Abordări în dezvoltarea sistemelor ”întrebare – răspuns”

Liviu Carcea
Technical University of Moldova
carcea@mail.utm.md

***Rezumat* – acest articol este o încercare de analiză a abordărilor ale sistemelor ”întrebare-răspuns”(SIR). În ultimii ani se observă un interes deosebit pentru sistemele “întrebare-răspuns”, care a condus la o schimbare calitativă a funcţionării acestor sisteme. Direcţiile principale de cercetare ţin de formularea întrebărilor în limbaj natural, utilizarea motoarelor de căutare, evaluarea calităţii răspunsurilor şi au provocat apariţia abordărilor deosebite de cele tradiţionale.**

Cuvinte-cheie: abordare, sisteme ”întrebare-răspuns”, regăsirea informaţiei, interogare, motor de căutare.

I. INTRODUCERE

Sistemele de tip ”întrebare–răspuns” sunt complexe şi utilizează în general mai multe moduri de analiză şi tipuri de resurse. Crearea suportului informatic pentru acest gen de sisteme necesită implicare a specialiştilor din domeniile atât informaticii cît şi a lingvisticii.

Obiectivele ştiinţifice ale cercetărilor din acest domeniu sunt atât cercetarea cât şi analiza abordărilor problemei regăsirii informaţiei relevante la interogarea expusă şi crearea unui sistem automat de informare a populaţiei în limbaj natural, inclusiv şi 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.

Evaluarea calităţii răspunsurilor şi respectiv asigurarea găsirii lor sunt însă probleme de cercetare, pentru care abordările tradiţionale au devenit insuficiente [7]. Astfel, implementările motoarelor de căutare moderne recurg din ce în ce mai des la tehnicile PLN (Procesarea Limbajului Natural), acestea fiind utilizate în toate etapele fluxului de prelucrare, începând de la nivelul specificării întrebării şi până la extragerea fragmentului de text relevant. Concomitent au apărut şi campaniile de evaluare în domeniul regăsirii inteligente a a informaţiei. Acestea constituie astăzi priorităţi ale cercetării de avangardă dedicată spaţiului digital al cunoaşterii. Ele au fost organizate iniţial în SUA (MUC-*Message Undestanding Coference*, TREC-*Text Retrieval Conference*, DUC-*Document Undestanding Conference*). În Europa, manifestarea similară este CLEF(*Cross Language Evaluation Forum*), care are ca subiect în primul rând limbile Uniunii Europene, inclusiv şi cele cu resurse electronice limitate. În acest context proiectul naţional din România SIR-RESDEC lansat în 2007 a răspuns unei priorităţi europene, propunându-şi realizarea unui sistem de ”întrebare-răspuns” în limbaj natural.

Conform planului de acţiuni „eEurope – O societate informaţională pentru toţi”, în Uniunea Europeană a fost propusă lista de 20 de servicii publice în formă electronică prestate cetăţenilor. În Moldova la fel se acordă o atenţie sporită dezvoltării serviciilor electronice. Sistemele automatizate de informare prezintă o parte-componentă a setului de servicii electronice

II. METODE FOLOSITE IN DEZVOLTAREA SISTEMELOR “INTREBARE-RASPUNS”

Cele mai eficiente metode de descoperire şi de achiziţie a informaţiei o reprezintă, în prezent, **motoarele de căutare**. Scopul acestora este de a servi utilizatorului un set de articole în care acesta să poată găsi informaţia care îi este necesară. De multe ori, articolele servite de motoarele de căutare nu îndeplinesc deziderатul de a răspunde cerinţei utilizatorului, de a conţine un răspuns care să poată fi considerat satisfactor. Deasemenea, ele nu oferă răspunsul concret la problema utilizatorului, ci doar un set de articole din care utilizatorul trebuie să extragă singur informaţia căutată.

Pasul următor în domeniul achiziţiei informaţiei e constituit de dezvoltarea sistemelor capabile să răspundă la întrebări formulate de utilizator în limbaj natural. Dezideratul principal al unui astfel de sistem este să asigure un răspuns la întrebarea utilizatorului care sa îndeplinească urmatoarele trei condiţii:

- **să fie corect,**
- **să fie formulat tot în limbaj natural uman şi**
- **să fie suficient de succint.**

Un sistem de răspuns la întrebări necesită o procesare mai complexă a limbajului natural decât sistemele de achiziţie de documente (cum sunt unele motoare de căutare a informaţiei).

În teorie, procesarea limbajului natural e un subiect foarte atractiv, datorită aplicabilităţii sale în domenii ca cel al interacţiunii om-maşină. În practică se constată, însă, o serie de limitări majore, datorate, mai ales, modului diferit în care o afirmaţie în limbaj natural poate fi interpretată şi a multitudinii de sensuri pe care cuvintele constituente le pot lua.

Pentru construirea unui sistem de răspuns la întrebări avem două variante:

- Abordare de tip **shallow**, bazată pe cuvinte cheie. În această metodă se folosesc cuvinte cheie pentru a gasi pasaje şi propoziţii în text care ar putea candida drept răspunsuri valide la întrebări. Aceste potenţiale răspunsuri urmează să fie analizate mai apoi în profunzime pentru a se stabili dacă sunt răspunsuri reale sau nu. Acestă metodă poate fi folosită cu succes în cazul întrebărilor scurte, factuale, când se caută nume, date, locaţii, cantităţi.

- Abordarea de tip **deep**, ce implică o analiză mai sofisticată, o procesare sintactică, semantică şi contextuală. Există o serie de metode ce pot fi încadrate în această categorie: *abduction, named-entity recognition, relation detection,* etc.

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

III. MODELE DE ABORDĂRI

Una din cele mai importante etape în sistemele întrebare/răspuns este analiza întrebărilor. De fapt întrebările sunt exprimate în limbaj natural spre deosebire de căutările clasice în care interogările sunt prezentate prin cuvinte cheie [3]. Rezultă că una din etapele de bază constă în extragerea cuvintelor cheie din întrebare pentru a formula una sau mai multe interogîri. Altă funcţie a analizei întrebării este determinarea tipului răspunsului aşteptat. Deobicei tipul este formalizat avînd forma unei entităţi. De exemplu: PERSOANA, NUMARUL, LOCUL etc.

Astfel întrebarea " *Cîte facultăţi are UTM* ? " va conduce la cuvintele cheie "*facultăţi/UTM*" şi se va aştepta un răspuns de tip "*numarul (*de facultăţi*)*".

În general analiza întrebării se face pe trei nivele. Pentru început sunt detectate entităţile întrebării. Apoi se va determina tipul întrebării şi deci şi tipul răspunsului. În final se aplică o analiză sintactică pentru a determina legăturile dintre elementele întrebării.

În [6] este prezentat un sistem care răspunde la întrebări formulate în limbaj natural în baza unui corpus de date cu posibilităţile următoare:

- Dezambiguizarea întrebărilor.
- Se va ţine cont de contextul general al întrebării.
- Extragera răspunsurilor.
- Asigurarea interacţiunii cu utilizatorul ce va ajuta la justificarea răsunsului.
- Reformularea răspunsului în caz de fuziune a documentelor.
- Răspunsul va fi obţinut în timp real.

Pentru a obţine aceste obiective s-a făcut apel la tehnicile următoare:

- Căutarea informaţiei şi procesarea limbajului natural
- Raţionament în bază de caz (Case-Based Reasoning)
- Tehnicile inteligenţei artificiale (multiagent)

Menţionăm faptul că sistemul nu ţine cont de profilul utilizatorului şi care reduce posibilităţile sale destul de avansate.

Aplicaţia PIQUANT prezentată la TREC12 [2] a păstrat toate performanţele versiunilor precedente la care a adăugat noi psibilităţi. Prelucrarea începe cu analiza întrebării, care la rîndul său implică mai mai multe acţiuni cum ar fi analiza sintactică, identificarea entităţilor şi altele.

Analiza întrebării produce un QFRAME, care conţine tipul răspunsului dorit, tipul întrebării, cuvintele cheie din întrebare şi o formă sintactică simplă. La etapa următoare sistemul transmite QFRAME generatorului QPLAN, care de fapt generează lista agenţilor disponibili. Această listă este transmisă modulului "*Rezoluţie şi răspuns*", care alege răspunsul adecuat în funcţie de tipul întrebării.

PIQUANT utilizează o tehnică numită *QA-par Dossier* pentru a justifica răspunsul. Totuşi sistemul este limitat datorită lipsei de interacţiune cu utilizatorul ceea ce nu-i permite să ia o decizie în caz de ambiguitate.

Sistemul descris în [1] are ca principiu de funcţionare utilizarea multiplelor surse de date, de fapt doi agenţi independenţi de căutare: unul pentru Internet şi altul pentru corpusul TREC. Acest lucru îi permite sistemului să obţină scoruri înalte de certitudine pentru răspunsuri. Sistemul conţine modulele următoare: analiza întrebării, selecţia documentului, identificarea entităţilor şi extragerea răspunsului. Dat fiind numărul foarte mare de documente pe Web şi redundanţa lor ridicată sistemul are şanse mari de a găsi răspunsul exact. Totuşi lipsa interacţiunii cu utilizatorul şi aplicarea unui algoritm unic pentru a răspunde la întrebări de tip diferit prezintă un obstacol pentru rezolvarea problemei evaluării răspunsului. În [4] se ţine cont de rolul cuvîntului în întrebare. Autorii menţionează următoarele trei roluri:

- **Focus** – este entitatea asupra căreia punctează întrebarea şi pentru care o caracteristică sau o definiţie este căutată. Prin definiţie acest element trebuie să fie prezent şi în textul răspunsului. În întrebarea "*Ce sport practică Zinedine Zidane*?" focusul este "*Zinedine Zidane"***;**
- **Domeniul –** care nu este întotdeauna prezent în întrebare, dar de obicei precizează tipul răspunsului. În întrebarea precedentă domeniul este "*sport*";
- **Verbul principal** – este vorba de verbul prezent în întrebare şi care are un rol important în răspuns atunci când introduce un fapt sau o acţiune. În întrebarea precedentă verbul principal este "*practică*".

Prelucrarea întrebărilor în [7] conţine următorii paşi. După primirea întrebării este apelat serviciul web care se ocupă de clasificarea întrebărilor pentru a obţine tipul de răspuns care trebuie căutat. La următorul pas folosind informaţia adnotată după procesare, întrebarea este transformată în interogări într-un limbaj formal înţeles de motorul de căutare. Autorii folosesc doi algoritmi pentru a genera două interogări diferite, ambele conţinând ca termen de căutare clasa întrebării. Se menţionează faptul că în faza de indexare au fost indexate odată cu paragrafele şi clasele corespunzătoare acestora, pentru ca reducerea spaţiului de căutaresă să se facă direct din faza de regăsire documentară.

Etapa de regăsire documentară este similară cu cea descrisă în [5], adică cuvintele documentelor au fost filtrate în funcţie de descrierea lor morfo-sintactică şi ulterior normalizate la forma lor lemă. Autorii au folosit 2 algoritmi de generare a interogărilor formale către motorul de căutare pentru ca sistemul să poată întoarce pentru unele întrebări mesajul că niciun răspuns nu a fost găsit.

Analiza abordărilor de dezvoltare a SIR demonstrează utilizarea complexă şi profundă a tehnologiilor moderne de prelucrare a textelor în limbaj natural.

IV. CONCLUZII

În societatea modernă, care este o societate informaţională, informaţiei îi revine un rol important. Obţinerea operativă a informaţiei a devenit o necesitate zilnică pentru mulţi cetăţeni ai Moldovei.

În cadrul proiectelor instituţionale de cercetare ştiinţifică şi în scopul creării unei societăţi informaţionale în Republica Moldova este lansat proiectul "Cercetarea în domeniul de Regăsire a Informaţiei în scopul creării sistemului electronic de informare publică", realizat de către cercetătorii Laboratorului de Inginerie a Limbajului Uman (LILU) al catedrei Informatica Aplicată facultatea Calculatoare, Informatica şi Microelectronica a Universităţii Tehnice a Moldovei.

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.

Crearea sistemului necesită o implicare a specialiştilor în lingvistică şi informatică, care activează în echipa noastră. Pentru sistemul dat vor fi utilizate metodele lingvistice de procesare a propoziţiilor, dar şi metode statistice de regăsire a informaţiei în volume mari de texte.

Deşi există deja sisteme/medii care oferă posibilitatea de a pune întrebări în limbaj natural şi de a obţine răspunsuri extrase dintr-un volum de texte, nu există deocamdată un mediu complet care să satisfacă toate cerinţele în aplicaţiile practice. Mai mult, nu cunoaştem nici o aplicaţie care să folosească limba română. Astfel, sistemul propus prezintă un grad de noutate şi complexitate semnificativ şi are ca scop dezvoltarea de aplicaţii reale, şi anume, un sistem de informare publică. Potenţialii beneficiari sunt practic toţi cetăţenii Republicii Moldova din ţară sau din afara graniţelor, dar şi cetăţeni străini care vor avea nevoie de informaţie din ţara noastră.

În baza abordărilor de elaborare şi dezvoltare a SIR, inclusiv şi cele netradiţionale, dar şi a cerinţelor formulate de către comandatarii proiectului un astfel de sistem trebuie să conţină:

- Unitatea de analiză a întrebărilor în limba română şi de formare a interogării pentru căutarea răspunsului;
- Unitatea de căutare a textelor relevante pentru întrebarea dată;
- Unitatea de detectare a fragmentelor textelor ce conţin răspunsul posibil;
- Unitatea de formare şi vizualizare a răspunsului;
- Unitatea de analiză a întrebărilor, răspunsurilor şi de evaluare a lor de către utilizatori.

Sistemul necesită implicarea specialiştilor din diferite domenii. Unitatea de analiză a întrebărilor şi de formare a răspunsului, sunt realizate aplicând metode de analiză lingvistică, pe când unitatea de căutare a textelor relevante şi unitatea de detectare a fragmentelor textelor ce conţin răspunsul se realizează prin metode statistice.

Unitatea regăsirii informaţiei relevante la interogarea expusă este cea principală în sistemul dat şi afectează imens performanţa lui. Astfel, în proiectul propus, se va:

1) cerceta metodele de regăsire a informaţiei (Information Retrieval - IR),

2) elabora un sistem de informare care va căuta răspunsuri în colecţia de documente, utilizând tehnologiile avansate de regăsire a informaţiei,

3) examina impactul mai multor factori care afectează performanţa sistemului prin exploatarea lui experimentală on-line.

Sistemele automate de informare a populaţiei pot fi utilizate pe larg de orice instituţie guvernamentală şi non-guvernamentală, în departamentele de relaţii cu publicul în scopul maximizării volumului informaţiei obţinute de populaţie în paralel cu minimizarea cheltuielelor 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 pentru, de exemplu, înregistrarea întreprinderilor mici, privatizarea locuinţelor, tranzacţiilor comerciale legate de imobil etc.

BIBLIOGRAFIA

[1] Chalendar, G., Dalmas, T. LIMSI-CNRS, The Question. Answering System QALC at LIMSI Experiments in using Web and Wordnet, In Proceedings of the 11 th Text REtrieval Conference (TREC-10), 2002. http://trec.nist.gov/pubs/trec11/t11_proceedings.html

[2] Chu-Carrol, J., Prager, J., Welty, C.,Czuba, K., Ferrucci, D. IBM T.J. Watson Research Center IBM's PIQUANT in TREC2003 (TREC-12),2003 http://trec.nist.gov/pubs/trec12/t12_proceedings.html

[3] Crestan, E., Lemaire, E., Claude de Loupy, Ressources pour un syteme de Question/Reponse. Actes de TALN 2004. http://edutice.archives-ouvertes.fr

[4] Grappy, A., Ligozat, A., Grau, B. Evaluation de la reponse d'un systeme de question –reponse et de sa justification. COnference en Recherche d'Information et Application, CORIA 2008, Traitement automatique de langues et résumé automatique p. 273-288.

[5] Ion, R., Ştefănescu, D., Ceauşu, A. and Tufiş D. (2009) RACAIţs QA System at the Romanien-Romanien QA@CLEF2008 Main Task, Lecture Notes in Computer Science, vol. 5706, september, p. 393-400.

[6] Merdaoui, B., Frasson, C. QUERI: Un systeme de question-reponse collaboratif et interactif. http://edutice.archives-ouvertes.fr

[7] 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.