scopul de a spori frazarea întrebării şi de a transforma întrebarea în interogare pentru extragerea de informații (motor de căutare).

Aceste etape permit modulului de procesare a întrebării, ca să transmită în final, un set de termeni de interogare către modulul de prelucrare a documentelor, care le foloseşte pentru a îmbunătăți procesul regăsirii informațiilor relevante.

Analiza întrebării este, de asemenea un pas, care conduce la depistarea esenței. Din păcate, clasificarea întrebării şi cunoaşterea tipului său nu sunt suficiente pentru a găsi răspunsuri la toate întrebările. Întrebările de tip “Ce” , în special, pot fi destul de ambigue din punct de vedere al informațiilor cerute de întrebare. În vederea abordării acestei ambiguități, o componentă suplimentară, care analizează problema şi identifică esența ei este necesară. Esența întrebării a fost definită în [4] ca fiind un cuvînt sau secvență de cuvinte care indică ce informații se cer în întrebare. De exemplu, întrebarea "Care este cel mai lung fluviu din țara Galilor?" are accentul "cel mai lung fluviu ". Dacă atât tipul întrebării (din componenta de clasificare a întrebării) cît şi esența (accentul) sunt cunoscute, sistemul este capabil să determine mai uşor tipul de răspuns necesar. Identificarea esenței se poate face folosind regulile modelului de potrivire, care se bazează pe clasificarea tipurilor de întrebări. [5]

Răspunsul la întrebare este o alternativă de regăsire a informațiilor, care depistează informații detaliate, mai rapid decît documente. Un sistem QA are o întrebare în limbaj natural la intrare, converteşte întrebarea într-o interogare şi o transmite către următorul modul al sistemului IR. Cînd un set de documente necesare este depistat, sistemul QA extrage un răspuns pentru această întrebare. Există diferite metode de identificare a răspunsurilor. Una dintre ele utilizează un set predefinit de clase de entități. Avînd în vedere o întrebare selectată, sistemul QA o clasifică în acele clase care se bazează pe tipurile de entități pe care le caută, identifică entitatea în documente, şi o selectează pe cea mai aşteptată din toate entitățile din aceeaşi clasă ca şi întrebarea. Există diferite tipuri de metode disponibile pentru a clasifica întrebările. În continuare va fi descrisă o tehnică importantă pentru clasificarea întrebării.

Întrebări funcționale:

Întrebări de tip CÎND: Întrebările de tip Cînd încep cu cuvintul cheie" Cînd " şi sunt după natura lor întrebări de timp. Modelul general pentru Întrebările Cînd, este Cînd (fac | au făcut | AUX) NP VP X ", în cazul în care AUX, NP, şi VP sînt verbe auxiliare, fraze substantivale, şi expresii verbale. "|" Indică operația booleană OR, iar "X" poate fi orice combinație de cuvinte care joacă un rol nesemnificativ în determinarea tipului de răspuns. Exemplu: **Când a fost publicată legea cu privire la ştiință şi inovare al RM?**

Întrebări de tip UNDE: Întrebările de tip UNDE încep cu cuvîntul cheie "UNDE" şi se referă la locații. Acestea pot reprezenta obiecte naturale cum ar fi munți, granițe geografice, obiecte create de om, cum ar fi tractoare, cărți, stadioane sau o locație virtuală cum ar fi Internet sau loc fictiv. Modelul general pentru întrebări Unde este UNDE (do|does|did| AUX) NP VP X?”

Examplu: Unde este Italia?

- **Întrebări de tip CARE:** Modelul general pentru întrebări de tip CARE este CARE NP X ? Răspunsul aşteptat la astfel de întrebări este decis de către tipul entității NP.

- **Întrebările de tipul (CINE/a cui/CINE):** Întrebările din această categorie au modelul general **(CINE/a cui/CINE)** [do|does|did|AUX] [VP] [NP] X? Aici [cuvîntul] indică prezența opțională a termenului cuvînt în model. Aceste întrebări, de obicei, întreabă despre un individ sau instituție.

Exemplu: CINE a scris “Luceafărul”?

- **Întrebările (DE CE)**: Întrebările de acest fel, de obicei, întreabă despre o cauză ori explicație. Modelul general pentru Întrebările DE CE este: DE CE [do|does|did|AUX] NP [VP] [NP]” X”.

Exemplu: DE CE este necesar respectarea regulelor de higienă?

- **Întrebările de tip (CUM):** au două tipuri de modele de sintaxă: ‘‘CUM [do/does/did/AUX] NP VP X?” or ‘‘CUM [big|fast|long|many|much|far] X?” Pentru primul model, tipul răspunsului este explicația unui proces, în timp ce al doilea returnează un număr ca răspuns.

Exemplu: Cum poate fi obținut un concediu de creație?

- **Întrebările de tip (CE):** aceste întrebări au cîteva tipuri de modele. Cele mai generale expresii regulate pentru întrebările Ce pot fi scrise astfel ‘‘CE [NP] [do/does/did/AUX] [functional-words] [NP] [VP] X?

Examplu: Ce este necesar pentru a privatiza un imobil?

IV DESCRIEREA SISTEMULUI

Sistemul este creat cu scopul informării în limba română a populației, și funcționează pe baza metodei regăsirii informației relevante la interogarea formulată. Acest sistem prelucrează întrebările formulate de către utilizatori în limba româna (propoziții interogative) și extrage răspunsul dintr-o colecție de documente (legi).

Potențialii beneficiari sunt, practic, toți cetățenii Republicii Moldova din țară sau din afara hotarelor țării, dar și cetățenii străini care au nevoie de informații ce țin de legislația țării noastre.

Sistemul este format din următoarele module:

- Modulul de procesare a întrebării, unde au loc următoarele operații:
  - Analiza părților întrebării;
  - Formarea listei de cuvinte-cheie;
  - Reformularea întrebării în răspuns;
- Modulul de căutare a aliniatelor cu răspunsul posibil:
  - Extragerea aliniatelor – candidați;
  - Calcularea scorului pentru fiecare candidat;
  - Selectarea candidatului cu scorul maximal, care este afișat ca potențial răspuns.

Interfața, care are un aspect prietenos utilizatoului conține modulul care oferă posibilitatea de a selecta documentul în care să se caute răspunsul; fereastra pentru formularea întrebării; și două opțiuni de precizare a modului de scriere a întrebării(întrebarea a fost scrisă cu semne diacritice sau fără).

Modulul de procesare a intrebării funcționează în modul următor:

Se presupune că fiecare întrebare începe cu partea interogativă. De exemplu, “Ce este ...”, “În care document ...”, “Unde pot găsi ...”

Astfel sunt prevăzute patru elemente a părții interogative:

1. Prepoziția ("de","la", "de catre","in","");
2. Pronume ("ce", "cine", "care", "unde", "cînd", "când","cît timp", "ce documente“, “”);
3. Particula ("nu","se","nu se", "");
4. Verb ("sunt", "sînt", "va fi", "va", "are", "poate fi", "");

În baza acestora este formulat un algoritm de analiză a întrebării:

$intreb = $prep." ".$pronq." ".$part." ".$gverb; în baza căruia sistemul analizează întrebarea, considerînd că restul întrebării formează cuvintele-cheie. În exemplele ce urmează este descris modul de interpretare a întrebării, și extragerea sau analiza părților ei:

$intreb = $prep." ".$pronq." ".$part." ".$gverb; restul întrebării formează cuvintele-cheie

Exemple:

Ce este ajutorul de minimis?

pronume verbul cuvinte-cheie

Ce stabileşte legea cu privire la ajutorul de stat?

pronume verbul cuvinte-cheie

După analizarea părților întrebării are loc reformularea întrebării. Ea se reformulează în:

$restmax." ".$partmax." ".$gverbmax." ".$restverbmax." ".$prepmax;

cuvinte-cheie + particula + verbul + verb de baza + prepoziția

Exemple:

Ce stabileşte legea cu privire la ajutorul de stat?

pronume verbul cuvinte-cheie

Se reformulează în:

legea cu privire la ajutorul de stat stabileşte …

Căutarea răspunsului

- La momentul actual răspunsul se caută numai într-un document. Utilizatorul trebuie să selecteze documentul cu legea care îl interesează.
- În documentul selectat se caută cuvintele cheie obținute în urma analizei înrebării puse.
- Căutarea se efectuiază pe aliniate.
- Dacă într-un aliniat se găsește un cuvînt cheie acest aliniat se memorizează în lista răspunsurilor posibile. [6]

Problemele depistate:

- În urma analizei efectuate, putem concluziona că la tipul de întrebări unde sînt verbele auxiliare: a fi, a avea, și cele modale: a putea, a trebui, uneori primim un răspuns corect , alteori nu-l primim. La analiza morfologică a elementelor întrebării se produc erori la verbe, și chiar verbul de bază. Acolo unde avem verbele de bază: a fi, a avea, a putea, a trebui, el este considerat drept verb auxiliar și aceasta duce la erori în continuare. La fel sînt extrase prepozițiile, care nu joacă nici un rol în extragerea răspunsului.
- În întrebările cu verbul la timpul viitor se extrage greșit verbul și această eroare se extinde asupra restului întrebării.
- La tipul de întrebări “care masuri nu constituie ajutor de stat ?, pe ce termen se elibereaza permisul de sedere?” deasemeni nu primim răspunsuri exacte. Adică întrebările care au un substantiv în fața verbului provoacă dificultăți la extragerea părților întrebării, cuvintelor cheie și extragerea răspunsului.

Unele răspunsuri chiar dacă sînt cele corecte , au un scor mai mic decît cel extras ca primul răspuns, și utilizatorul este nevoit să-l caute în alte documente.

IV PERFECTIONAREA FUNCȚIONALITĂȚILOR

Sistemul a fost perfecționat la nivel de următoarele module:

- **Modulul de procesare a întrebării**, unde au loc următoarele operații:
  - Analiza morfologică a întrebării:
  - Analiza părților întrebării:

În varianta inițială se forma structura întrebării.

În varianta propusă se analizează structura întrebării, şi a fiecărei parţi de vorbire a propoziţiei pentru a determina corect elementele întrebării:

- Formarea listei de cuvinte-cheie;
- Reformularea întrebării în răspuns;

- **Modulul de căutare a aliniatelor** cu răspunsul posibil:
  - Extragerea aliniatelor – candidaţi şi memorizarea în fişierul cu răspunsuri
  - Calcularea scorului pentru fiecare candidat;
- **Modulul de selectare a răspunsului**.
  - Formarea răspunsului selectat;
  - Selectarea candidatului cu scorul maximal, afişarea ca răspuns potenţial.

Algoritmul modificat conţine următorii paşi

1. Se citeşte lista cu documente şi se afişează lista legilor în care se va căuta răspunsul. Se înregistrează data şi ora cînd a avut loc interogarea. Se citesc şi parsează datele de intrare.
2. Procesarea întrebării se efectuează în trei etape consecutive în rezultatul cărora are loc formalizarea interogărilor în limbajul natural. Astfel, etapele includ:
3. analiza morfologică a părţilor întrebării şi se afişează aceste rezultate.

Se analizează structura întrebării pentru fiecare cuvînt dacă este prepoziţie, pronume interogativ, grup nominal, grup verbal, cuvintele_cheie, pe rînd , memorizate în fişierul cu rezultate.

Această parte a programului a fost îmbunătăţită faţă de varianta precedentă care producea erori la analizarea părţilor propoziţiei. Acest lucru a fost posibil prin introducerea secvenţei de program care va analiza grupurile nominale (substantivele), ceea ce conduce la obţinerea unor răspunsuri mai calitative la anumite tipuri de întrebări, şi anume cele ce conţin grupul nominal înaintea grupului verbal.

Următorul pas este reformularea întrebării în propoziţie pentru a căuta acest răspuns în textul documentului.

4. După ce s-a reformulat întrebarea, s-au găsit cuvintele-cheie are loc căutarea răspunsului parcurgînd fragmente de texte (capitolele, articolele şi alineatele), în baza cuvintelor cheie obţinute în urma procesării întrebării.

În cazul în care 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 potrivit răspuns la întrebarea sa, sistemul calculează ponderea (scorul) pentru fiecare fragment memorat cu ajutorul formulei care ia în considerare paragraful în care se efectuează cautarea, ponderea paragrafului, numărul de ordine a frazei-cheie în lista frazelor-cheie, lungimea fragmentului paragrafului (numărul de caractere) înainte de fraza-cheie găsită de la primul caracter pînă la fragmentul găsit şi lungimea cuvîntului-cheie găsit (numărul de caractere).

5. Se memorizează răspunsurile în fişierul de ieşire şi se sortează conform scorului.
6. Se memorizează răspunsurile în fişierul cu raspunsuri, pentru a putea fi afişate în pagina html. şi ele sînt aranjate conform scorului calculat. Sistemul afişează răspunsul care are ponderea maximă, însă celelalte răspunsuri sunt înregistrare într-un fişier disponibil pentru utilizator prin accesarea link-ului apărut pe aceeaşi pagină. Sistemul afişează în acelaşi timp titlul, capitolul şi articolul din actul legislativ în care a fost găsit răspunsul.

V CONCLUZII

În scopul creării unei societăţi informaţionale în Republica Moldova este realizat proiectul ”Cercetarea în domeniul de Regăsire a Informaţiei în scopul creării sistemului electronic de informare publică”. Obiectivele ştiinţifice ale proiectului propus sunt de a cerceta problema regăsirii informaţiei relevante la interogarea expusă şi de a crea un sistem automat de informare a populaţiei în limba română. Sistemul de informare va prelucra întrebările formulate de către persoane conform gramaticii limbii române (propoziţii interogative) şi va extrage răspunsul din textele relevante.

În urma analizei răspunsurilor s-au depistat erori în analiza morfologică a întrebării, îndeosebi la verbele care au şi rol de verb auxiliar sau modal. În special verbele: *a fi, a avea, a putea, a trebui*. Aceste erori duceau la răspunsuri incorect sau nesatisfăcătoare.

Întrebările care au un substantiv în faţa verbului provocau dificultăţi la extragerea părţilor întrebării, cuvintelor cheie şi extragerea răspunsului.

În urma perfecţionării algoritmului de analiză aceste probleme au fost înlăturate şi sistemul oferă răspunsuri relevante.

BIBLIOGRAFIE

[1] Carcea L. Abordări în dezvoltarea sistemelor ”întrebare – răspuns” In Proceedings of the 7th International Conference on „”Microelectronics and Computer Science” (ICMCS-2011), Chişinău, 2011

[2] Ştefănescu, D., Ion, R., Ceauşu, A., Tufiş, D. Sistem întrebare-răspuns antrenabil pentri limba română. În lucrările conferinţei ”Resurse lingvistice şi instrumente pentru prelucrarea limbii române” Bucureşti, 6-7 mai 2010 p.153-164.

[3] M. Murata, M.Utiyama. & H.Isahara, (2000). Question Answering System Using Similarity-Guided Reasoning, IPSJ SIG Technical Reports, NL-135-24

[4] D.Moldovan, Lasso: A Tool for Surfing the Answer Net. In Proceedings of the Eighth Text Retrieval Conference (TREC-8) 1999.

[5] R Botnaru, V. Bobicev Studiul tipurilor de întrebări din sistemul de întrebare-răspuns. In Proceedings of the 7th International Conference on ”Microelectronics and Computer Science” (ICMCS-2011), Chişinău.

[6] V. Bobicev Preprocesarea textelor în sistemele de tip ” întrebare-răspuns”. . In Proceedings of the 7th International Conference on ”Microelectronics and Computer Science” (ICMCS-2011), Chişinău.