Metode utilizate la elaborarea unui sistem întrebare – răspuns (QA)

Victoria MAXIM
Technical University of Moldova
maxivica@yahoo.com

***Abstract* — In information retrieval and natural language processing (NLP), question answering (QA) is the task of automatically answering a question posed in natural language. To find the answer to a question, a QA computer program may use either a pre-structured database or a collection of natural language documents (a text corpus such as the World Wide Web or some local collection). QA research attempts to deal with a wide range of question types including: fact, list, definition, How, Why, hypothetical, semantically constrained, and cross-lingual questions. Search collections vary from small local document collections, to internal organization documents, to compiled newswire reports, to the World Wide Web. QA is regarded as requiring more complex natural language processing (NLP) techniques than other types of information retrieval such as document retrieval, thus natural language search engines are sometimes regarded as the next step beyond current search engines. [3]**

***Index Terms* — Answer Extraction, Information Retrieval, Typical Question Answering System, Question Answering, Question Classification.**

I. INTRODUCERE

Sistemele de întrebare - răspuns (în engleză "question answering systems", sau **sisteme QA**) sunt considerate ca fiind următorul pas în evoluţia motoarelor de căutare a informaţiei. Sistemele de tip QA sunt caracterizate prin faptul că primesc un set de întrebări în limbaj natural şi, pe baza unei colecţii de documente, trebuie să extragă răspunsul sau răspunsurile. Această colecţie poate varia de la o simplă colecţie locală până la întregul World Wide Web.

II. GENERALITĂŢI

La ora actuală Internetul reprezintă, fără îndoială, cea mai mare bază de cunoştinţe, aflată într-o continuă extindere şi actualizare. El este, în acelaşi timp, una dintre cele mai accesibile locaţii în care aceste cunoştinţe pot fi consultate. Dar gradul de dezvoltare a Internetului are şi aspecte negative: datorită multitudinii de informaţii disponibile, găsirea informaţiei necesare la un moment dat poate fi dificilă sau şi nesigură.

Cele mai eficiente metode de descoperire şi de achiziţie a informaţiei o reprezintă, în prezent, ***motoarele de căutare***. Scopul acestora este de a oferi utilizatorului un set de articole sau pagini web în care acesta să poată găsi informaţia care îi este necesară. De multe ori articolele oferite de motoarele de căutare nu îndeplinesc dezideratul utilizatorului de a obţine un răspuns satisfacator. Deasemenea, ele nu oferă răspunsul concret la problema utilizatorului, ci doar un set de pagini web, din care utilizatorul trebuie să extragă singur informaţia căutată.

Pasul următor in 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 să îndeplinească urmatoarele trei condiţii:

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

Un sistem de răspuns la întrebări necesită o procesare a limbajului natural mult mai complexă decât sistemele de achiziţie de documente.

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 dificultăţi 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. Sistemele de răspuns la întrebări, văzute ca un subdomeniu al procesarii limbajului natural, moştenesc problemele acestora.

III. METODE UTILIZATE LA ELABORAREA UNUI SISTEM QA

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

- ***Abordare de tip shallow, bazată pe cuvinte cheie.*** În această metodă se folosesc cuvinte cheie pentru a găsi pasaje şi propoziţii în text care ar putea reprezenta răspunsuri valide la întrebări. Aceste potenţiale răspunsuri urmează să fie analizate apoi mai 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.

Arhitectura generală a unui sistem de tip întrebare-răspuns

Dacă la începuturile inteligenţei artificiale, în anii 1960, cercetătorii erau fascinaţi de ideea de a putea construi sisteme capabile să răspundă la întrebări aparţinând unor domenii restrânse (closed domains), în prezent dezvoltarea Internetului şi paşii făcuţi în ceea ce priveşte achiziţiei informaţiei (information retrieval - IR) şi a tehnicilor de procesare a limbajului natural (natural language processing - NLP), precum şi cererea pentru acces facil la informaţie, a dus la creşterea interesului pentru sisteme care să ofere răspunsuri din domenii largi (open domains).

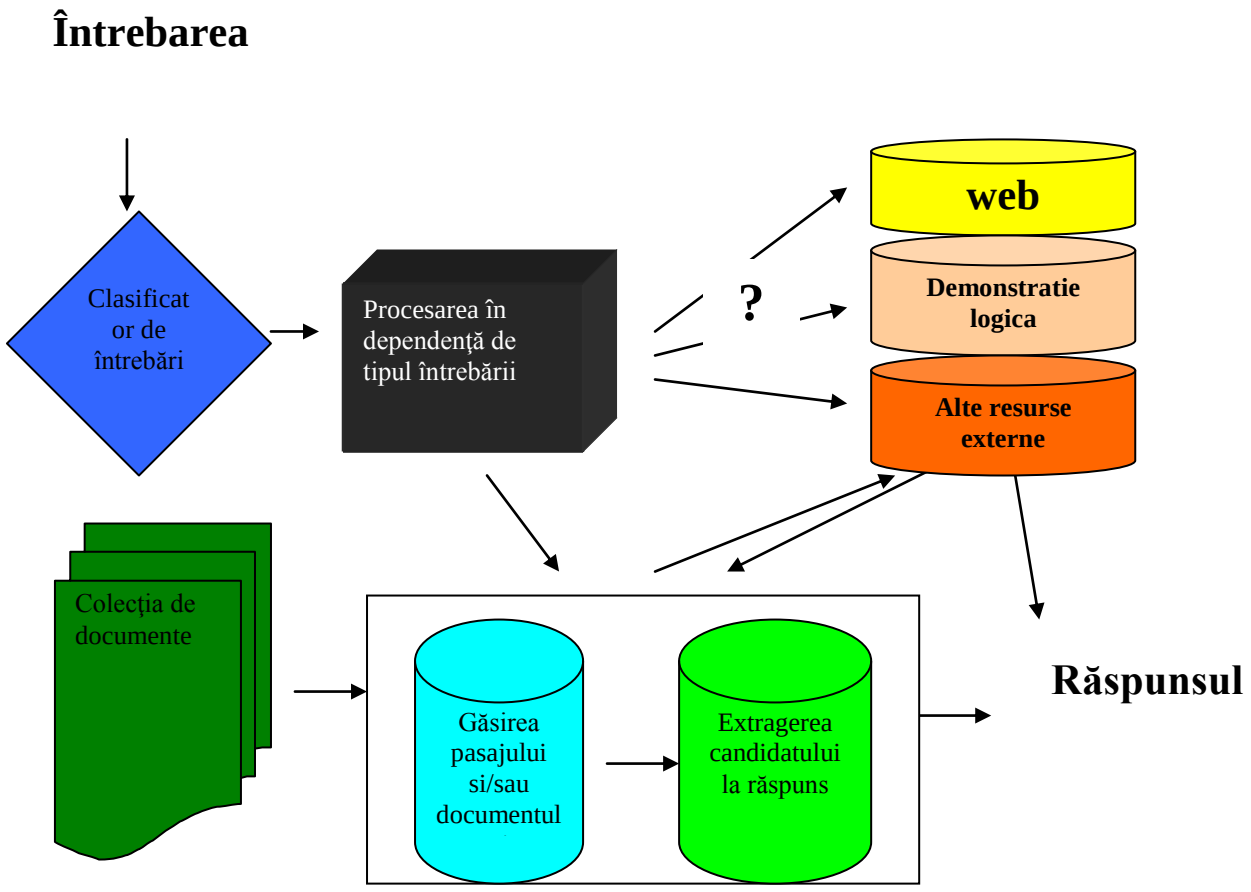


Figura 1. Un sistem QA tipic [2]

Un sistem de răspuns la întrebări bazat pe o colecţie de documente are, în mod tipic, trei componente principale:

1. ***Modulul de analiză a întrebării*** – transformă întrebările formulate în limbaj natural uman în interogări pentru motorul de achiziţie de documente;
2. ***Modulul de achiziţie de articole*** – caută în colecţia de articole articolele relevante pentru întrebarea formulata de utilizator, pe baza datelor primite de la modulul de analiză a întrebării;
3. ***Modulul de extragere a răspunsului*** – din colecţia de articole returnate de modulul de achiziţie de articole, extrage un răspuns succint şi care constituie răspunsul în limbaj natural uman la întrebarea utilizatorului. Dacă un astfel de răspuns nu există în colecţia de documente considerată de modulul de achiziţie

de articole, e de preferat ca sistemul să nu răspundă la întrebare, în loc de a întoarce un răspuns eronat.

Chiar dacă Internetul este un mediu plin de informaţii din toate domeniile, găsirea unui răspuns la o întrebare simplă poate fi uneori o sarcină dificilă. Unele din dificultăţile ce pot să apară în dezvoltarea unui astfel de sistem:

Formularea corectă a interogărilor. Transformarea unei întrebări din limbaj natural într-o interogare pentru un motor de căutare este o sarcină dificilă. Dacă întrebarea e prea genrală, va fi extras un număr prea mare de documente. Deasemenea, temele descrise în colecţia de documente extrase e posibil să nu conţină tocmai răspunsul la întrebarea utilizatorului. Dacă sunt extrase prea multe documente, timpul de procesare va fi sporit. Dacă setul de cuvinte căutate e prea mic, e posibil să nu fie găsit articolul care răspunde la întrebare. De aceea se cere ca setul de cuvinte folosite în intergoare să fie bine formulat, pentru a fi returnate, pe cât posibil, doar documente care să conţină informaţie utilă.

Chiar dacă e găsit setul de cuvinte corecte pentru a realiza o interogare care să aibă posibilitatea să întoarcă articole cu informaţie utilă, motorul de căutare poate întoarce un număr foarte mare de articole care să nu răspundă întrebării utilizatorului. De exemplu, în cazul întrebării „Cine a fost primul om în spaţiu?", modulul de achiziţie de informaţii va primi, printre altele, şi următoarele cuvinte cheie după care să realizeze căutarea: „primul om în spaţiu". Motorul de căutare e posibil să aibă indexate pagini despre „primul turist în spaţiu" sau despre „cel mai bătrân om în spaţiu" şi să le returneze şi pe acestea, împreună cu articolele care conţin răspunsul corect, şi anume cele despre Iuri Gagarin, primul cosmonaut în spaţiu.

Informaţii false. Chiar dacă întrebarea e bine formulată, motorul de achiziţie de articole întoarce articole care corespund subiectului întrebării, e posibil ca unele dintre aceste articole să conţină informaţii greşite. Acesta este unul dintre cazurile cele mai nefavorabile, deoarece sistemul nu are cum să îşi dea seama care răspunsuri sunt corecte şi care nu.

Resurse limitate. Când este construit un sistem de răspuns la întrebări trebuie să se ţină cont de limitările impuse de lucrul cu cantităţi mari de informaţii. Este neindicat să se trimită sistemului un set prea mare şiruri de cuvinte pentru interogări. Cu toate ca motoarele de căutare actuale sunt suficient de rapide şi întorc răspunsuri la obiect, căutarea în liste prea mare de articole consumă mult prea mult timp, şi utilizatorul sistemului nu este dispus să aştepte minute pentru ca sistemul să îi ofere un răspuns.

IV. CLASIFICAREA INREBĂRILOR

Un sistem QA caută expresia sau enunţul (sau pasajul, documentul, setul de documente) care este răspunsul exact la o întrebare.

• întrebările au multe nuanţe;

• majoritatea căutărilor sunt axate pe întrebări faptice;

• răspunsurile pot fi enunţuri sau expresii;

• răspunsurile complete ar trebui să fie găsite într-o sdingură sursă.

Tipuri de întrebări:

- ✓ Faptice : Cine l-a omorât pe Martin Luther King?
- ✓ Sarcini : Cum pot să aplic pentru un paşaport?
- ✓ Opinii : Care a fost cel ma bun film anul acesta?
- ✓ Definiţii : Cine este Jane Goodall?
- ✓ Liste : În ce filme s-a produs Jude Law?
- ✓ Explicaţii : Care a fost motivul războiului din Koreea?
- ✓ Da-Nu : Este legal să mergi la culoarea roşie a semaforului?

Exemple de întrebări:

1. *Aspartame mai este cunoscut şi ca?*
2. *La ce vârstă Rossini a încetat să mai scrie opere?*
3. *În ce zi este sărbătorită Ziua Boxului?*
4. *Defineşte Thalassemia.*
5. *Cât de mare este Galaxia noastră în diametru?*
6. *Cât de rece trebuie să fie un frigider?*
7. *CPR este acronimul la ce?*
8. *Cât durează gestaţia la om?*
9. *Ce lungime are râul Nistru?*

Exemple de întrebări faptice:

Q: Când s-a născut Mozart?
A: 1756.
Q: Ce ste un nanometru?
A1: O miliardime de metru.
A2: O milionime de milimetru.
Q: Când a avut loc Marea Depresie?
A1: Anii 1930.
A2: 1931.
A3: 1932.
Q: Cine este Abesalom?
A1: lider afro-american, primul negru capitan de navă de vânătoare de balene.
A2: Fiul lui David (biblic), care la- trădat pe tatăl său.

Clasificarea întrebărilor faptice după tipul răspunsului aşteptat/scontat.

Există 2 abordări de bază:

1. Clasificarea

Avantaj – uşor de înţeles/implementat, sigur.
Dezavantaj – nu oferă o altă informaţie decât categoria.
Expresii uzuale
Avantaj – oferă informaţii suplimentare la categorie.
Dezavantaj – foarte incert/nesigur.
Categorii de clasificare: etichete PDV, cuvinte, WordNet, cuvinte WH, arbori sintactici, taguri NE.
Expresii uzuale

- cuvinte simple WH
- compuse tipologia QA

precizie/acurateţe 90%.

2. Procesarea în dependenţă de tipul întrebării

Rescrierea întrebării pentru web

Transformarea întrebării într-o cerere, combinarea probelor multiple

Demonstra'ii logice

- Tentativa de a raţiona despre întrebare
- Filtrarea răspunsului
- Generarea răspunsurilor ce par corecte, dar sunt imposibile.

• Analiza întrebării pentru modele (for patterns)

- Modele din întrebare sugerează modelele din răspuns

Resurse externe

• web (problematice)
- rezumate Web
- Validarea Răspunsului
- Mărirea datelor experimentale

• etichetoare PDV, extractoare de NE, separatoare de grupuri nominale
• Gazetteers, ontologii, tezaure
- WordNet, ConceptNet
• Demonstraţii logice
• Întrebări soluţionate anterior

V. EXTRAGEREA RĂSPUNSULUI

Extragerea Răspunsului (mai simplu)

• Extragera răspunsului care conţine tipul corect numit al entităţii dint enunţul superior.
• Extragerea răspunsurilor din enunţurile superioare N, şi votarea.
• Extragerea răspunsurilor candidate din enunţurile superioare N, validarea pe Web.

Marcarea/etichetarea răspunsului

• Trataţi etichetarea răspunsului drept marcarea unei entităţi numite.
• răspunsurile adesea nu sunt de tipul unei entităţi numite, de exemplu, celebrele Wh-întrebări.
• Răspunsurile nu sunt previzibile şi nu au întotdeauna indicatori previzibili.

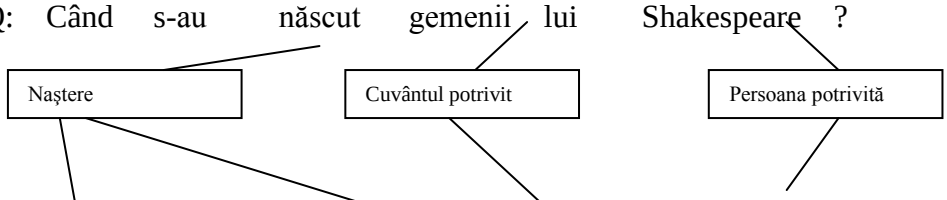
Caracteristicile răspunsurilor nu sunt direct secvenţiale şi sunt adesea prea lungi.

Caracteristicile unui tip de întrebări pot să nu se refere la alte tipuri de întrebări

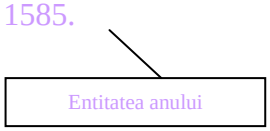
Etichetarea răspunsului (Easy)

• Determinati tipul de răspuns la întrebare
• Preluaţi un enunţ corect
• Întoarceţi entitatea numită corespunzătoare

Q: Când s-au născut gemenii lui Shakespeare ?



A: 2 ani mai târziu are loc naşterea gemenilor lui Shakespeare Judith şi Hamnet, fată şi băiat, botezaţi în anul 1585.



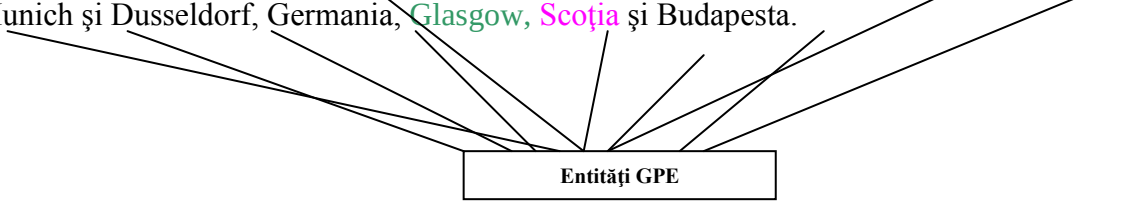
Etichetarea răspunsului (mai complicat)

• Determinati tipul de răspuns la întrebare
• Preluaţi un enunţ corect
• Întoarceţi entitatea numită corespunzătoare

Q: Unde se află Glasgow ?

A: Recesiunea a atins Glasgow mai târziu, precum şi restul Scoţiei.

A: Reducerile companiei America afectează toate zborurile din oraşele SUA spre Zurich, Elveţia, Munich şi Dusseldorf, Germania, Glasgow, Scoţia şi Budapesta.



Modele de răspuns

• Modelele de răspuns se pot afla în textul de vecinătate imediată a unui răspuns la o întrebare faptică
• În funcţie de tipul de întrebare
• Independent de o întrebare specifică

Exemple de modele de răspuns:

Modelul "inventator"

<NUME> , inventat de <RĂSPUNS>

<NUMELE>al cui <RĂSPUNS>

<NUME> a fost inventat de <RĂSPUNS>

<RĂPUNS> a inventat <NUME>

Exemplu:

Şi demonstraţia s-a întâmplat la 115 ani după ce Edison a inventat becul electric.

Modelul "anul de naştere"

Dodi Fayed s-a născut în 1956.

În viaţa sa scurtă şi moartea tragică Jessica Dubroff (1988 – 1996) a devenit o metaforă pentru orice, pornind de la idealismul timpuriu, până la excesele Neew-Age.

VI. CONCLUZII

Sistemele de extragere a răspunsurilor la întrebări în limbajul natural uman se înscriu în categoria sistemelor de achiziţie de informaţii.

Putem caracteriza drept cunoaştere starea unui sistem informaţional cuplat cu un ambient. Prin sistem informaţional se poate înţelege atât un individ (în sensul de persoană privită ca unitate distinctă faţă de alte persoane) care interacţionează în mod direct şi conştient cu ambientul şi care îşi construieşte prin mijloace proprii (observare, cogniţie, etc.) universul de cunoaştere, dar şi un sistem tehnic ce primeşte informaţia de la diverşi agenţi umani şi o procesează sau o pune, mai departe, la dispoziţia altor indivizi. Internetul poate fi descris ca un astfel de sistem informaţional care, interacţionând cu mediul prin intermediul agenţilor umani, îşi construieşte o bază proprie de cunoaştere.

REFERINTE BIBLIOGRAFICE

[1] http://en.wikipedia.org/wiki/Question_answering

[2] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.14.143&rep=rep1&type=pdf

[3] Dragomir R. Radev, John Prager, and Valerie Samn. Ranking potential answers to natural language questions. In Proceedings of the 6th Conference on Applied Natural Language Processing, Seattle, WA, May 2000.

[4] John Prager, Eric Brown, Anni Coden, and Dragomir Radev. Question-answering by predictive annotation. In Proceedings, 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, July 2000.

[5] Hirschman, L. & Gaizauskas, R. (2001) Natural Language Question Answering. The View from Here. Natural Language Engineering (2001), 7:4:275-300 Cambridge University Press.

[6] Lin, J. (2002). The Web as a Resource for Question Answering: Perspectives and Challenges. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002).

[7] Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S., Israel, D., Jacquemin, C., Lin, C-Y., Maiorano, S., Miller, G., Moldovan, D., Ogden, B., Prager, J., Riloff, E., Singhal, A., Shrihari, R., Strzalkowski, T., Voorhees, E., Weishedel, R. Issues, Tasks and Program Structures to Roadmap Research in Question Answering (QA).