Formularul B pentru proiecte complexe
fontul Arial 11 sau Times New Roman 12, spatiere – la un rind
Titlul proiectului (maxim 200 caractere)

Sistem de Întrebare-Răspuns în limbile Română şi Engleză cu Spaţii Deschise de Căutare (SIR-RESDEC)

Cuvinte cheie: (max 5 cuvinte)

Limbaj natural, ontologie, inferenta, extragere de cunostinte, servicii web

Rezumatul proiectului (maxim 1 pagina)

Din punctul de vedere al utilizatorului, cea mai naturala forma de interactiune intre acesta si masina este comunicarea in limba sa materna. Sistemele de intrebare-raspuns (SIR) sunt sisteme inteligente de regasire documentara ce folosesc tehnici ale prelucrarii limbajului natural si metode de rationament automat pentru a gasi informatia solicitata, in limbaj natural, de catre utilizator. Daca spatiul de cautare al raspunsului este structurat, omogen si relativ stabil (asa cum este de pilda o baza sau o banca de date) sistemele de intrebare-raspuns se numesc circumscrise. De regula, structura cunoasterii la care poate raspunde un SIR circumscris este modelata printr-o baza de cunostinte (ontologie a domeniului) ce abstractizeaza conceptele si relatiile dintre ele in cadrul universului de discurs in care trebuie circumscris dialogul om-masina. Daca insa spatiul de cautare este nestructurat, eterogen si dinamic (asa cum este de exemplu multimea documentelor de pe Internet accesibila prin World Wide Web) sistemele de intrebare-raspuns se numesc deschise. Abstractizarea cunoasterii in acest caz se poate realiza la un nivel foarte general prin asa-numitele ontologii de nivel superior cuplate cu ontologii lexicale pentru limba naturala folosita la interogare. Tehnicile de cautare si furnizare a raspunsului in cadrul sistemelor de intrebare-raspuns deschise (cel mai adesea construite peste motoare de cautare de tip Google, Yahoo, Altavista, MSN Search, AOL, Ask.com, Excite, Infoseek, etc) sunt fundamental diferite de cele folosite in sistemele circumscrise, posibilitatile lor inferentiale sunt extrem de limitate iar informatia utila este de cele mai multe ori scufundata in volume mari de documente nerelevante. Cercetarile in domeniul extraordinar de dinamic al Web-ului semantic sunt orientate spre automatizarea adnotarii semantice standardizate a documentelor si documentarii serviciilor publicate pe web, asigurarea interoperabilitatii semantice intre diferitele modelari conceptuale a continutului acestor resurse eterogene de informatii si servicii, astfel incat motoarele de cautare, si implicit sistemele de intrebare raspuns deschise, sa poata identifica cat mai precis informatia relevanta pentru o cerere arbitrara de informatie. Cea mai noua tendinta in abordarea sistemelor deschise de intrebare-raspuns consta in exploatarea tehnologiilor Web-ului semantic pentru a construi in mod dinamic colectii de documente relevante pentru un anumit univers de discurs si apoi, cu sprijinul ontologiilor de domeniu (existente apriori sau construite dinamic prin unificarea unor ontologii partiale si eterogene) utilizarea metodelor specifice sistemelor circumscrise.

Proiectul de fata propune realizarea unui sistem avansat, parametrizabil si interlingual de intrebare în limba romana si raspuns in limba romana sau engleza relativ la o colectie dinamica de documente continand un numar arbitrar de mare de texte (la limita, intregul web). Sistemul va include in arhitectura sa componente specifice unui sistem deschis de intrebare-raspuns dar va fi capabil sa isi rafineze precizia şi completitudinea raspunsurilor pe masura ce domeniul de discurs este mai bine precizat. Sustinut de ontologii de nivel superior si ontologii lexicale, sistemul, in mod dinamic va clasifica interactiunea cu utilizatorul intr-un domeniu de discurs deschis sau circumscris (existand o ontologie suport si o colectie relevanta de documente nestructurate dar indexate lexical). Domeniile alese ca studii de caz sunt domeniul legislativ si domeniul bioinformatic pentru care sunt avute in vedere colectii de documente nestructurate, de mari dimensiuni, anume ansamblul de documente in limba romana al Aquis Communautauire si respectiv colectia de documente in limba engleza a PubMed. Pentru ambele domenii de discurs exista in domeniul public ontologii partiale ce urmeaza a fi extinse si lexicalizate in limba romana de catre membrii consortiului acestui proiect.

Relevanta proiectului (maxim 1 pagina)

Proiectul propus se incadreaza in directia de cercetare nr. 1 "Tehnologia informatiei si Comunicatii", obiectivul specific de cercetare fiind 1.4 "Inteligenta Artificiala, robotica si sisteme autonome avansate" iar sub-obiectivele relevante fiind:

1.4.2 Dezvoltarea de sisteme bazate pe semantică în spațiul web; realizarea interoperabilității semantice între resurse eterogene de informație şi servicii, între diferitele tipuri de conținut, între diferitele limbi naturale;

1.4.4 Dezvoltarea de baze de cunoştinţe infrastructurale (ontologii de domenii, ontologii lexicale pentru limbile de interes) multimodale şi multimedia;

1.4.6 Dezvoltarea de sisteme de interacțiune naturală om – calculator minimal dependente de universul discursului;

Proiectul valorifica o serie de rezultate de exceptie obtinute de membrii consortiului in domeniul tehnologiilor lingvistice multilinguale (castigarea primelor doua competitii mondiale de aliniere lexicala de la Edmonton - Canada 2003 si Ann Arbor - SUA 2005, castigarea competitiei de rezolutie a anaforei ARE din cadrul DAARC, Lagos-2007, obtinerea locului 3 la competitia de implicatie/deductie lexicala TE-2007, anuntat la ACL-Praga in iunie 2007, dezvoltarea unor resurse lexicale multilinguale de referinta, etc) ca si al sistemelor de rationament in mediul web, aducand laolalta competente romanesti recunoscute peste hotare, in domenii complementare, in scopul realizarii unui program de actiune competitiv, aliniat la obiectivele internationale de referinta. Proiectul va contribui la dezvoltarea cunostintelor in domeniul achizitiei de cunostinte din texte nestructurate, al tehnologiilor de regasire inteligenta a informatiei relevante, va contribui la diseminarea bunelor practici si a standardelor sau recomandarilor internationale in exploatarea resurselor web-ului semantic, va contribui la sporirea utilizarii si a vizibilitatii limbii romane in spatiul Intenet

Scopul proiectului consta in dezvoltarea unor modele si tehnici noi de exploatare a cunostintelor continute in texte electronice, implementarea unui sistem de intrebare-raspuns in limba romana pentru volume mari de documente, sistem inedit atat in Romania cat si in strainatate, dezvoltarea unor metode de achizitie a cunostintelor factuale, existente in, sau deduse din texte nestructurate, dezvoltarea unei ontologii lexicale pentru limba romana cu mare acoperire lingvistica si aliniata cu ontologia lexicala de referinta pentru limba engleza. Proiectul va implementa doua aplicatii de mare actualitate si impact stiintific si socio-economic: un sistem de intrebare-raspuns in domeniul bioinformatic, pentru uzul direct al cercetatorilor din domeniu si medicilor si un sistem de intrebare-raspuns pentru uzul cetateanului dar si al specialistului in domeniul legislatiei Comunitatii Europene. Sistemele folosite ca demonstrator al conceptelor si modelelor dezvoltate in acest proiect vor putea fi facil adaptate la domenii conexe (de pilda legislatia Romaniei).

Aplicatia de bioinformatica va aborda interogari complexe din punct de vedere semantic privind maladii, gene, medicamente, tratamente si relatiile dintre acestea. Pentru aceasta este necesara realizarea interoperabilitatii semantice intre surse de informatii eterogene, pe baza careia vor putea fi combinate interogari textuale ale PubMed cu interogari catre alte surse de informatii bioinformatice din spatiul Web.

Descrierea proiectului din punct de vedere stiintific si tehnic, incluzind gradul de noutate si posibilitatea aplicarii rezultatelor cercetarilor (maxim 5 pagini)

Prezentarea succinta a stadiului realizarilor S/T din domeniu, la nivel national si international, raportat la tema proiectului

Volumul din ce in ce mai mare de cunostinte existente pe web a creat o competitie formidabila intre furnizorii de continut electronic pentru gasirea celor mai eficiente metode de stocare, indexare si regasire cat mai precisa a documentelor electronice relevante de pe Internet. Industria "captarii atentiei" cum mai este numita industria furnizarii de servicii web si continut electronic, centrata pe consumatorul din ce in ce mai pretentios si mai presat de lipsa timpului, are ca principala provocare, in raport cu beneficiarii sai, realizarea unor motoare de indexare si cautare din ce in ce mai performante, capabile sa foloseasca criterii semantice in intelegerea cererilor de regasire si depistarea documentelor relevante. Cea mai naturala modalitate de specificare a unei cereri de informatie ramane limbajul natural.

Realizarea de sisteme deschise sau semi-deschise (circumscrise) de intrebare-raspuns constituie obiective de prim rang ale cercetarii si dezvoltarii tehnologice in industria web, principala dificultate constand in posibilitatea interogarii volumelor imense de cunostinte de catre utilizatori in limba lor materna. Pentru cateva limbi de larga circulatie internationala exista deja o serie de sisteme de intrebare-raspuns in limbaj natural, cele mai performante fiind in limba engleza (e.g. START- dezvoltat la MIT de colectivul condus de prof. Boris Katz, FALCON - dezvoltat la UTD de colectivul condus de prof. Dan Modovan si prof. Sanda Harabagiu, AskJeeves, Yahoo Answers, etc.). Competitiile anuale TREC (Text REtrieval Conference) sponsorizate de Institutul National de Standarde al SUA (NIST) si Departamentul Apararii al SUA (DoD) au constituit si constituie forumurile americane cele mai prestigioase pentru evaluarea perfomantelor sistemelor avansate de regasire a informatiilor in colectii mari de documente. Sistemul de intrebare raspuns de referinta pentru limba engleza este sistemul FALCON, castigatorul competitiei TREC-9 cu un scor de relevanta a raspunsurilor (MRR) de 0,58 (maxim 1). Un sistem similar, PowerAnswer-2, dezvoltat de acelasi grup de cercetatori de la UTD a castigat competitia TREC din 2005, de data aceasta cu un scor de relevanta de 0,534. In Japonia, se organizeaza incepand din 1999 competitiile NTCIR (spnsorizate de Institutul National de Informatica din Japonia) de evaluare a sistemelor de intrebare raspuns pentru limbile din Asia. In Europa, incepand cu anul 2003 se organizeaza competiile CLEF (Cross Language Evaluation Forum), sponsorizate de Comisia Europeana, de evaluare a sistemelor multilinguale (altele decat pentru limba engleza) de intrebare-raspuns.

La competitia CLEF din 2006 a fost inclusa pentru prima data intre limbile de concurs si limba romana doi dintre participanti find membri ai consortiului prezent (ICIA si UAIC). Cel mai bun rezultat (MRR) pentru limba romana, obtinut de unul din membrii consortiumului (ICIA), a fost de 0,1316, modest in comparatie cu nivelul modial atins pentru limba engleza, dar competitiv in raport cu alte limbi nou venite. La competitia CLEF 2007, limba romana a fost din nou prezenta in competitie iar ICIA si UAIC s-au inscris in concurs. Rezultatele, inca neoficiale, arata insa o imbunatatire substantiala fata de rezultatele de anul trecut (scorul MRR fiind in jurul valorii 0,28). Unul din obiectivele majore ale acestui proiect este ca prin noi metode si algoritmi sistemul nostru de intrebare raspuns (monolingv sau cros-lingual) final sa aiba un scor MRR de peste 0,5.

Contributia proiectului la dezvoltarea cunostintelor in domeniu, inclusiv noutate si complexitate a solutiilor propuse

Criteriile de performanta asumate, ca si aplicatiile propuse prin acest proiect vor contribui la dezvoltarea cunostintelor in domeniul prelucrarii limbii romane prin crearea si/sau extensia resurselor lingvistice de baza pentru limba romana, proiectarea si implementarea unor algoritmi performanti, capabili sa prelucreze zeci de milioane de cuvinte in cateva minute, dezvoltarea de metode de stocare, indexare si regasire a unor volume foarte mari si dinamice de documente.

Sistemul de intrebare raspuns in limba romana sau cros-lingual (intrebare in romana, spatiul de cautare continand documente in limba engleza) propus a fi dezvoltat prin acest proiect, precum si aplicatiile pe legislatia europeana (Acquis Communautaire) si pe documentele din PubMed nu au echivalent in cercetarea romaneasca. Prin particularitatile limbii romane, sistemul propus va avea o complexitate mai mare decat in cazul limbii engleze, constituind o noutate si pe plan mondial. Complexitatea proiectului este accentuata nu numai de problemele specifice prelucrarii limbajului natural ci si de volumele foarte mari de documente ce vor fi prelucrate lingvistic, indexate si stocate in vederea exploatarii ulterioare (extragere de cunostinte noi, interogare in limbaj natural), precum si de implementarea aplicatiilor avute in vedere. De pilda, documentele din Acquis Communautaire ce au fost traduse in limba romana (si al caror numar continua sa creasca) contin peste 50,000,000 de articole lexicale iar numarul de rezumate din PubMed, relevante pentru aplicatia de bioinformatica, este de peste 10 milioane.

Obiectivele generale si specifice ale proiectului

Obiectivul general al proiectului este dezvoltarea unor tehnologii, produse si servicii inovative care sa permita utilizatorilor interactiunea cu calculatorul prin intermediul limbii romane.

Mai precis, proiectul propune realizarea unui sistem de intrebare-raspuns monolingv (romana) cat si cros-lingual (intrebare in limba romana, raspunsul extras din documente elaborate in limba engleza) in spatii de cautare deschise sau circumscrise, sistem parametrizabil atat in raport cu limba cat si cu domeniul de discurs. Pentru realizarea acestui obiectiv, proiectul pune in evidenta patru sub-obiective:

1. colectarea, adnotarea automata (adnotare morfosintactica, lematizare, depistarea entitatilor denumite, a grupurilor sintactice si a dependentelor intre cuvinte, dezambiguizare semantica automata) si corectarea unor volume mari de documente in doua domenii de discurs foarte diferite: legislatia europeana si literatura stiintifica in domeniul bio-medica. Prima aplicatie se adreseaza cetateanului comun ca si expertului in legislatie internationala, in timp ce a doua aplicatie de interogare bazata pe semantica a informatiilor din domeniul biomedical se adreseaza cercetatorilor din domeniu, precum si medicilor;
2. colectarea ontologiilor publice de domeniu, relevante pentru aplicatiile tinta, extensia lor si echivalarea conceptelor acestora cu sinseturile retelelor semantice lexicale ale limbilor romana si engleza;
3. demonstrarea functionalitatii sistemului de intrebare-raspuns in doua universuri de discurs foarte diferite: domeniul legislativ si respectiv literatura stiintifica in domeniul bio-medical; cele doua aplicatii mentionate vor permite utilizatorilor interogarea in limba romana asupra continutului corpusurilor de la punctul 1, sistemul furnizand raspunsuri in limba romana si/sau engleza. De asemenea, dorim sa dezvoltam o metoda prin care sa putem construi automat o baza de cunostinte dintr-un text adnotat (conform cu punctul 1) pe care sa o folosim apoi la justificarea raspunului;
4. dezvoltarea unei interfete WEB pentru SIR-RESDEC care sa faca sistemul accesibil utilizatorilor. Acestia vor putea interoga in romana colectia de documente si vor putea obtine raspunsuri in romana sau engleza la intrebarile lor.

Detalierea activitatilor in corelatie cu obiectivele propuse si prezentarea rezultatelor S/T corespunzatoare activitatilor prevazute. Schema de realizare privind rolul şi responsabilităţile fiecărui participant pentru realizarea proiectului, cu defalcarea pe activităţi (pentru fiecare activitate se va prezenta necesarul de om/luna pentru realizarea activitatii)

Din punct de vedere arhitectural, sistemul SIR-RESDEC va consta din urmatoarele trei mari componente:

1. un modul de prelucrare a intrebarii care asigura: a) procesarea primara a textului intrebarii (segmentare, lematizare, dezambiguizare morfo/lexicala, detectarea dependentelor sintactice si semantice intre unitatile lexicale ale intrebarii, dezambiguizare semantica a cuvintelor polisemantice pe baza ontologiei suport); b) identificarea focusului intrebarii si a tipului de raspuns cautat; c) identificarea cuvintelor cheie din intrebare care se vor cauta in colectia de documente impreuna cu variante lexicale derivate si/sau sinonimice ale lor; d) identificarea tipului de intrebare (factoid, definitie, lista etc.); e) forma standard de interogare utilizabila de motorul de cautare
2. un motor de cautare a colectiei de documente care constituie sursa de informatie ce trebuie interogata. Utilizand forma standard de interogare furnizata de modulul de prelucrare a intrebarii, se extrag documentele (iar in cadrul acestora paragrafele) care contin raspunsul la intrebare;
3. un modul de extragere a raspunsului care ruleaza pe rezultatele motorului de cautare si care “decupeaza” segmentul/segmentele de text din paragrafele relevante care constituie raspunsul/raspunsurile la intrebare.

Procesarea primara a intrebarii apeleaza la cele mai avansate metode si tehnici ale prelucrarii limbajului natural pentru a asigure analiza morfo-sintactica, identificarea dependentelor sintactico-semantice intre constituentii frazali ai intrebarii, prelucrarea extragramaticalitatilor (rezolvarea anaforelor si a elipselor), dezambiguizare semantica a cuvintelor polisemantice pe baza ontologiei suport. Membrii consortiului au dezvoltat majoritatea modulelor necesare pentru prelucrarea textelor in limba romana si engleza si aceste instrumente vor fi perfectionate si utilizate in cadrul acestui proiect. Focusul intrebarii se extrage prin aplicarea asupra intrebarii a unor sabloane sintactice, continand variabile, cu legarea variabilelor respective la cuvintele corespunzatoare din intrebare. Tipul de raspuns

cautat (dimensiune, volum, suprafata, data, persoana, companie, etc.) se determina din focus iar tipul de intrebare se afla de obicei prin clasificarea sabloanelor care identifica focusul.

Motorul de cautare este responsabil de gasirea acelor paragrafe, existente in documentele spatiului de cautare, care contin cu o mare probabilitate raspunsul cautat. Se utilizeaza in general motoare de cautare a documentelor bazate pe scoruri de relevanta ale cuvintelor cautate pentru documentele in care se afla (TF/IDF) cu fraze de interogare cu cuvinte/expresii cheie si conectori booleeni (AND, OR, NOT) cum ar fi de exemplu Lucene (vezi http://lucene.apache.org/).

Extragerea raspunsului este operatia cea mai importanta a unui SIR si este responsabila de gasirea unui sau mai multor fragmente de text (din toata colectia de documente accesibila sistemului!) care fie reprezinta raspunsul corect (si complet) al intrebarii fie il implica (textual entailment). Un prim nivel de extragere a raspunsului este dat de motorul de cautare care furnizeaza paragrafele documentelor in care se regasesc cele mai multe din cuvintele cheie ale intrebarii si de asemenea entitati de tipul focusului. Acest nivel va fi rafinat in sensul gasirii acelui fragment de text care sa constituie explicit sau implicit raspunsul la intrebarea pusa.

Etapa 1. (dec. 2007) Evaluarea cerinţelor şi funcţionalităţii sistemelor de interogare inteligentă bazată pe semantică. Analiza surselor de informaţii şi ontologiilor din domeniile aplicatiilor, precum şi a tehnologiilor de interogare a acestora.

Activitatea 1.1 Analiza tipurilor de resurse lingvistice si prelucrari textuale necesare unui sistem de intrebare raspuns in limbaj natural – ICIA (2 om/luna)

Activitatea 1.2 Evaluarea cerinţelor şi funcţionalităţii sistemelor de interogare inteligentă bazată pe semantică în domeniul biomedical – ICI (1 om/luna)

Activitatea 1.3 **Analiza arhitecturii principalelor sisteme de intrebare raspuns in limbaj natural - UAIC (1 om/luna)**

Activitatea 1.4. **Determinarea cerinţelor utilizatorilor sistemelor curente, punerea în evidenţă a limitărilor şi respectiv a principalelor funcţionalităţi dorite de utilizatori, dar dificil de implementat cu tehnologiile existente** - ICIA (2 om/luna), UAIC (1 om/luna)

Activitatea 1.5. Analiza surselor de informaţii şi ontologiilor din domeniul biomedical, precum şi a tehnologiilor de interogare a acestora - ICI (0,5 om/luna)

Activitatea 1.6. **Analiza tipurilor de intrebari in limbaj natural pentru aplicatia de legislatie europeana** - UAIC (1 om/luna)

Rezultatele etapei: *a) Raport de cercetare asupra celor mai avansate SIR in limbaj natural; b)Raport asupra surselor de informatii si ontologiilor existente in domeniile aplicatiilor*

Etapa 2. (iunie 2008) Colectarea si analiza surselor primare de informaţii din domeniile aplicatiilor, construirea unor lexicoane specifice domeniilor de discurs şi indexarea multicriteriala a colectiilor de documente rezultate

Activitatea 2.1 Colectarea si "curatirea" textelor de informatie nerelevanta pentru aplicatia de legislatie - ICIA (1 om/luna), UAIC (1 om/luna)

Activitatea 2.2 Colectarea şi analiza surselor primare de informaţii din domeniul biomedical (de exemplu colecţia de rezumate biomedicale Pubmed) – ICI (2 om/luna)

*Activitatea 2.*3.Construcţia de lexicoane de termeni specifici domeniului juridic (engleza, romana) ICIA (6 om/luna) , UAIC (2 om/luna)

*Activitatea 2.*4. Construcţia de dicţionare de termeni specifici domeniului biomedical (sinonime de nume de gene, proteine, termeni medicali, maladii etc) – ICI (3 om/luna)

Activitatea 2.5 Analiza unui sistem de indexare public si fara restrictii de utilizare in cercetare (Lucene) si adaptarea acestuia in conformitate cu necesitatile generale ale proiectului - ICIA (3 om/luna), UAIC (3 om/luna)

Rezultatele etapei: *a) Colectii textuale de mari dimensiuni continand documente relevante pentru aplicatiile proiectului; b) Raport asupra continutului colectiilor construite (date cantitative si calitative asupra continutului, documentarea primara); c) Lexicoane specifice domeniilor de discurs; d) Sistem de indexare multicriteriala a colectiilor mari de documente*

Etapa 3. (dec. 2008) Construirea corpusurilor aplicatiilor si a ontologiilor specifice. Adnotarea

ontologica a corpusurilor

Activitatea 3.1 Constructia corpusurilor aplicatiei legislative: prelucrarea lingvistica de bază a colectiilor de documente construite in etapa 2, adnotarea şi indexarea acestora – ICIA(2,5 om/luna), UAIC (1 om/luna)

Activitatea 3.2 Construirea unor ontologii specifice (partiale) pentru aplicațiile de legislatie utilizînd tezaurele sau ontologiile specializate din domenile respective (EUROVOC, Legal Ontology) , UAIC (2 om/luna)

Activitatea 3.3 Constructia corpusurilor aplicatiei de bioinformatica: prelucrarea lingvistica de bază a colectiilor de documente construite in etapa 2, adnotarea şi indexarea acestora – ICIA(2,5 om/luna)

Activitatea 3.4 Construirea unor ontologii specifice (partiale) pentru aplicația de bioinformatică utilizînd tezaurele sau ontologiile specializate din domenile respective (Gene Ontology, MeSH, OBO - Open Biomedical Ontologies, etc.) – ICI (2 om/luna)

Activitatea 3.5 Imbogatirea ontologiei lexicale pentru limba romana (Ro-WordNet) cu sinseturi ce au corespondente cu conceptele din ontologiile de domeniu - ICIA(2),

Activitatea 3.6 Punerea in corespondenta a sinseturilor din ontologiile lexicale pentru limba romana (Ro-WordNet) si limba engleza (Princeton-WordNet) cu conceptele ontologiilor de domeniu pentru aplicatiile SIR-RESDEC - ICIA(1)

Activitatea 3.7. Dezvoltarea de instrumente de adnotare ontologica a corpusurilor (de exemplu bazate pe Wordnet şi ontologiile de domeniu) – ICIA(2), UAIC(1), ICI(2)

Activitatea 3.8. Evaluarea statistica a adnotarii ontologice a corpusurilor ICI(1)

Rezultatele etapei: *a) Doua corpusuri textuale de mari dimensiuni, codificate XML si adnotate pana la nivelul 3 prevazut de XCES (cel mai fin: cuvant, proprietati morfo-lexicale, lema, dependente sintactice, concept), indexate multicriterial; b) Ontologii (partiale) ale domeniilor aplicatiilor; c) Ontologii lexicale extinse pentru limba romana si engleza; d) Modul software de adnotare ontologica automata a corpusurilor; e) Raport asupra corpusurilor si ontologiilor*

Etapa 4. (iunie 2009) Implementarea arhitecturii nucleu a sistemului SIR-RESDEC

Activitatea 4.1 Proiectarea sabloanelor generice pentru intrebari in spatii de cautare deschise si construirea unei baze de date cu sabloane specifice domeniilor de discurs ale aplicatiilor ICIA (2), UAIC (2), ICI(1)

Activitatea 4.2 Implementarea unui clasificator automat al intrebarilor utilizatorilor (intrebari de tip factual, de tip definitie, de tip explicatie, intrebari enumerative etc.) si al elementelor focale ale intrebarilor (topic, focus, tipul raspunsului asteptat) ICIA (2)

Activitatea 4.3 Constructia semi-automata a unei baze de cunostinte pe baza corpusurilor si ontologiilor realizate in etapa 4; extragerea de dependențe utilizînd un parser de dependențe pentru instanțierea relațiilor din ontologie; instrumente de extragere de cunoştințe bazate pe reguli ICIA (3), ICI (1)

Activitatea 4.4 Proiectarea unui modul de inferenta pentru gasirea, cu ajutorul bazei de cunostinte de la 4.3, a raspunsurilor neexplicite si a fragmentelor de text ce sustin rationamentul (utile pentru argumentarea raspunsurilor inferate) ICIA (3), UAIC(2), ICI(1)

Activitatea 4.5. Instrumente de interogare inteligentă bazată pe semantică şi ontologii a surselor de informații biomedicale – ICI(2)

Rezultatele etapei: *a) Implementarea arhitecturii nucleu a sistemului SIR-RESDEC; b) Raport de implementare*

Etapa 5. (dec. 2009) Implementarea sistemului SIR-RESDEC si a celor doua aplicatii (prototip)

Activitatea 5.1 Specificarea limbajului de interogare bazat pe logică (limbajul de interogare va trebui sa poată referi textul primar, adnotări ale acestuia, termeni din ontologie asociați – şi toate acestea combinate în mod arbitrar) - ICI(2), ICIA (2)

Activitatea 5.2 Implementarea completa a sistemului SIR-RESDEC de interogare in limba romana a surselor de informații relevante pentru aplicatiile proiectului - ICIA (5), UAIC(3), ICI(2)

Activitatea 5.3 Implementarea prototipului aplicatiei: QA-Acquis pentru un subset reprezentativ al Acquis-Communautaire, respectiv intreaga colectie de documente din Acquis-Communautaire care au fost traduse in limba romana. ICIA (1), UAIC(1)

Activitatea 5.4 Implementarea prototipurilor aplicatiei: QA-PubMeD pentru un subset reprezentativ al bazei de date Pubmed de rezumate de articole din domeniul biomedical ICI (2)
Activitatea 5.5 Rafinarea implementării SIR-RESDEC şi a interfeței acestuia. ICIA (2), ICI(1)
Rezultatele etapei: *a) Implementarea sistemului SIR-RESDEC; b) Prototipurile celor doua aplicatii* QA-Acquis si QA-PubMeD; c) *Raport de experimentare*
Etapa 6. (sept. 2010) Finalizarea QA-Acquis si QA-PubMeD şi testarea lor intensivă
*Activitatea 6.1C*ompletarea ontologiilor utilizate pentru intelegerea intrebarilor si gasirea raspunsurilor ICIA (4), ICI (2), UAIC(2)
*Activitatea 6.2 P*rocesarea şi indexarea componentei de limba engleza a Acquis-Communautaire (precum si a documentelor noi ce vor fi tradude in limba romana) ICIA (4), UAIC(1.5)
*Activitatea 6.3 P*rocesarea şi indexarea întregii baze de date Pubmed de rezumate de articole din domeniul biomedical (care conține practic toate rezumatele publicațiilor din literatura bio-medicală – de ordinul a 15 milioane de rezumate) – ICIA (4), ICI (2)
Activitatea 6.4 Testarea SIR-RESDEC (ICIA 1) QA-Acquis (UAIC 1), QA-PubMed (ICI 1)
Activitatea 6.5 Implementarea SIR-RESDEC ca serviciu web. Accesul la aceste servicii va fi furnizat pe baza de licenta ICIA (2)
Rezultatele etapei:
a) Implementarea sistemului cadru SIR-RESDEC si a celor doua aplicatii QA-Acquis si QA-PubMeD*; b)Raport de testare si evaluare; c) Serviciu web de intrebare-raspuns in limbaj natural pentru aplicatiile* QA-Acquis si QA-PubMeD
Viabilitatea si riscurile proiectului
Experienta capatata de ICIA si UAIC prin participarea la doua competitii ale sistemelor de intrebare-raspuns CLEF2006 si CLEF2007 cu rezultate notabile (cele doua sisteme au avut cele mai bune rezultate pentru limba romana), constituie un punct de pornire extrem de solid in cercetarea acestui proiect. Sistemele s-au bazat pe solutii diferite dar complementare, existand premisele realizarii in colaborare a unui sistem mult mai performant. Acest proiect isi propune sa dezvolte noi metode de extragere indexare si regasire a documentelor potential relevante pentru o anumita intrebare ca si noi metode de extragere a raspunsului (partea cea mai deficitara in sistemele anterioare). Avand in vedere realizarile precedente ale ICIA si UAIC in prelucrarea primara a textelor in limbaj natural, resursele lexicale dezvoltate pentru limba romana (dispunem de o retea semantica lexicala Ro-WordNet cu aproximativ 40000 de concepte care acopera in mare parte fondul lexical al limbii romane si in plus sunt echivalate cu conceptele ontologiilor SUMO&MILO) si experienta bogata a ICI in dezvoltarea si utilizarea unor sisteme inferentiale complexe, apreciem ca riscurile privind realizarea in bune conditii a proiectului sunt minime.

Riscurile privind aplicatiile prevazute in proiect pot aparea in primul rand din problemele legate de dreptul de proprietate intelectuala a documentelor ce vor fi colectate pentru constituirea spatiilor de cautare a raspunsurilor la intrebarile utilizatorilor. Avand insa in vedere ca demonstratoarele se vor baza pe documente publice (corpusul AcquisCommunautaire si arhiva PubMed) pentru care deja a fost obtinut acordul de utilizare in scopuri de cercetare, consideram ca cel putin pentru aceste aplicatii nu vor fi probleme majore. Astfel de volume de documente sunt manipulate cu mijloacele tehnologiei motoarelor de cautare curente in care ICIA are o experienta foarte buna prin utilizarea si adaptarea sistemului public Lucene (http://lucene.apache.org/). Alternativ, pentru aplicatia bio-medicala este posibila combinarea motorului de cautare PubMed bazat pe cuvinte cheie cu prelucrarea aprofundata oferita de SIR-RESDEC dezvoltat in proiect.

Modalitățile de valorificare a rezultatelor – potențiali beneficiari
Aplicatia din domeniul legislatiei comunitare (Acquis Communautaire) va fi realizata ca serviciu web public, la dispozitia tuturor cetatenilor din Romania sau strainatate. Aplicatii similare, se vor putea realiza pe siturile publice ale Parlamentului Romaniei, ale ministerelor, ale primariilor, si in general ale tuturor actorilor publici din domeniul legislativ.

Aplicatia de bioinformatica se va adresa atat cercetatorilor din domeniul biologiei, cat si

medicilor care au nevoie de un acces rapid la cunostintele noi in legatura cu cazuistica mai complexa cu care se confrunta in practica medicala.

De asemenea, partenerul ICI este in prezent implicat in doua proiecte interdisciplinare (cu parteneri din domeniul biomedical) care studiaza mecanismele moleculare ale cancerului pancreatic si respectiv de colon. Experienta acestor proiecte a demonstrat necesitatea unui sistem mult mai sofisticat de interogare decat sistemele existente. Sistemul dezvoltat in proiect va fi utlizat si pentru analiza rezultatelor proiectelor mai sus mentionate.

Diseminarea rezultatelor

Rezultatele cercetarilor si realizarile software vor fi prezentate la conferintele de profil in fiecare faza a proiectului. Intentionam sa participam la cele doua competitii internationale de referinta in domeniul SIR, CLEF si TREC, pentru a putea evalua performantele sistemului de intrebare-raspuns bazat pe deductie in comparatie cu cele mai bune SIR din lume.

Vor fi elaborate lucrari stiintifice si vor fi transmise la conferintele internationale relevante pentru cercetarile din cadrul acestui proiect cum ar fi: Language Resources and Evaluation Conference, European Association for Computational Linguistics Conference, COLING Global WordNet Conference, European Conference on Artificial Intelligence, etc.

Pe langa participarea la conferinte si competitii internationale rezultatele proiectului vor fi publicate si in reviste cu impact (reviste ISI, reviste ale Academiei Romane). Vor fi organizate prezentari si demonstratii publice: la Scoala de vara Eurolan 2009, la workshopurile nationale ale Consortiului pentru Informatizarea Limbii Romane, la Parlamentul Romaniei, la Ministerul de Externe, la institutii publice cu responsabilitati in domeniul legislativ. Se vor realiza prospecte publicitare si CD-uri de demonstratie pentru a fi distribuite potentialilor utilizatori interesati.

Pentru sistemul SIR-RESDEC ca si pentru cele doua aplicatii se vor face demersurile pentru protectia intelectuala si se vor depune cereri de patentare in tara si strainatate.

Impactul generat de proiect (maxim 1/2 pagina)

Impactul economic al proiectului

Avand in vedere ca solutia de realizare a sistemului SIR-RESDEC este minimal dependenta de universul de discurs, vor putea fi implementate aplicatii specializate la orice agent economic sau social interesat de a-si promova expertiza, serviciile sau produsele proprii unei mari mase de utilizatori nespecialisti in navigarea pe Internet. Utilizarea limbii romane pe Internet va creste semnificativ (acesta fiind un indicator major al serviciilor localizate oferite de marii jucatori pe piata e-Bussiness). Gradul de informare al decidentilor romani in luarea deciziilor economice va spori prin utilizarea unui sistem de intrebare-raspuns in limba romana, bazat pe semantica si deci mai precis decat sistemele bazate pe cuvinte cheie. Aplicatia in domeniul bio-informatic va permite cercetatorilor in domeniu dar si medicilor identificarea mai exacta a informatiilor relevante scufundate intr-un ocean documentar, dandu-le posibilitatea de a se focaliza mai bine in studiile, diagnozele si tratamentele efectuate.

Impactul social al proiectului

Impactul social al acestui proiect va fi substantial, prin faptul ca va permite unei mari categorii de oameni accesul la legislatia Comunitatii Europene (precum si la alte sisteme de reglementari nationale sau locale) in propria limba, ridicand substantial gradul de informare si constientizare sociala.

Impactul asupra mediului

Ca orice aplicatie in tehnologia informatiei, impactul direct al SIR-RESDEC asupra mediului este nul. Indirect insa, o aplicatie specializata pe probleme de mediu, poate sprijini prin informarea corespunzatoare actiunile de prevenire si combatere a efectelor nocive asupra mediului ca si a calamitatilor si dezastrelor naturale.

Managementul proiectului. Alcatuirea consortiului (maxim 1,5 pagini)

Experienta coordonatorului in domeniu si in managementul proiectelor nationale/ internationale

Coordonatorul de proiect, prof. Dan Tufis, membru corespondent al Academiei Romane, este un cercetator cu indelungata experienta de cercetare in domeniul prelucrarii limbajului natural, cu realizari cunoscute pe plan national si international. Prof. Dan Tufis, impreuna cu profesorul Dan Cristea, conducatorul echipei UAIC, au realizat in anii '80 primul sistem de dialog in limba romana IURES,

pentru universuri de discurs foarte bine precizat si implementat pe minicalculatoarele CORAL si INDEPENDENT. Ulterior, sistemul IURES a fost implementat si pe calculatoare personale (PC-IURES) si a constituit primul sistem romanesc de inteligenta artificiala omologat international si vandut la export. Prof. Dan Tufis are o vasta experienta in coordonarea proiectelor nationale si internationale (a se vedea lista proiectelor in CV). In tara a coordonat 10 proiecte mari, incluse in planul national de cercetare, 2 programe fundamentale ale Academiei Romane precum si numeroase teme de cercetare incluse in planul regulat de cercetare al Academiei Romane. A coordonat pentru partea romana 15 proiecte europene (Copernicus, INCO, Esprit, ACTS, FP5, FP6, COST), 10 proiecte internationale bi- sau trilaterale cu parteneri din Anglia, Bulgaria, Elvetia, Franta, Germania, Grecia, Rusia, Serbia, Turcia, SUA. In prezent prof. Dan Tufis coordoneaza echipa de cercetare a ICIA in mai multe proiecte internationale si nationale:

-proiectele UE RomNet-Era ("Romanian Inventory ant Networking for Integration in ERA", nr. contract 510475), FP6 ProLearn (NoE on Professional Learning, nr. contract 507310; partener asociat), KnowledgeWeb (NoE on Semantic WEB, nr. contract 507482; partener asociat), proiectul BSEC (Black See Economic Cooperation) WISE, proiectul AUF (Agentia Universitatilor Francofone) " Collocations en contexte: extraction et analyse contrastive" si proiectul COST A31 “Stability And Adaptation of Classification Systems in a Cross-Cultural Perspective"COST A31.

-proiectul ROTEL din cadrul programului national Cercetare de Excelenta (CEEX-modulul I).

Prof. Dan Tufis a fost unul dintre initiatorii programului european pe 10 ani CLARIN (Common Language Resource Infrastructure) inclus in strategia europeana a cercetarii (European Research Infrastructure Roadmap) aprobata de Parlamentul European in noiembrie 2006. In programul CLARIN (prof.Dan Tufis este Vice-Presedinte al Comitetului Stiintific al CLARIN), inclus in sectiunea Infrastructuri a FP7, pentru primul apel a fost elaborat o propunere de proiect in care sunt inclusi ca parteneri ICIA si UAIC. O parte dintre obiectivele proiectului SIR-RESDEC vor beneficia direct de Programul CLARIN si proiectele ce vor fi realizate in cadrul acestui program in primul rand sub aspectul standardizarii codificarii resurselor specifice ce vor fi create si a serviciilor web prin care aplicatiile SIR-RESDEC vor fi puse la dispozitia utilizatorilor.

Experienta partenerilor in domeniu si realizarea de proiecte nationale/ internationale

Consortiul pentru proiectul SIR-RESDEC este acelasi cu cel al proiectului CEEX ROTEL aflat in derulare si care a fost apreciat pana in prezent ca unul dintre cele mai de succes proiecte CEEX ceea ce constituie o premiza favorabila a bunei colaborari intre participant si a desfasurarii proiectului in cele mai bune conditii. Obiectivele SIR-RESDEC au fost stabilite avand in vedere valorificarea unor rezultate foarte bune, cu impact international, obtinute in ROTEL.

Grupul de cercetare in prelucrarea limbajului natural de la Facultatea de Informatica a Universitatii "A.I.Cuza" , Iasi, coordonat de prof. Dan Cristea este nu numai unul dintre cele mai vechi din Romania, dar este totodata unul din cele mai active si mai productive, bucurandu-se de o foarte buna apreciere internationala si nationala. In prezent, acest grup de cercetare este implicat in doua proiecte CEEX (ROTEL si InterOb) in proiectele UE FP6 LT4eL (Language technology for e-Learning, coordonator al partii romane, nr. contract IST 027391) KnowledgeWeb (NoE on Semantic WEB, nr. contract 507482; partener asociat) si proiectul INTAS 05-104-7633:RoLTech *(*Platform For Romanian Language Technology: Resources, Tools And Interfaces, coordonator al intregului proiect).

Cercetarile in domeniile Semantic Web, al sistemelor de reprezentare a cunostintelor si inferenta si al bio-informaticii (analiza resurselor de informatii in biologia moleculara, genetica etc), desfasurate de echipa de cercetare din ICI, condusa de dr. Liviu Badea sunt printre cele mai cunoscute in comunitatea stiintifica interna si internationala, cu rezultate deosebit de apreciate in strainatate si comunicari la cele mai prestigioase conferinte ale domeniului inteligentei artificiale (IJCAI, ECAI, ECML, ILP etc.). Incepand cu anul 1993, ICI a fost implicat in peste 40 de proiecte europene, la realizarea multora dintre ele participand si colectivul implicat in acest proiect. In prezent colectivul din ICI, implicat in acest proiect, participa la proiectul european REWERSE (Reasoning on the Web with Rules and Semantics, nr. contract 506779) ale carui obiective sunt extrem de relevante pentru realizarea SIR-RESDEC si a aplicatiei in bioinformatica.

Metodele/modalitatile de conducere, coordonare şi comunicare pentru realizarea proiectului

Coordonatorul proiectului (ICIA) va reprezenta consortiul si va asigura legatura cu participantilor la proiect cu institutia finantatoare. Va asigura respectarea termenelor si a livrarilor prevazute in programul de lucru al proiectului in cele mai bune conditii de calitate stiintifica si tehnica. Coordonatorul proiectului va urmari realizarea deciziilor privind solutiile tehnice, diseminarea si exploatarea rezultatelor proiectului precum si a protectiei drepturilor intelectuale asupra acestor rezultate, luate prin consens de catre reprezentantii echipelor ce formeaza consortiul. Va asigura gestiunea financiara a proiectului intr-un mod deschis si eficient, cu respectarea tuturor reglementarilor in vigoare. Pentru bunul mers al proiectului, pe langa comunicarea continua prin telefon si/sau e-mail, se vor organiza intalniri de lucru periodice (trimestriale, si in plus, oridecateori va fi nevoie) la care vor participa obligatoriu coordonatorii celor trei echipe de cercetare (sau in cazuri deosebite, loctiitorii desenati ai acestora) precum si membrii fiecarei echipe a consortiului, responsabli ai uneia sau mai multor activitati curente. Intalnirile regulate sau fortuite vor fi anuntate cu cel putin o saptamana inainte iar locul lor de desfasurare se va stabili de comun acord.

Conducerea stiintifica a proiectului va fi realizata de Comitetul Stiintific al proiectului constituita din coodonatorii fiecarei echipe participante in consortiu. Pentru fiecare etapa a schemei de realizare a proiectului, coodonatorii fiecarei echipe vor desemna responsabili tehnici pentru activitatile fazei (Comitetul Tehnic al fazei), care vor urmari realizarea in cele mai bune conditii a activitatilor prevazute, vor prezenta dari de seama la intalniri si vor intocmi rapoartele tehnice pe baza carora membrii Comitetului Stiintific vor intocmi documentatia stiintifico-tehnica aferenta raportarilor de faza catre institutia finantatoare.

In cazul, putin probabil, al lipsei de consens, decizia va fi luata prin votul membrilor Comitetului Stiintific si al Comitetului Tehnic al fazei. In caz de indecizie, votul Coordonatorului de proiect va stabili decizia finala care va deveni obligatorie pentru toti membrii consortiului.