Dan Ştefănescu, Radu Ion, Alexandru Ceauşu, and Dan Tufiş. Sistem întrebare-răspuns antrenabil pentru limba română.
In Adrian Iftene, Horia-Nicolai Teodorescu, Dan Cristea, and Dan Tufiş (eds.) Resurse Lingvistice şi Instrumente pentru Prelucrarea Limbii Române, pp. 153-164, Iaşi, Romania, septembrie 2010. Universitatea "Al.I. Cuza" Iaşi, Editura Universităţii "Al.I. Cuza" Iaşi. ISSN 1843-911X.

SISTEM ÎNTREBARE-RĂSPUNS ANTRENABIL PENTRU LIMBA ROMÂNĂ

DAN ŞTEFĂNESCU, RADU ION, ALEXANDRU CEAUŞU, DAN TUFIŞ

Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română

{danstef, radu, aceausu, tufis}@racai.ro

Rezumat

Lucrarea prezintă un sistem întrebare-răspuns dezvoltat la *Institutul de Cercetări pentru Inteligenţă Artificială* – ICIA în cadrul unui proiect naţional şi evaluat independent în contextul competiţiei europene CLEF. Evaluarea a fost realizată în cadrul exerciţiului *ResPubliQA* pentru limba română. Este descris modul de combinare a diferiţilor factori de relevanţă pe baza cărora sistemul identifică paragrafele cele mai relevante ca răspunsuri la întrebările formulate în limbaj natural. Sistemul este disponibil on-line pe pagina de servicii web a ICIA. El este însă complet antrenabil, funcţionalitatea sa fiind independentă de registrul lingvistic ce caracterizează datele de antrenare.

1. Introducere

Cercetările privind prelucrarea automată a limbajului natural (PLN), domeniu central al inteligenţei artificiale, produc rezultate cu impact din ce în ce mai mare în societatea globalizată de fenomenul Internet. Sistemele de întrebare-răspuns (ÎR) în limbaj natural, în vogă în perioada anilor ‘70-80, au revenit în centrul cercetărilor PLN dar, de data aceasta având ca obiectiv identificarea răspunsurilor la întrebări arbitrare în spaţii de căutare incomparabil mai mari, la limită întregul web. Conţinutul informaţional al acestui spaţiu virtual este atât de mare încât se consideră că orice solicitare raţională de informaţie îşi poate găsi măcar un răspuns pe Internet. Evaluarea calităţii răspunsurilor şi respectiv asigurarea găsirii lor sunt însă probleme de cercetare, pentru care abordările tradiţionale au devenit insuficiente. Astfel, implementările motoarelor de căutare moderne recurg din ce în ce mai mult la tehnici PLN, acestea fiind utilizate în toate etapele fluxului de prelucrare, începând de la nivelul specificării întrebării şi până la extragerea fragmentului de text relevant dintr-unul sau mai multe documente. Odată cu utilizarea tehnicilor PLN au apărut şi campaniile de evaluare în domeniul regăsirii inteligente a informaţiei. Acestea constituie astăzi priorităţi ale cercetării de avangardă dedicată spaţiului digital al cunoaşterii. Ele au fost organizate mai întâi în SUA (MUC-*Message Understanding Conference*, TREC-*Text Retrieval Conference*, DUC-*Document Understanding Conference*, devenită TAC-*Text Analysis Conference*). În Europa, manifestarea similară este CLEF (Cross Language Evaluation Forum), ajunsă în anul 2009 la a 10-a ediţie. Având ca subiect al analizei în primul rând limbile Uniunii Europene, începând cu anul 2006 limbile europene “cu resurse electronice limitate” (română, bulgară, cehă, greacă, portugheză etc.) au devenit “subiecte” de concurs.

În acest context, proiectul naţional SIR-RESDEC (lansat în 2007) a răspuns unei priorităţi europene, propunându-şi realizarea unui sistem de ÎR în limbaj natural la nivelul celor mai avansate sisteme ale cercetării internaţionale. Consorţiul SIR-RESDEC format din cercetători de la ICIA, UAIC şi ICI şi-a concentrat eforturile în

direcţia realizării unor sisteme de ÎR pentru limbile română şi engleză, în domeniul legislaţiei Uniunii Europene şi respectiv în domeniul geneticii umane. Grupul de cercetare al ICIA şi-a concentrat eforturile în direcţia realizării sistemului de ÎR pentru limba română în domeniul legislativ, având la dispoziţie corpusul JRC-Acquis (Steinberger et al., 2006). În restul articolului de faţă vom descrie acest sistem.

2. *Considerente preliminare*

Pentru a testa performanţele sistemului, ICIA s-a înscris în anul 2009 în competiţia QA@CLEF la secţiunea *ResPubliQA*[1] (Peñas et al., 2009), urmând tradiţia participărilor la competiţiile CLEF încă din 2006. Sarcina sistemelor dezvoltate de echipele înscrise la *ResPubliQA* a fost să identifice automat paragrafe relevante pentru răspunsuri la întrebări formulate în limbaj natural, în domeniul juridic acoperit de corpusul paralel de lucru al competiţiei. Pentru prima oară, evaluările sistemelor de întrebare-răspuns în limbaj natural au putut fi comparate interlingual, întrucât întrebările de test (500) au fost aceleaşi în 8 limbi (bască, bulgară, engleză, franceză, germană, italiană, română şi spaniolă) răspunsurile trebuind a fi căutate în corpusul paralel (JRC-Acquis) al legislaţiei europene „Acquis Communautaire" disponibil în toate limbile Uniunii Europene. Alinierea la nivel de paragraf a corpusului pentru toate limbile implicate a oferit posibilitatea evaluării răspunsurilor sistemelor indiferent de limba de interogare. În plus, tot în premieră, organizatorii ResPubliQA au calculat, pentru fiecare limbă, performanţele unui sistem de regăsire documentară (RD) de ultimă generaţie, dar fără componente de PLN. S-a urmărit în acest mod evaluarea cantitativă a rolului tehnologiilor de prelucrare a limbajului natural faţă de tehnicile standard utilizate în regăsirea informaţiilor. Notând cu $A_{ÎR}$ acurateţea sistemului de ÎR (v. Secţiunea 6) şi cu A_{RD} acurateţea sistemului de RD, atunci raportul $M = \frac{A_{ÎR}}{A_{RD}}$ cuantifică meritul tehnicilor de prelucrare a limbajului natural. O cifră de merit supraunitară semnifică faptul că prelucrarea limbajului natural îmbunătăţeşte performanţa unui sistem de regăsire documentară. Deşi pare intuitiv ca pentru orice sistem de ÎR cifra de merit M să fie supraunitară, organizatorii ResPubliQA au constatat că dintre 28 de sisteme evaluate doar 14 au avut o cifră de merit supraunitară (v. Table 10 în (Peñas et al., 2009)).

Cu experienţa dobândită în competiţiile CLEF precedente (Puşcaşu et al., 2007; Tufiş et al., 2008c; Ion et al., 2009a) şi cerinţele specifice ale competiţiei din 2009, obiectivele tehnice principale au fost perfecţionarea modulului de regăsire a paragrafelor relevante şi respectiv a celui de reordonare a acestora, pe baza unei analize complexe a relevanţei paragrafelor candidat. Un impact semnificativ l-a constituit implementarea unei metode similare metodei de optimizare MERT (Och, 2003), în cadrul etapei de reordonare a paragrafelor. Am păstrat abordarea din anii precedenţi în ceea ce priveşte construcţia sistemului, însă diferitele module care îl alcătuiesc au fost implementate ca servicii şi/sau aplicaţii Web (Tufiş et al., 2008b):

- *serviciul de analiză a întrebării*[2] cu ajutorul căruia se clasifică fiecare întrebare ataşându-i-se o etichetă ce indică tipul de răspuns pe care acea întrebare îl cere;

[1] http://celct.isti.cnr.it/ResPubliQA/

[2] http://shadow.racai.ro/JRCACQCWebService/Service.asmx?WSDL

- *serviciile de generare a cererilor*[3] cu ajutorul căruia o întrebare în limbaj natural este transformată în interogări în limbaj formal compatibile cu motorul de căutare;
- *serviciul de interogare a motorului de căutare*[4] se ocupă de partea de regăsire a paragrafelor relevante pentru o interogare în limbaj formal furnizată la intrare;
- *modulul de reordonare a paragrafelor* preia rezultatele furnizate de motorul de căutare sub forma unei liste de paragrafe, calculează scoruri suplimentare de relevanţă pentru fiecare paragraf şi, în funcţie de o interpolare liniară (obţinută aplicând o optimizare de tip MERT) a acestora, asignează scoruri paragrafelor. Paragraful sau paragrafele cu scorurile cele mai ridicate sunt întoarse utilizatorilor.

În faza de indexare a corpusului JRC-Acquis, am luat în considerare doar textul propriu-zis al documentelor, acesta fiind în prealabil preprocesat cu ajutorul TTL (Ion, 2007) dezvoltat la ICIA. Textul a fost segmentat la nivel de unitate lexicală în funcţie de terminologia Eurovoc, adnotat la parte de vorbire, şi lematizat.

3. Identificarea terminologiei

Având în vedere caracterul juridic specializat al corpusului de lucru, o etapă importantă a fost identificarea şi tratarea ca unităţi lexicale a unor anumite expresii sau termeni multi-cuvânt. Acest lucru a fost realizat cu ajutorul tezaurului multilingv Eurovoc[5] (Ştefănescu and Tufiş, 2006). Descriptorii Eurovoc sunt termeni tehnici care trebuie să apară cu consecvenţă în toate documentele juridice pentru toate limbile implicate. Recunoaşterea acestora atât în faza de analiză a întrebărilor cât şi faza de preprocesare a documentelor (ce trebuie ulterior indexate) devine esenţială pentru performanţele oricărui sistem QA pe acest corpus.

În consecinţă am realizat un modul care, după preprocesarea corpusului, recunoaşte termenii Eurovoc şi generează unităţi lexicale corespunzătoare (în acest fel, unui termen multi-cuvânt îi corespunde o singură unitate lexicală). Etapa de identificare a terminologiei se desfăşoară de-a lungul a 6 etape: (i) termenii din Eurovoc sunt identificaţi în corpus, în forma lor de tezaur; (ii) pentru fiecare formă ocurenţă identificată la pasul 1 se extrage secvenţa de leme implicată şi se adaugă unui inventar; (iii) se identifică în corpus toate ocurenţele secvenţelor de leme din inventarul construit la pasul 2; (iv) termenii astfel identificaţi sunt aduşi la forma ocurenţă din corpus; (v) fiecărui termen identificat i se asignează un descriptor morfo-sintactic – cum termenii sunt de fapt grupuri nominale, descriptorul asignat termenului este acelaşi cu descriptorul centrului grupului nominal; (vi) fiecărei ocurenţe a unui termen i se asignează ca lemă descriptorul corespunzător Eurovoc.

De exemplu, termenul *adunare parlamentară* apare în corpus cu formele flexionate: *adunarea parlamentară*, *adunările parlamentare*, *adunărilor parlamentare*. Toate aceste unităţi lexicale primesc ca lemă descriptorul Eurovoc *adunare parlamentară*, iar ca descriptor morfosintactic, descriptorul corespunzător centrului grupului (i.e., *adunare*, *adunarea*, *adunările*, *adunărilor*).

[3] http://shadow.racai.ro/QADWebService/Service.asmx?WSDL

[4] http://www.racai.ro/webservices/search.asmx?WSDL

[5] http://en.wikipedia.org/wiki/Eurovoc

4. Clasificarea automată a paragrafelor şi întrebărilor

Specificaţiile competiţiei au definit 5 tipuri de întrebări posibile: (i) *factoid* (factual) – întrebări care cer ca răspuns persoane, locaţii, instituţii, momente în timp, etc.; (ii) *definition* (definiţie) – întrebări care cer ca răspuns o definiţie; (iii) *procedure* (procedură) – întrebări care cer ca răspuns o procedură juridică; (iv) *reason* (motiv) – întrebări care cer ca răspuns un motiv, o cauză; (v) *purpose* (scop) – întrebări care cer ca răspuns un scop, un obiectiv. Numărul redus de clase şi faptul că răspunsul corect pentru o întrebare nu se poate întinde pe mai multe paragrafe au condus la ideea clasificării paragrafelor în funcţie de probabilitatea lor de a răspunde la un tip de întrebare sau altul. Etichetarea unui paragraf cu tipul de întrebare la care acel paragraf ar putea răspunde cel mai bine, oferă, în mod evident, posibilitatea reducerii complexităţii etapei de identificare a documentelor relevante, pe care o vom numi etapă de *regăsire documentară*. Acolo unde tipul întrebării este corect identificat, răspunsul este căutat, în principal, în paragrafele de tip identic cu cel al întrebării. Acest lucru a presupus construcţia unui modul de clasificare a întrebărilor care a fost optimizat şi antrenat pentru tipurile de întrebări asociate corpusului de tip juridic aşa cum vom arata mai jos.

Problema de clasificare a paragrafelor este similară cu cea a selecţiei propoziţiilor dintr-un text în vederea generării automate a rezumatului acelui text (Ion et al., 2009b). Se observă însă că cele două probleme diferă în ceea ce priveşte numărul de clase care trebuiesc considerate şi tipul entităţilor (pe de o parte paragrafe, iar de cealaltă, fraze) ce trebuie supuse procesului de clasificare. În cazul nostru, clasele pe care le-am considerat diferă uşor de cele furnizate de organizatorii competiţiei în specificaţiile date. Astfel, clasele *reason* şi *purpose* au fost unite, clasa obţinută având denumirea de *reason-purpose*. Motivul constă în dificultatea dezambiguizării automate între cele două clase. Pentru a îmbunătăţi precizia în faza de regăsire a paragrafelor relevante, am adăugat o altă clasă, etichetată *delete*, cu scopul de a elimina astfel, încă din faza de căutare, acele paragrafe care nu ar fi putut conţine răspunsuri corecte pentru vreo întrebare. În această categorie intră paragrafele ce conţin titluri (e.g., „Articolul 1”), părţi de tabele ale căror formatare nu s-a mai păstrat, semnături, etc.

Pentru faza de antrenare a clasificatorului am utilizat o colecţie de aproximativ 800 de paragrafe etichetate manual cu următoarele etichete: *factoid*, *definition*, *procedure*, *reason-purpose* şi *delete*, iar pentru a testa precizia, am folosit doar 89 de paragrafe. Metoda de clasificare la care am recurs este aceeaşi pe care am folosit-o şi pentru clasificarea întrebărilor în anii precedenţi: *principiul maximizării entropiei* (Ratnaparkhi, 1998). Trăsăturile pe care le-am luat în considerare s-au bazat pe cuvinte cheie, descriptori morfo-sintactici, punctuaţie, lungimea propoziţiei. Primele 5 cuvinte au fost considerate trăsături, ele având un puternic caracter de discriminare pentru clasele alese. Ca trăsături de tip morfo-sintactic, am luat în considerare descriptorul morfo-sintactic al verbului principal, o altă trăsătură fiind existenţa sau absenţa în cadrul propoziţiei a unui substantiv propriu. Alte trăsături sunt numărul de virgule din propoziţie, numărul de ghilimele (indicând numărul de citate), semnul de punctuaţie cu care se termină propoziţia. Din punct de vedere ortografic, o trăsătură importantă este numărul cuvintelor din propoziţie care încep cu majusculă. Trăsăturile legate de lungime includ numărul propoziţiilor în paragraf şi lungimea paragrafului în cuvinte.

Precizia clasificatorului pe datele de test a fost de 94% (Ion et al., 2009b; Ştefănescu, 2010). Deşi gradul de încredere statistică a evaluării preciziei este redus datorită numărului mic de 89 de paragrafe conţinute de datele de test, rezultatele finale obţinute folosind clasele asignate paragrafelor au fost mult îmbunătăţite. În condiţiile în care doar primele 50 de paragrafe întoarse de motorul de căutare au fost luate în considerare în etapele următoare ale fluxului de prelucrare, am constatat că, folosind clasificarea paragrafelor, în majoritatea cazurilor răspunsurile corecte s-au aflat în printre acestea.

Pentru clasificarea întrebărilor cele mai multe din sistemele actuale QA folosesc un modul specializat pentru a determina ce tip de răspuns ar trebui căutat în corpusurile avute la dispoziţie. Desigur, clasificarea se poate realiza în mai multe feluri, cele mai simple metode folosind reguli sub forma unor expresii regulate. În cazul nostru, am apelat din nou la clasificatorul bazat pe maximizarea entropiei. Acesta a fost antrenat pentru a identifica 8 clase: *reason-purpose*, *procedure*, *definition*, *location*, *name*, *numeric*, *temporal* şi *factoid*. Clasa *location* cuprinde întrebările care necesită ca răspuns o locaţie; clasa *name* conţine întrebările care cer ca răspuns un nume de persoană, organizaţie sau numele unei entităţi (e.g., comisie, ţară, etc.); clasa *temporal* cuprinde acele întrebări care cer drept răspuns o dată calendaristică sau un interval de timp; clasa *numeric* conţine întrebările care cer drept răspuns un număr („*Câţi membri sunt în comisia de ...*”), iar clasa *factoid* conţine toate întrebările de tip factoid care nu sunt în clasele tocmai descrise. Trăsăturile pe care le-am folosit sunt următoarele: primul cuvânt de tip *WH* din întrebare, primul verb principal din întrebare, primul substantiv din întrebare, descriptorii morfo-sintactici ai tuturor substantivelor, verbelor, adjectivelor, adverbelor şi numeralelor din întrebare, ordinea de apariţie a primului verb şi a primului substantiv în cadrul întrebării analizate. Desigur, extragerea acestor trăsături survine abia după preprocesarea întrebării.

Pentru antrenare, am plecat de la exemplele furnizate de organizatori şi am construit 200 de întrebări (incluzând şi acele exemple) pentru care am ataşat manual clasa din care fac parte (Ştefănescu, 2010). Modelul a fost construit considerând doar trăsăturile care apăreau de cel puţin două ori în datele de antrenare, în final acesta având o acurateţe de 99% pe aceste date. Deşi evaluarea performanţelor folosind datele de antrenare („*biased evaluation*”) trebuie evitată pentru aplicaţiile de învăţarea automată, ea ne-a permis totuşi să ajungem la o selecţie mai bună de trăsături. După cum era de aşteptat, acurateţea a scăzut pentru cele 500 de întrebări furnizate odată cu startul competiţiei, însă la numai 97.2%. Scorul nu este atât de surprinzător pe cât pare, datorită diferenţelor mari dintre trăsăturile caracteristice claselor considerate. De altfel, am ales clasele astfel încât clasificatorul să nu aibă probleme în identificarea lor corectă. Ne interesează în principal să nu avem erori de clasificare, chiar dacă rămânem cu clase nu foarte rafinate. Cu alte cuvinte, încercăm să restrângem cât mai mult spaţiul de căutare fără însă a îngădui prea multe erori de clasificare (Ştefănescu, 2010).

5. Sistemul QA

Sistemul nostru este implementat ca un flux de prelucrare realizat peste arhitectura serviciilor şi aplicaţiilor web dezvoltate la ICIA. Sistemul este optimizat pentru a maximiza un scor de relevanţă global $S(p)$ folosit la identificarea paragrafului cel mai plauzibil a constitui răspunsul adecvat la o întrebare adresată sistemului. Scorul global

$S(p)$ se calculează ca o combinaţie liniară a unor scoruri asignate paragrafelor în funcţie de criteriile de relevanţă în raport cu o întrebare furnizată de un utilizator:

$$S(p) = \sum_i \lambda_i s_i, \quad \sum_i \lambda_i = 1 \qquad (1)$$

unde s_i ($s_i \in [0,1]$) este unul din următoarele scoruri de relevanţă:

- s_1 este 1 când clasificarea întrebării corespunde cu cea a paragrafului, în caz contrar, valoarea sa fiind 0. După cum am arătat, clasele modulului de clasificare a paragrafelor (5 la număr) diferă de cele ale modulului de clasificare a întrebărilor (8) dar între ele există o corespondenţă bine definită (Ion et al., 2009b);

- s_2 este un scor de similaritate lexicală bazat pe lanţuri lexicale, coeziune lexicală şi identificarea perechii verb principal–argument; este scorul care încearcă să cuantifice gradul de similaritate lexicală dintre o întrebare şi paragrafele întoarse de motorul de căutare ca fiind relevante pentru acea întrebare. Se calculează după următoarea metodă: pentru o întrebare Q şi un paragraf candidat P, construim două liste conţinând lemele ce corespund cuvintelor conţinut din întrebare, respectiv din paragraf: LQ şi LP. Numim lemele din lista LQ *cuvinte cheie*. Pe baza lor asignăm lui P un scor de relevanţă. Astfel, s_2 se calculează ca un produs a trei alte scoruri ce caracterizează: (i) distanţa semantica dintre lemele celor două liste (*DS*), (ii) coeziunea cuvintelor cheie în paragraf (*CC*) şi (iii) identificarea în paragraf a unui posibil cuplu verb-argument extras din întrebare (*VA*). Aceste scoruri au fost descrise pe larg în (Ştefănescu, 2010) şi nu vom mai insista asupra lor. Scorul final se calculează ca produsul celor trei scoruri:

$$s_2 = DS \times CC \times VA$$

- s_3 este un scor similar cu scorul BLEU (Papineni et al., 2002) care avantajează paragrafele în care cuvintele cheie ale întrebării apar în aceeaşi ordine ca în întrebare; ca şi scorul precedent, se calculează considerând lemele cuvintelor conţinut din întrebare, respectiv paragraf. Ideea implementării acestui scor (Ion et al., 2009b) are la bază observaţia că de foarte multe ori, în realitate, formularea răspunsului corespunzător unei întrebări conţine o parte din acea întrebare. Principiul comparării n-gramelor a fost folosit în cazul de faţă pentru a evalua similaritatea dintre întrebare şi paragrafele candidat.

- s_4 şi s_5 sunt scorurile de relevanţă pentru paragraf şi document întoarse de motorul de căutare.

5.1. Prelucrarea întrebării şi alegerea răspunsului

După ce sistemul primeşte la intrare o întrebare, ea este trimisă serviciului web TTL pentru a fi preprocesată. Este apelat apoi serviciul web care se ocupă de clasificarea întrebărilor pentru a obţine tipul de răspuns care trebuie căutat. În următorul pas, folosind informaţia adnotată după preprocesare, întrebarea este transformată în interogări într-un limbaj formal înţeles de motorul de căutare. Folosim 2 algoritmi pentru a genera două interogări diferite, ambele conţinând ca termen de căutare clasa întrebării. Trebuie menţionat că în faza de indexare au fost indexate odată cu paragrafele şi clasele corespunzătoare acestora, pentru ca reducerea spaţiului de căutare să se facă direct din faza de regăsire documentară.

Pentru fiecare din cele două interogări generate, motorul de căutare întoarce două liste L_1 şi L_2 conţinând fiecare 50 de paragrafe sortate după scorul de relevanţă descris de ecuaţia (1). Răspunsul întors de sistem este acel paragraf care se găseşte atât în L_1 cât şi în L_2 şi care satisface:

$$\text{argmin}_p(\text{rang}_1(p) + \text{rang}_2(p)), \quad \text{rang}_1(p) \leq K, \text{rang}_2(p) \leq K, K \leq 50 \qquad (2)$$

unde $rang_1(p)$ este poziţia paragrafului p în L_1, iar $rang_2(p)$, poziţia paragrafului p în L_2. Dacă nu există un paragraf care să satisfacă condiţiile din (2), atunci sistemul semnalizează că nu poate găsi un răspuns la întrebarea dată, întorcând şirul de caractere NOA (*no answer* – nu există răspuns). În urma experimentelor, cele mai bune rezultate au fost obţinute pentru K=3.

Sistemul dezvoltat este antrenabil, ponderile λ_i folosite în interpolarea liniară a scorurilor de relevanţă fiind obţinute cu ajutorul unei tehnici de optimizare similare cu metoda MERT. Metoda de antrenare pentru găsirea celui mai probabil răspuns a fost folosită pe cele 200 de întrebări utilizate şi în cazul clasificatorului pentru întrebări şi are următoarele etape: (i) rularea sistemului pentru cele 200 de întrebări şi păstrarea primelor 50 de paragrafe întoarse de motorul de căutare pentru fiecare întrebare, ordinea paragrafelor fiind dată DOAR de scorurile de paragraf întoarse de motor (s_4); (ii) calcularea scorurilor s_i, $i = \overline{1,5}$, pentru fiecare din paragrafele extrase; (iii) pentru fiecare combinaţie de ponderi λ, cu $\sum_{i=1}^{5} \lambda_i = 1$ şi un pas de incrementare de 10^{-2}, se calculează scorul *Mean Reciprocal Rank* (MRR) (Radev et al., 2002) pentru întreg setul de 200 de întrebări, lista paragrafelor întoarse fiind sortată după ecuaţia (1); (iv) reţinerea combinaţiei de ponderi λ care maximizează scorul MRR.

Cele două moduri de a genera cereri în limbaj formal conduc la rezultate diferite ale motorului de căutare. Putem considera că avem de a face cu două sisteme ale căror rezultate sunt combinate pentru a obţine răspunsul final. Astfel, cele două sisteme au fost individual optimizate, în ceea ce priveşte ponderile λ, fără a avea în vedere, în vreun moment, posibilitatea ca un sistem să întoarcă răspunsuri NOA. După antrenarea celor două sisteme, observând că sistemul de evaluare al competiţiei *ResPubliQA* favorizează răspunsurile NOA faţă de cele incorecte, am luat decizia de a introduce ecuaţia (2) pentru a păstra doar acele răspunsuri pentru care avem un grad de încredere ridicat. La momentul transmiterii rezultatelor, setul de ponderi λ utilizat nu ţinea cont de tipul întrebărilor furnizate la intrare. Vom arăta în secţiunea de evaluare că putem obţine rezultate îmbunătăţite antrenând setul de ponderi pentru fiecare clasă de întrebări.

5.2. Generarea de cereri formale pentru motorul de regăsire a documentelor

Etapa de regăsire documentară este similară cu cea descrisă în (Ion et al., 2009a), folosind şi de această dată motorul de căutare Lucene (Hatcher and Gospodnetić, 2004). Cuvintele documentelor au fost filtrate în funcţie de descrierea lor morfo-sintactică astfel încât să rămână pentru indexare doar cuvintele conţinut. Mai mult, cuvintele au fost normalizate la forma lor lemă. Eventualele erori de dezambiguizare morfo-lexicală sau lematizare au fost luate în calcul şi pentru a compensa astfel de erori (mai ales în cazul unităţilor lexicale absente din dicţionarul sistemului TTL) am indexat şi forma ocurenţă pe aceeaşi poziţie cu lema. În acest fel, un termen poate fi căutat atât ca lemă, cât şi în forma sa ocurenţă. Am procedat la construirea a doi indecşi pentru colecţia de

documente: un index pentru paragrafe şi unul pentru documente. Dată fiind o anumită cerere, motorul întoarce o listă alcătuită din paragrafele cele mai relevante în raport cu acea cerere. Astfel, avem deja două scoruri calculate pentru un paragraf: scorul de relevanţă pentru paragraf în indexul construit pe paragrafe (s_4) şi cel de relevanţă pentru documentul din care face parte paragraful în indexul construit pe documente (s_5).

Abilitatea sistemului nostru de a întoarce pentru unele întrebări mesajul că nici un răspuns nu a fost găsit se datorează combinării rezultatelor diferite obţinute folosind cei doi algoritmi de generare a interogărilor formale către motorul de căutare.

ALGORITMUL DE GENERARE A CERERILOR FORMALE – VARIANTA 1 ia în considerare toate cuvintele conţinut ale întrebării (substantive, verbe, adjective şi adverbe) cu care construieşte o disjuncţie de termeni, ce sunt practic lemele acelor cuvinte. Singura condiţie impusă asupra includerii unui termen în cerere este ca scorul TF-IDF (Salton and Buckley, 1988) al acelui termen să fie peste un anumit prag (Ion et al., 2009b; Ştefănescu, 2010). În urma experimentelor, am stabilit ca valoarea acestui prag să fie 9. Folosirea acestui scor vizează eliminarea din interogare a termenilor comuni (cu frecvenţă de ocurenţă mare în documente) şi deci, cu capacitate discriminatorie mică.

ALGORITMUL DE GENERARE A CERERILOR FORMALE – VARIANTA 2 foloseşte, asemeni variantei 1, informaţia obţinută prin preprocesarea întrebării. Varianta precedentă a acestui algoritm, dezvoltată pentru căutările pe corpusul Wikipedia (Ion et al., 2009a) a fost extinsă şi optimizată pentru noul corpus. Ca şi în versiunea precedentă, algoritmul foloseşte verbele principale ale propoziţiei şi grupurile nominale identificate în faza de preprocesare pentru a genera cererea formală către motorul de căutare. Noutatea constă în faptul că pentru fiecare grup nominal sunt construiţi doi termeni ai cererii. (i) Primul este o expresie obţinută prin concatenarea cu spaţiu între ele a tuturor lemelor cuvintelor de tip conţinut în ordinea apariţiei lor în grupul nominal. Setăm apoi 2 parametri pentru expresia formată. Primul, pe care îl vom numi *parametru de proximitate*, se referă în principal la numărul de cuvinte ce pot fi intercalate între termenii expresiei. Valoarea sa este egală cu 1 plus numărul cuvintelor funcţionale ce se găsesc în grupul nominal. Al doilea parametru, pe care îl vom numi *parametru de amplificare*, defineşte importanţa expresiei ca termen al interogării. Dacă acest parametru este setat la o valoare *n*, acest lucru înseamnă că identificarea ulterioară a termenului echivalează cu identificarea a *n* termeni obişnuiţi. Am setat acest parametru la o valoare egală cu numărul unităţilor lexicale din expresia formată, pentru a favoriza paragrafele care conţin grupuri nominale identice cu cele din întrebare. (ii) Al doilea termen este o expresie Booleană formată doar din conjuncţia lemelor corespunzătoare cuvintelor conţinut din grupul nominal.

Observăm că primul termen este mai restrictiv pentru că impune o ordine de apariţie şi o anumită poziţionare în documente a cuvintelor din întrebare. În schimb, pentru al doilea termen condiţiile sunt relaxate. Urmărim ca atunci când cuvintele se găsesc în texte, dar nu strict în condiţiile impuse de primul termen, să putem extrage totuşi acele fragmente care le conţin. Ca şi în cazul versiunii precedente, pentru fiecare verb principal din întrebare (cu excepţia verbelor difuze semantic[6]) se generează un termen corespunzând lemei acelui verb. Operatorul boolean folosit implicit de Lucene este SAU (OR) şi

[6] Verbe cu putere discriminatorie redusă (a *fi*, a *avea*, a *putea*, etc.) întrucât utilizarea lor este comună în orice domeniu.

astfel, cererile sunt disjuncţii logice. Pe de altă parte, există şi situaţia în care nici un paragraf nu conţine o anumită expresie. În acest caz, cu ajutorul expresiilor conjunctive se încearcă identificarea celor care conţin cât mai multe cuvinte conţinut ce apar şi în întrebare.

Cele două cereri furnizate către motorul de căutare conduc la obţinerea a două seturi de paragrafe ca rezultate. Putem vorbi astfel de două sisteme, chiar dacă, în foarte multe din componentele ce le compun, aceste sisteme sunt similare.

Una din cele mai utilizate practici pentru a îmbunătăţi performanţele unui sistem QA este aceea de a îmbogăţi termenii mono-cuvânt din interogările formale cu sinonime extrase din lexicoane sau ontologii lexicale precum Pinceton WordNet (Fellbaum, 1998). Conţinând aproximativ 60.000 de mulţimi sinonimice, varianta românească a WordNet-ului (Tufiş et al., 2008a) este o excelentă resursă ce ar putea fi utilizată în acest scop. Cu toate acestea, datorită limbajului juridic specializat, caracteristic corpusului JRC-Acquis, am ales să nu facem uz de ontologia lexicală la nivelul generării interogărilor, ci într-o etapă ulterioară în care a fost necesară calcularea unui scor de similaritate lexicală dintre întrebare şi paragrafele având probabilitate ridicată de a conţine răspunsul corect.

6. Evaluare şi concluzii

Evaluarea acurateţei sistemelor înscrise în competiţia ResPubliQA a fost făcută de organizatori folosind un program ce implementează formula (3):

$$A = \frac{1}{n}(n_{RC} + n_{NOA} \times \frac{n_{RC}}{n}) \quad (3)$$

unde *n* reprezintă numărul total de întrebări (500), n_{RC} este numărul de răspunsuri corecte, iar n_{NOA} reprezintă numărul întrebărilor la care am răspuns cu NOA. După cum se observă, formula de evaluare punctează suplimentar (termenul al doilea al formulei) capacitatea unui sistem de a discerne situaţiile în care nu are certitudinea necesară de a da un răspuns corect şi decide să emită un răspuns de tip „nu ştiu” (NOA=„no answer”). Ipoteza organizatorilor a fost că acurateţea în aceste situaţii ar fi aceeaşi cu acurateţea răspunsurilor corecte ($\frac{n_{RC}}{n}$). Cu alte cuvinte, atunci când sistemul decide că „nu ştie” răspunsul corect, el este la fel de precis ca în cazurile în care furnizează un răspuns. Din analiza efectuată după primirea rezultatelor competiţiei reiese că în cazul sistemului nostru, precizia cu care răspunsurile greşite au fost evitate prin soluţii NOA a fost chiar mai mare decât precizia răspunsurilor corecte.

Pentru evaluare, ICIA a trimis organizatorilor competiţiei două seturi de rezultate. Primul set, ICIA091RORO, conţine rezultatele obţinute prin combinarea ieşirilor celor două variante ale sistemului fără a folosi clasele întrebărilor ca termeni ai interogărilor formale, în timp ce al doilea set, ICIA092RORO, foloseşte aceste clase ca termeni. Utilizarea claselor întrebărilor ca termeni în interogările generate a condus la rezultate semnificativ îmbunătăţite datorită faptului că în faza de indexare a paragrafelor am adăugat un câmp care să conţină clasa acestora.

Faza de antrenare desfăşurată pe cele 200 de întrebări (v. formula (1)) a condus la obţinerea următoarelor ponderi λ:

Tabelul 1: Ponderile pentru cele două variante de interogare ale sistemului

	λ_1	λ_2	λ_3	λ_4	λ_5
varianta 1 (interogări generate în funcţie de scorul TF-IDF)	0,22	0,1	0,1	0,19	0,39
varianta 2 (interogări generate folosind structura de grupuri propoziţionale a întrebărilor	0,32	0,14	0,17	0,25	0,12

Fiecare din variantele sistemului a primit la intrare cele 500 de întrebări ale competiţiei, pentru ca apoi să întoarcă o lista a celor mai relevante 50 de paragrafe pentru fiecare întrebare, folosind parametrii din tabelul de mai sus. Dintre acestea, în cazul în care un paragraf a satisfăcut ecuaţia (2) acesta a fost furnizat ca răspuns. În caz contrar, răspunsul generat a fost NOA. Tabelul 2 conţine evaluările oficiale ale rezultatelor trimise de ICIA: ICIA091RORO şi ICIA092RORO. Scorul de acurateţe (formula 3) pentru sistemul ICIA092RORO a fost cel mai bun dintre toate cele 44 de seturi de răspunsuri (28 trimise de participanţi, plus încă 16 scoruri de referinţă calculate de organizatori). Scorul de acurateţe pentru sistemul ICIA092RORO a fost al 4-lea.

Tabelul 2: Rezultatele oficiale ale ICIA la competiţia ResPubliQA pe limba română

	ICIA091RORO)	**ICIA092RORO**
Întrebări la care a fost dat un răspuns	393	344
Întrebări la care NU fost dat un răspuns	107	156
Întrebări la care răspunsul a fost CORECT	237	260
Întrebări la care răspunsul a fost INCORECT	156	84
Întrebări la care răspunsul a fost NOA	107	156
scorul c@1	0,58	0,68

Interesant de menţionat faptul că evaluarea cifrei de merit $M = \frac{A_{\hat{I}R}}{A_{RD}}$ (v. (Peñas, et al., 2009) şi Secţiunea 2) a pus în evidenţă mai obiectiv performanţa sistemelor înscrise în concurs în raport cu limba de întrebare şi răspuns. Cele două sisteme ale ICIA au obţinut cele mai mari cifre de merit (ICIA092RORO=1,55, respectiv ICIA091RORO=1,32). Următoarele cifre de merit au fost obţinute de sistemele de ÎR pentru limba italiană (1,24) şi limba spaniolă (1,175). În raport cu scorul de acurateţe c@1, aceste două sisteme s-au clasat pe locurile 8 şi respectiv 12. Aceste rezultate sunt foarte interesante şi ele demonstrează un lucru mai puţin evident: pentru limba engleză câştigul de acurateţe adus de tehnologiile de limbaj natural faţă de tehnologiile standard de regăsire a informaţiei este relativ mic (15%); acest lucru se explică prin faptul că limba engleză are o morfologie foarte simplă dar şi prin faptul că cele mai avansate tehnici de regăsire documentară au fost dezvoltate, testate şi optimizate pe documente scrise în limba engleză. Explicaţia se confirmă şi prin faptul că analiza scorurilor de acurateţe ale sistemelor pentru diferite limbi cu morfologie semnificativă (germană, franceză, spaniolă, italiană, bulgară) au fost mai mici decât scorurile de acurateţe pentru limba engleză, deşi unele dintre sistemele pentru limba engleză au fost elaborate de aceleaşi echipe care au implementat sistemele pentru limbile lor naţionale. De asemenea, ierarhizarea în funcţie de cifra de merit M, plasează cel mai bun sistem de ÎR pentru limba engleză[7] pe locul al 6-lea, în timp ce sistemul pentru limba germană, cu al 17-lea scor c@1, are cifra de merit 1,158 care-l plasează pe poziţia a 5-a.

[7] Sistemul uned092 a obţinut al doilea scor de acurateţe (c@1) 0,61, în timp ce cifra sa de merit (M) a fost de 1,151.

Am testat şi ipoteza conform căreia antrenarea ponderilor λ în funcţie de clasa întrebării va conduce la îmbunătăţirea rezultatelor. Am folosit încă odată cele 200 de întrebări pentru antrenare, partiţionate în funcţie de clasă, şi astfel am obţinut rezultatele din tabelul 3. Scorurile c@1 şi M s-au îmbunătăţit cu peste 3% faţă de cel mai bun rezultat precedent.

Tabelul 3: Ponderile rezultate după antrenarea pe fiecare clasă

		λ_1	λ_2	λ_3	λ_4	λ_5
varianta cu interogări generate în funcţie de scorul TFIDF	Factoid	0,1	0	0,2	**0,4**	0,3
	Definition	0,2	0,15	0,05	0,15	**0,45**
	Reason-Purpose	0,1	0	0,15	0,3	**0,45**
	Procedure	0,1	0	0,15	0,15	**0,6**
varianta cu interogări generate folosind structura de grupuri propoziţionale a întrebărilor	Factoid	0,15	0	**0,3**	**0,3**	0,25
	Definition	0,05	**0,5**	0,15	0,1	0,2
	Reason-Purpose	0,2	0	**0,4**	0,2	0,2
	Procedure	0,15	0,1	0,25	0,2	**0,3**

În final, trebuie să menţionăm că sistemul nostru este disponibil on-line[8], sub forma unei aplicaţii web ce apelează, pentru fiecare interogare, diferitele servicii web ce intră în arhitectura sistemului. În viitorul apropiat intenţionăm să ne concentrăm eforturile în direcţia dezvoltării unor capabilităţi cros-linguale care să dea posibilitatea utilizatorilor să interogheze sistemul în limbaj natural fie în limba română, fie în engleză, iar sistemul să întoarcă rezultatele în oricare din aceste limbi. De altfel, secţiunea în limba engleză a corpusului JRC-Acquis a fost deja preprocesată şi avem în vedere folosirea aplicaţiilor de traducere automată dezvoltate la ICIA (Ceauşu, 2009; Irimia, 2009) pentru a traduce fie întrebările în limbaj natural, fie cererile formale către motorul de căutare.

Mulţumiri. Activitatea de cercetare descrisă a fost sprijinită de CNMP – MECT prin proiectul naţional PNII SIR-RESDEC, D11007/18.09.2007.

Referinţe bibliografice

Ceauşu, A. (2009) *Tehnici de traducere automată şi aplicabilitatea lor limbii române ca limbă sursă*, Bucureşti, România: Teză de doctorat, Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română.

Fellbaum, C. (1998) *WordNet: An Electronic Lexical Database*, MIT Press.

Hatcher, E. and Gospodnetić, O. (2004) *Lucene in Action*.

Ion, R. (2007) *Metode de dezambiguizare semantică automată. Aplicaţii pentru limbile engleză şi română*, Bucureşti: Teză de doctorat, Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română.

Ion, R., Ştefănescu, D., Ceauşu, A. and Tufiş, D. (2009a) *RACAI's QA System at the Romanian-Romanian QA@CLEF2008 Main Task*, *Lecture Notes in Computer Science*, vol. 5706, September, p. 393–400.

[8] http://www2.racai.ro/sir-resdec/

Ion, R., Ştefănescu, D., Ceauşu, A., Tufiş, D., Irimia, E. and Barbu-Mititelu, V. (2009b) *A Trainable Multi-factored QA System*, Proceedings of CLEF2009 Workshop, Corfu, Greece.

Irimia, E. (2009) *Metode de traducere autmată prin analogie. Aplicaţii pentru limbile română şi engleză*, Bucureşti: Teză de doctorat, Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română.

Och, F.J. (2003) *Minimum Error Rate Training in statistical machine translation*, Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Sapporo, Japan, 160-167.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002) *BLEU: a method for automatic evaluation of machine translation*, Proceedings of the ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, Philadelphia, USA, 311-318.

Peñas, A., Forner, P., Sutcliffe, R., Rodrigo, Á., Forascu, C., Alegria, I., Giampiccolo, D., Moreau, N. and Osenova, P. (2009) *Overview of ResPubliQA 2009: Question Answering Evaluation over European Legislation*, Proceedings of the QA@CLEF workshop, Sept. 30 - Oct. 3.

Puşcaşu, G., Iftene, A., Pistol, I., Trandabăţ, D., Tufiş, D., Ceauşu, A., Ştefănescu, D., Ion, R., Orăşan, C., Dornescu, I., Moruz, A. and Cristea, D. (2007) *Cross-Lingual Romanian to English Question Answering at CLEF 2006*, Lecture Notes in Computer Science, Available: ISBN: 978-3-540-74998-1.

Radev, D.R., Qi, H., Wu, H. and Fan, W. (2002) *Evaluating Web-based Question Answering Systems*, Demo section, LREC 2002, Las Palmas, Spain.

Ratnaparkhi, A. (1998) *Maximum Entropy Models for Natural Language Ambiguity Resolution*, Philadelphia, PA, USA: PhD Thesis, University of Pennsylvania.

Salton, G. and Buckley, C. (1988) *Term-weighting approaches in automatic text retrieval*, Information Processing and Management, pp. 513-523.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. and Varga, D. (2006) *The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages*.

Ştefănescu, D. (2010) *Extragere inteligentă de informaţii din corpusuri multilingve*, Bucureşti: Teză de doctorat, Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română.

Ştefănescu, D. and Tufiş, D. (2006) *Aligning Multilingual Thesauri*, Proceedings of The 5th Language Resources and Evaluation Conference (LREC), Genoa, Italy.

Tufiş, D., Ion, R., Bozianu, L., Ceauşu, A. and Ştefănescu, D. (2008a) *Romanian Wordnet: Current State, New Applications and Prospects*, Proceedings of the 4th Global WordNet Conference (GWC-2008), Szeged, Hungary, 441-452.

Tufiş, D., Ion, R., Ceauşu, A. and Ştefănescu, D. (2008b) *RACAI's Linguistic Web Services*, Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008), Marrakech, Morocco.

Tufiş, D., Ştefănescu, D., Ion, R. and Ceauşu, A. (2008c) *RACAI's Question Answering System at QA@CLEF 2007*, Lecture Notes in Computer Science: Advances in Multilingual and Multimodal Information Retrieval (CLEF 2007), pp. 3284-3291.