Sistemul Întrebare-Răspuns Online

Constantin BREAHNA Victoria BOBICEV
Universitatea Technica a Moldovei
constantin.breahna@yahoo.com vika@rol.md

***Abstract* — lucrarea dată prezintă un sistem întrebare-răspuns online. Acest sistem răspunde utilizatorilor la întrebări din domeniul medicinei, ce țin de boli, simptome şi metode de tratament. Întrebările sunt formulate de către utilizatori în limba română, prin interfața web. Apoi, acestea sunt analizate de către sistem şi răspunsul este prezentat utilizatorului. Problema regăsirii automate şi formulării răspunsului corect este una din cele mai dificile în domeniul de cercetare numit procesarea limbajului natural. Deseori, analiza sintactică a întrebării nu este suficientă, deoarece este necesară înțelegerea mai detaliată a cerințelor utilizatorului. Din cauza dificultăților de acest gen, sistemele existente răspund doar la un număr limitat de tipuri de întrebări.**

***Index Terms* — procesarea limbajului natural, lingvistica computațională, regăsirea informației, sisteme întrebare-răspuns, extragerea informației.**

I. INTRODUCERE

În prezent oamenii au acces la o cantitate extrem de mare de informație şi necesită instrumente care ar permite selecția datelor potrivite din acest volum uriaş de cunoştințe. Lingvistica computațională încearcă să ofere sisteme informatice menite să ajute utilizatorul să se descurce în multitudinea documentelor disponibile.

Întrebare – Răspuns (ÎR) este un domeniu de cercetare care combină diferite domenii, şi anume: Căutarea Informației, Extragerea Informației şi Procesarea Limbajului Natural [2].

II. STRUCTURA GENERALĂ A SISTEMELOR ÎR

Un sistem ÎR conține, de regulă, 3 module: modulul de procesare a întrebării, modulul de căutare a informației şi modulul de extragere a răspunsului [1].

Modulul de procesare a întrebării o analizează sintactic şi o clasifică în baza unei taxonomii deja existente. Tipul întrebării la rândul său indică tipul răspunsului aşteptat. Prin urmare, modulul dat transformă întrebarea în interogare pentru un motor de căutare. Tipurile de intrebari posibile sunt exemplificate in Tabelul 1.

Clasificarea întrebării şi cunoaşterea tipului ei nu sunt suficiente pentru găsirea răspunsul la toate întrebările. Întrebarea ”ce?” este ambiguă din punctul de vedere al informației prezentate în întrebare. Pentru a aborda această ambiguitate, este necesară o componentă care identifică centrul. Centrul întrebării este un cuvânt sau set de cuvinte care indică ce informație este cerută în întrebarea dată. De exemplu, întrebarea ”*Care este cel mai lung râu din SUA?*” are ca centru ”*cel mai lung râu*”.

Odată ce au fost identificate ”centrul” şi ”tipul întrebării”, modulul formează o listă de cuvinte-cheie care vor fi trimise componentului de căutare a informației.

Pot fi utilizate şi metode de extindere a setului de cuvinte-cheie a întrebării, de exemplu, adresarea la sursele lexicale online, cum ar fi ontologia WordNet. Seturile de sinonime din WordNet pot fi folosite pentru a extinde seturile de cuvinte-cheie prin cuvintele legate semantic. Acestea pot apărea în documente conținând răspunsul corect la întrebare.

TABELUL 1. EXEMPLE TIPURI DE ÎNTREBĂRI

Model	Întrebare	Tipul răspunsului
cine X?	Cine este preşedintele României?	Persoană
unde X?	Unde se afla Mariana Stanciu pe 17 aprilie?	Loc
cînd X?	Când a fost adoptată constituția?	Data
cum X?	Cum au fost construite piramidele din Egipt?	Manieră
dece X?	De ce ninge?	Motiv

Modulul de căutare a informației primeşte la intrare cuvintele cheie formate de modulul precedent şi returnează un set de documente relevante.

Sistemele de căutare a informației sunt, de regulă, evaluate în baza următoarelor două valori – precizie (precision) şi retragere (recall). Precizia se referă la raportul dintre documentele relevante returnate din numărul total şi numărul total de documente obținute. Retragerea este numărul de documente relevante returnate din numărul total de documente disponibile în colecția de documente căutate. În general, scopul sistemelor de extragere a informației este de a optimiza atât precizia cât şi retragerea.

$$Precizia = \frac{numărul\ de\ răspunusuri\ corecte}{numărul\ de\ întrebări\ răspunse}$$

$$Retragerea = \frac{numărul\ de\ răspunusuri\ corecte}{numărul\ de\ întrebări\ care\ urmează\ a\ fi\ răspunse}$$

Pentru sistemele întrebare-răspuns, optimizarea diferă considerabil. Un sistem ÎR efectuează post-procesarea documentelor returnate, importanța retragerii sistemului este mult mai semnificativă decît precizia lui. În alte cuvinte, modulul dat trebuie sa returneze cît mai multe documente în scopul de a nu pierde vre-un document cu răspunsul corect la această etapă.

Prin urmare, numărul de documente returnate de sistemul de căutare a informației poate fi foarte mare. Documentele returnate sunt filtrate şi ordonate de următorul modul – cel de extragere a răspunsului. Obiectivul principal al modulului dat este de a crea un set de paragrafe candidate, care conțin răspunsul sau răspunsurile posibile. Filtrarea paragrafelor este folosită

pentru a reduce numărul de documente candidate, şi de a reduce cantitatea textului din fiecare document. Paragrafele sunt ordonate dupa următorul principiu: cu cât mai aproape se află în cadrul lui cuvintele cheie, cu atât el este plasat mai aproape de începutul listei. Dacă cuvintele cheie dintr-o întrebare sunt dispersate prin tot volumul textului din paragraf, acesta va fi plasat mai departe în lista ordonată [3].

În fază finală, modulul de procesare a răspunsului realizează identificarea, extragerea şi validarea răspunsului din setul de paragrafe ordonate transmise din modulul de procesare a documentelor.

Există diverse reguli de extragere a răspunsului corect din răspunsurile candidate. Extragerea poate fi bazată pe măsurarea distanței dintre cuvintele cheie, numărul de cuvinte-cheie potrivite sau alte reguli similare.

Astfel, sistemul ÎR efectuează următoarele operațiuni:

1. Utilizatorul postează întrebarea în sistemul ÎR.
2. Analizatorul de întrebare determină centrul întrebării pentru a spori eficacitatea sistemului ÎR.
3. Clasificarea întrebării joacă un rol important în sistemul ÎR prin identificarea tipului întrebării şi tipului răspunsului aşteptat.
4. În cadrul reformulării întrebării, aceasta este parafrazată şi extinsă.
5. Componenta căutării informației este folosită pentru extragerea documentelor relevante cuvintelor-cheie din întrebare.
6. Documentele obținute sunt filtrate şi împărțite în paragrafe ce pot conține răspunsul.
7. Paragrafele filtrate sunt ordonate şi trecute prin modulul de procesare a răspunsului.
8. Bazându-se pe tipul întrebării şi alte tehnici de recunoaştere, răspunsurile candidate sunt identificate.
9. Un set de reguli definite este utilizat pentru a extrage doar frazele relevante care răspund la întrebare.
10. Răspunsul extras este validat şi prezentat utilizatorului.

III. SISTEME DE ÎNTREBARE – RĂSPUNS CUNOSCUTE

Sisteme BASEBALL [7] şi LUNAR [6] au fost printre primele de acest fel. BASEBALL putea răspunde la întrebările legate de jocul baseball din SUA; sistemul LUNAR răspundea la întrebările cu privire la analiza geologică a rocilor returnate de misiunile lunare Apollo. Ambele sisteme au fost destul de eficiente în domeniul lor. LUNAR a fost demonstrat la conferința ştiințifică lunară în 1971 şi a fost capabil să răspundă la 90% din întrebările puse. În următorii ani au fost dezvoltate mai multe sisteme ÎR circumscrise. Trăsătura comuna a acestor sisteme era folosirea bazelor de date sau sistemelor de cunoştințe scrise de mână de către experți în domeniul ales. Abilitățile lingvistice ale sistemelor BASEBALL şi LUNAR erau asemănătoare celor folosite de ELIZA, primul program – robot pentru chat [8].

START este primul sistem întrebare-răspuns bazat pe internet. Acest sistem a apărut, pentru prima dată, în 1993 şi acum poate fi accesat pe adresa http://start.csail.mit.edu. Sistemul Start a fost dezvoltat în Laboratorul de inteligenta artificială MIT de către Boris Katz. În căutarea răspunsului la întrebare, sistemul foloseşte atât baza de date locală, cât şi un şir de resurse din internet. Spre deosebire de motoarele de căutare, sistemul START are scopul de a oferi utilizatorului doar informația corectă, în loc de a furniza o simplă listă cu rezultate găsite. În prezent, sistemul poate răspunde la milioane de întrebări în limba engleză bazate pe localități (orașe, țări, coordonate, etc.), filme (titluri, actori, regizori, etc.), definiții şi multe altele din următoarele categorii:

- Întrebări explicative (What is fractal?)
- Întrebări legate de fapte (Who invented telegraph?)
- Întrebări legate de relații (What country is bigger, Russia or USA?)
- Întrebări cu răspunsuri multiple (Show me some poems by Alexader Puskin?)

Sistemul START foloseşte ”adnotarea limbajului natural” pentru a conecta căutătorii de informație la sursele date. Această tehnică utilizează propoziții şi fraze în limbajul natural – adnotări – ca descrieri ale conținutului, care sunt asociate cu segmente de informație. Un segment de informație este preluat atunci când adnotarea lui corespunde întrebării. Această metodă permite sistemului START să manipuleze toate tipurile de date, cum ar fi: texte, diagrame, imagini, clipuri video sau audio, seturi de date, pagini web şi altele.

Componenta de procesare a limbajului natural este formată din două module care folosesc aceeaşi gramatică. Modulul de înțelegere analizează textul şi codifică informația găsită în baza cunoştințelor existente. Având segmentul potrivit al bazei de cunoştințe, modulul de generare creează propoziții.

Sistemul RACAI a fost dezvoltat de către Institutul de Cercetări pentru Inteligența Artificială a Academiei Române. Pentru prima data sistemul RACAI a fost prezentat în competiția de întrebare-răspuns CLEF în anul 2006.

Sistemul RACAI [4] poate răspunde la întrebările formulate în limba română căutând răspunsul în documentele Wikipedia. Sarcina sistemului este de a oferi şirul cel mai scurt şi bine format sintactic, care va răspunde la întrebarea utilizatorului. Caracteristicile specifice ale sistemului sunt:

1. Pentru varianta cross-lingvală, traducerea interogării este realizată pe baza ontologiei lexicale RO-WordNet aliniată la nivel de serie sinonimică cu ontologia pentru limba engleză WordNet.
2. Pentru indexare şi regăsire este folosită platforma open-source LUCENE. Precizia ei este superioară unui motor general de căutare ce se reflectă substanțial în performanța sistemului ÎR.
3. Clasificarea întrebărilor se realizează de un sistem bazat pe modelul maximizării entropiei, antrenabil în raport cu o tipologie dată. Sistemul de clasificare este extrem de precis (la CLEF2007, din cele 200 de întrebări, 199 au fost corect clasificate în una din cele 8 categorii definite de organizatori – temporal, time interval, definition, measure, list, location, names şi explanations).
4. Pentru prelucrările lingvistice (segmentarea lexicală, adnotarea morfo-sintactică, lematizarea, analiza sintactică parțială) este folosită platforma de servicii web

dezvoltată de către Institutul de Cercetări pentru Inteligența Artificială a Academiei Române.

5. Extragerea celui mai probabil răspuns la o întrebare se realizează cu ajutorul unui algoritm recursiv de potrivire structurală între arborele de dependențe al întrebării şi arborele de dependență al candidatului de răspuns (metoda este originală şi în mare măsură responsabilă pentru performanța foarte bună a sistemului).

IV. DESCRIEREA SISTEMULUI CREAT

Sistemul creat în cadrul lucrării date răspunde la înrebările din domeniul medicinei. Acestea țin de boli, simptome şi metode de tratament.

Sistemul este constituit din modulele clasice descrise anterior, şi anume: modulul de analiză a întrebării, modulul de căutare a informației şi modulul de extragere a răspunsului.

Întrebarea este inițial procesată, iar apoi analizată din două aspecte: determinarea tipului întrebării şi a centrului. Cele două atribute sunt folosite pentru a extrage răspunsul. Centrul întrebării va fi folosit pentru formarea interogării pentru motorul de căutare. Din documentele obținute de către motorul de căutare vor fi extrase paragrafe mici de text ca răspunsuri conform tipului întrebării, şi prezentate utilizatorului.

Primul pas este analiza morfologică. Sistemul creat foloseşte algoritmul elaborat de către Universitatea Alexandru Ioan Cuza din Iaşi (UAIC) - UAIC Romanian Part of Speech Tagger. Acest Part Of Speech tagger românesc combină un model statistic cu unul bazat pe reguli. Dicționarul morfologic a fost extras în mare parte din DexOnline şi conține 1,25 milioane de cuvinte distincte. POS tagger-ul are o precizie de 96.6% testat pe o varianta corectată a corpusului "1984". Un exemplu de întrebare analizată morfologic este prezentat în figura 1.

	Cum	**se**	**tratează**	**afazia**	**?**
	↓	↓	↓	↓	
leme	cum	sine	trata	afazia	
părți de vorbire	adverb	pron. reflexiv	verb predicativ	nume comun	quest

FIGURA 1. ANALIZA MORFOLOGICĂ A CUVINTELOR IN ÎNTREBARE.

Următorul pas este clasificarea întrebării, sau determinarea tipului întrebării. Acest pas joacă un rol important, fiindcă depistarea corectă a tipului întrebării va determina tipul răspunsului oferit de program.

Pentru rezolvarea acestei sarcini este folosit un tabel de cuvinte-cheie, ce pot descrie clasa sau tipul întrebării. Cuvintele-cheie din întrebare se compară cu cuvintele-cheie a claselor din tabel şi astfel se determină tipul întrebării.

Modul funcționării acestei metode este demonstrat în baza întrebării: *Cum se tratează afazia?* Se analizează fiecare cuvînt din propoziția dată:

- Cuvîntul ”cum” este găsit în lista cuvintelor cheie a clasei ***manner***, respectiv, clasei ***manner*** i se măreşte scorul.
- Cuvîntul ”se” nu este găsit printre cuvintele cheie şi este sărit.
- Cuvântul ”tratează” este găsit in cuvintele cheie a clasei ***manner***, respectiv, clasei ***manner*** i se măreşte scorul.

În final clasa ***manner*** obține scorul cel mai mare şi este considerată clasa întrebării date. Tabelul 2 conține cîteva exemple de clase şi cuvinte-cheie asociate claselor date.

TABEL 2. CLASELE ÎNTREBĂRILOR ŞI CUVINTELE CHEIE ASOCIATE CU CLASELE DATE

Clasa întrebării	**Cuvintele cheie**
Definiton	ce, definiția, explică
Cause	care, cauzele, cauza, pricina
Manner	cum, tratează, tratament, tratarea

Următoarea etapă în căutarea răspunsului este definirea centrului întrebării. Acesta este definit de setul de cuvinte extrase din întrebare. Pentru determinarea centrul întrebării şi formarea interogării sunt parcurşi următorii paşi:

- excluderea semnelor de punctuație;
- excluderea cuvintelor interogative;
- excluderea cuvintelor funcționale;
- termenii rămaşi formează centrul;
- crearea interogării.

După determinarea centrului întrebării poate fi formulată interogarea.

Inițial sistemul caută răspunsul în baza de date a răspunsurilor stocate. Aceasta conține următoarele cîmpuri: titlul documentului; descrierea documentului; cuvintele-cheie ale documentului; sursa documentului; data adăugării/modificării; tipul întrebării; cuvintele-cheie care definesc tipul întrebării; întrebarea trimisă sistemului ÎR de către utilizator; numărul de adresări cu întrebarea dată. La momentul dat baza de date se completează manual, însă se preconizează crearea unui modul care va completa baza de date automat cu întrebările utilizatorilor şi răspunsurile corecte în procesul utilizării sistemului online.

Dacă în baza de date nu a fost găsit un răspuns potrivit, sistemul se adresează online la Wikipedia, o sursă vastă de informații din diverse domenii. Wikipedia oferă un set de API în diferite limbaje de programare şi formate de date.

Sistemul este creat cu ajutorul limbajului PHP. Datele din Wikipedia sunt extrase în formatul SOAP care se bazează pe limbajul XML. Documentul este analizat paragraf dupa paragraf, căutînd cuvintele-cheie şi indicatoarele tipului de întrebare. Paragraful cu cea mai înaltă rată de corespondență este selectat şi returnat utilizatorului.

V. CONCLUZIE

Lucrarea dată prezintă un sistem de întrebare-răspuns online. Sistemul dat răspunde la întrebările utilizatiorilor din domeniul medicinei, şi anume despre bolile, simptomele şi metodele de tratare. Întrebările se formulează în limba română de către utilizatori prin interfața web. Sistemul analizează întrebările puse, caută răspunsul şi oferă utilizatorului.

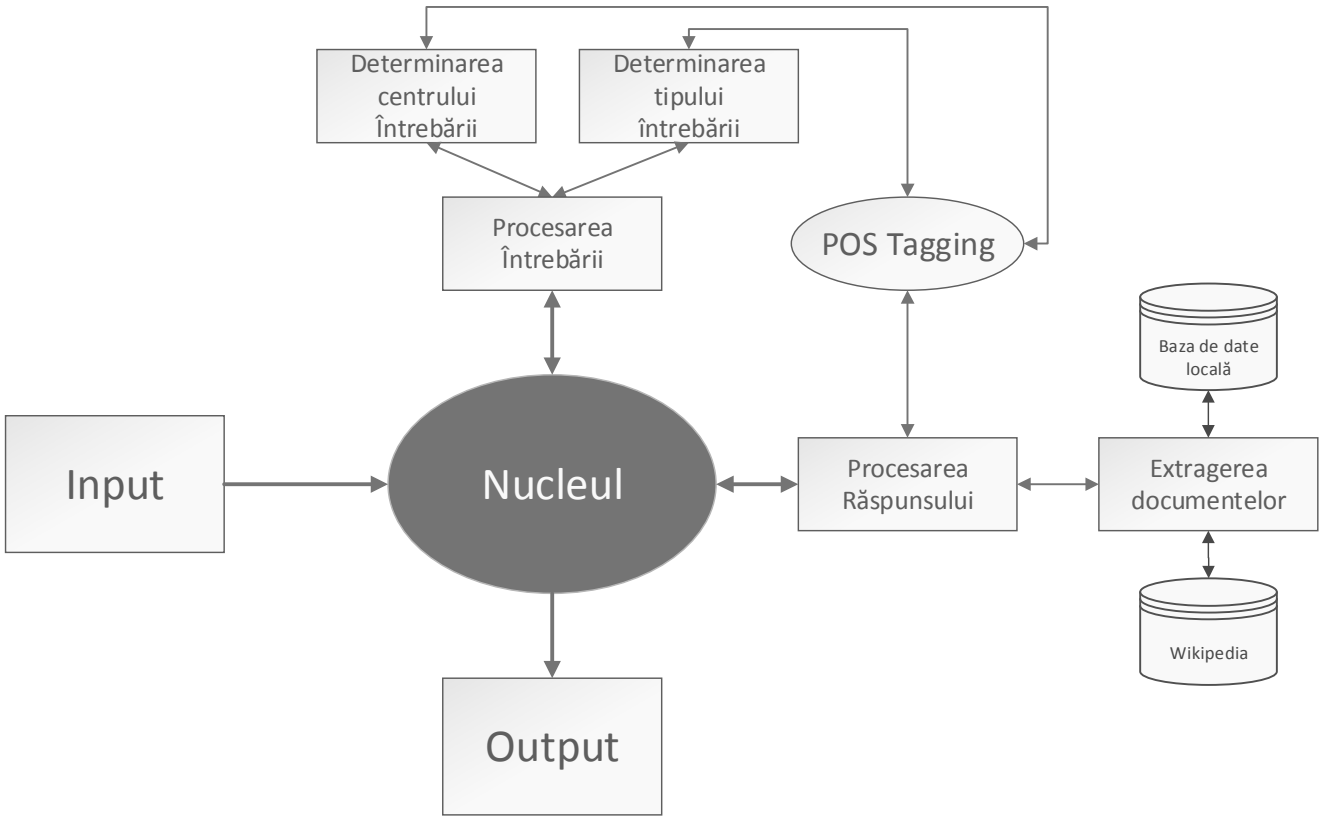

FIGURA 2. SCHEMA GENERALĂ A SISTEMULUI CREAT

Problema regăsirii automate şi formulării răspunsului corect este una din cele mai dificile în domeniul de cercetare numit procesarea limbajului natural. În multe cazuri, analiza sintactică a întrebării nu este suficientă. Astfel, este necesară înțelegerea tipului de răspuns solicitat de utilizator. Din cauza dificultăților de acest gen, sistemele existente răspund doar la un număr limitat de tipuri de întrebări

La momentul dat sistemul creat răspunde la întrebările de trei tipuri: definiția, simptomele şi tratamentul posibil a bolilor descrise în Wikipedia în limba română.

BIBLIOGRAFIA

[1] Constantin Orăsan, Doina Tatar, Gabriela Şerban, Dana Lupsa, Andrian Oneț. How to build a QA system in your back-garden: application for Romanian

[2] Dan Tufiș. Sisteme de întrebare – răspuns în limbaj natural petntru spații de căutare deschise - Seminarul internațional ”Instrumente pentru asistarea traducerii”.

[3] Radu Ion, Alexandru Ceauşu, Dan Ştefănescu, Dan Tufiş, Elena Irimia, Verginica Barbu Mititelu. Monolingual and Multilingual Question Answering on European Legislation – Research Institute for Artificial Intelligence, Romanian Academy, no.13, Romania.

[4] Radu Ion, Dan Ştefănescu, Alexandru, Ceauşu, Dan Tufiş. RACAI’S QA System at the Romanian-Romanian QA @ CLEF2008 Main Task - Research Institute for Artificial Intelligence, Romanian Academy, Septembrie, no.13, Romania.

[5] Boris Katz, Gary Borchardt and Sue Felshin. Natural Language Annotations for Question Answering. Proceedings of the 19th International FLAIRS Conference (FLAIRS 2006), 2006.

[6] Woods, W. A., Kaplan, R. M., and Nash-Webber, B. The lunar sciences natural language system: final report. Rep. No. 2378, Bolt Beranek and Newman Inc., Cambridge, Mass., 1972.

[7] Bert F.Green, Jr., Alice K.Wolf, Carol Chomsky, and Kenneth Laughery. BASEBALL: an Automatic Question-Answerer. Western joint IRE-AIEE-ACM computer conference, 1961.

[8] Weizenbaum, Joseph (January 1966), "ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine", Communications of the ACM 9 (1), 1966.