

Participarea la competiția de QA@CLEF-2006

Responsabilul științific al temei: **Dan Cristea**

Coordonatorii echipei de realizare a sistemului: **Adi Iftene**

Data acestui document: 6 aprilie 2006

I. Organizarea includerii limbii române (ca limbă sursă) în competiție

Dan – coordonarea sarcinii de organizarea a participării resurselor românești la competiție și prelucrarea constrângerilor temporale

- traducerea întrebărilor în românește:
 - o selectează setul de 200 întrebări de tradus → Ana
 - o prelucrează setul tradus și-l codifică XML → organizatori

Ana Masalagiu – traducere engleză-română

- preia de la Corina setul de întrebări și-l traduce
- colaborează cu Corina pentru codificarea XML și transmiterea setului la coordonatori

Termen: sf. aprilie

II. Realizarea sistemului de QA română-engleză

Adi – coordonarea sarcinii de QA RO-EN și recunoașterea numelor de entități

- material de test: corpusul și setul de întrebări din competiția trecută și cea de anul acesta
- rulează NER din GATE pe colecție
- confruntare cu rezultatele date de QA open-source
- transpunerea notației cu XSLT din offset la caracter (stand-alone) în inline
- o rulare pentru limba română
- categorii: location, date, person name, organization etc.

Cornel (an 2) – extractor de acronime, cu precizarea categoriei semantice (instituție, conferință etc.)

- colecția: CLEF-engleză
- de separat codul programului de resursă (setul de pattern-uri)
- de găsit o listă de abrevieri pentru română și engleză – de exemplu din dicționar
- de căutat un dicționar de acronime și abrevieri pentru română
- rezultat: <acronim> TAB <categorie> TAB <context>
- de apreciat P (eventual și R)

Maria – responsabilă cu resursa stop words

- lista de stop words pentru limba română și engleză

Ionuț – responsabil cu organizarea lanțului de procesări și determinarea tipului întrebărilor

- pos-tagging

- lematizare
- chunking??
- segmentare la propoziții
- listă de question patterns
- colecții: TREC și CLEF, plus întrebările colectate de noi

Diana – responsabilă cu resursele

- dicționar de echivalenți ro-en
- En-Ro alined WN
- Marcu ?? liste de echivalenți de traducere
- plasează în portalul Wiki toate resursele

Gabi Negară – motorul de indexare și căutare

- construiește un indexator, posibil ca o bază de date, în care interogarea se face SQL. colecția de texte pe care le indexează:
 - o POS-tăguite
 - o lematizate
 - o cu ID-uri de document și paragraf
 - o două module:
 - indexatorul: primește colecția de documente în format XML și lista de stop-words → întoarce indexul (bază de date) în care stop-words nu sunt indexate. Face o singură rulare;
 - motorul de căutare: la fiecare cerere primește o expresie logică de leme → un triplet: (scor, ID document, ID paragraf)
 - o vrea: schema documentului din intrare
 - o forma expresiei logice: fiecare rând ține o disjuncție de leme englezești, termenii disjuncției separați prin virgule
- vezi indexatoarele: Lucene și ... și Zprise. V. Information retrieval (Au)
-

Mădălina – responsabilă cu rezoluția anaforelor

- instalează RARE
- rulează RARE pe colecție
- proiectează reguli de rezoluție specifice

Cornel Barbu – responsabil cu dezambiguizarea semantică a cuvintelor din întrebare

- input: întrebarea adnotată XML (token cu pos, lemma)
- resurse:
 - corpusul adnotat la sensuri, de la Diana (SemCor)
 - RoWN de la Corina, cu documentația
 - dicționarul de la Diana
- output: expresia logică, pe care i-o transmite lui Gabi Negară

Diana Cotelea (an 3)

Iuliana Drăghici (an 3)

Alina Pițigoi (an 3)

- prima sarcină a echipei anului 3: montarea sitului Wiki al proiectului. Trăsături:
 - secțiune publică: informații generale, articole despre QA
 - secțiune privată cu acces pe bază de parole: documentația noastră, resursele, raporturi periodice etc.
- a doua sarcină: plasarea pe site a tuturor articolelor semnificative din domeniul QA și a unui raport *state-of-the-art* în care să se analizeze soluții de implementare pentru fiecare componentă a unui sistem QA.

Daniel Ohriniuc

- formularea răspunsului

Alex&Iustin

- rularea parserului FDG pe engleză și înjghebarea unuia pentru română
- transpunerea outputului în XML

Tratamentul constrângerilor temporale (discuția din 6 aprilie)

Gabi Mogoș – identificarea expresiilor temporale:

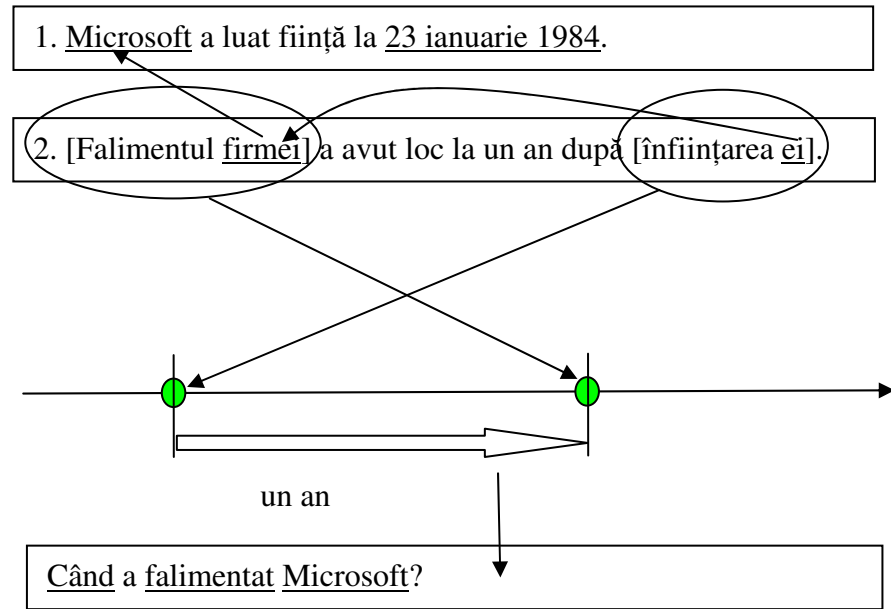
- primește de la **Adi** un set de exemple de codificări TIMEX
- culege o colecție (impresionant de mare!) de exemple în care apar expresii temporale
- **Gabi** construiește perechi de genul:

exemplu <TAB> codificarea TIMEX

```
anul trecut <TAB> <timex type="period"
start="01.01.2005" end="31.12.2005">anul
trecut</timex>
```

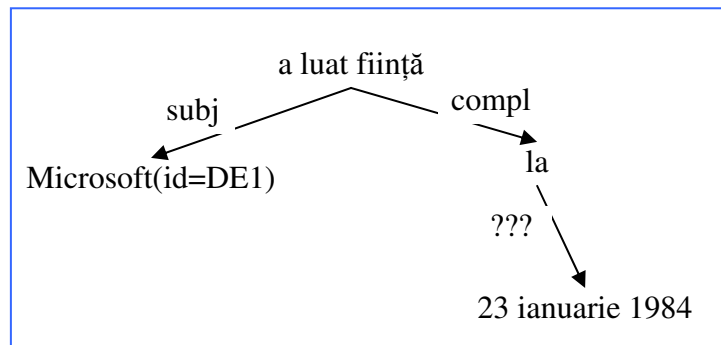
Termen: de stabilit.

- **Diana** aduce programele Perl de la cursul d-lui Ciortuz → **Cornel**
- **Cornel** dezvoltă acest nucleu cu alte pattern-uri care să acopere setul de exemple găsit de **Gabi**



Pași în raționament:

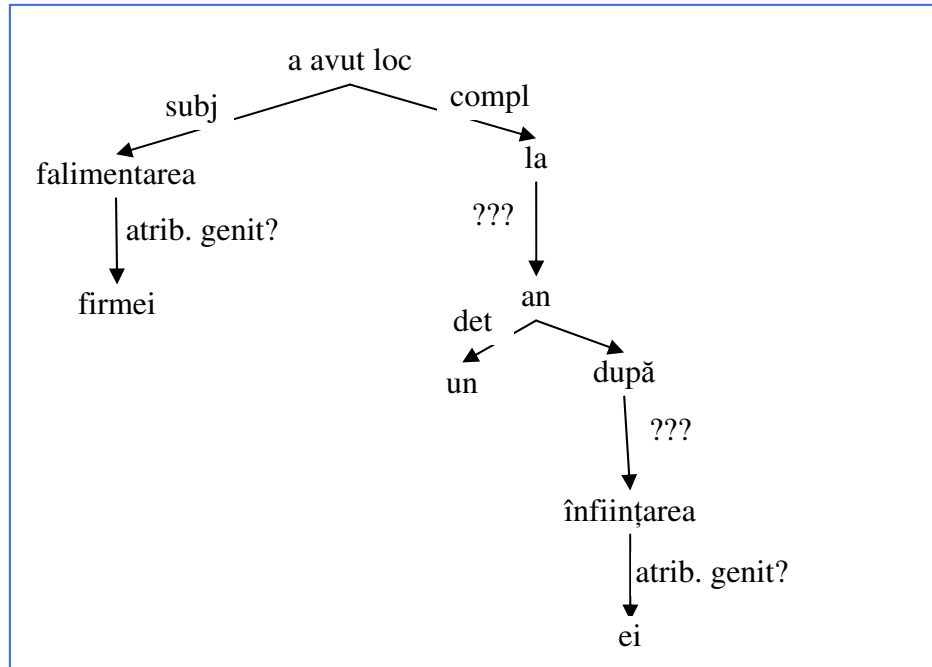
- Gabi Negară:** frazele 1 și 2 primesc scoruri de regăsire relativ mari față de altele: *Microsoft* este un indicator f. puternic (nume propriu), *falimentarea/faliment* este în imediata apropiere a unității în care apare Microsoft (cea de cel mai mare scor). → sunt ambele aduse de motorul de căutare
- Alex&Justin:** se parsează FDG 1:



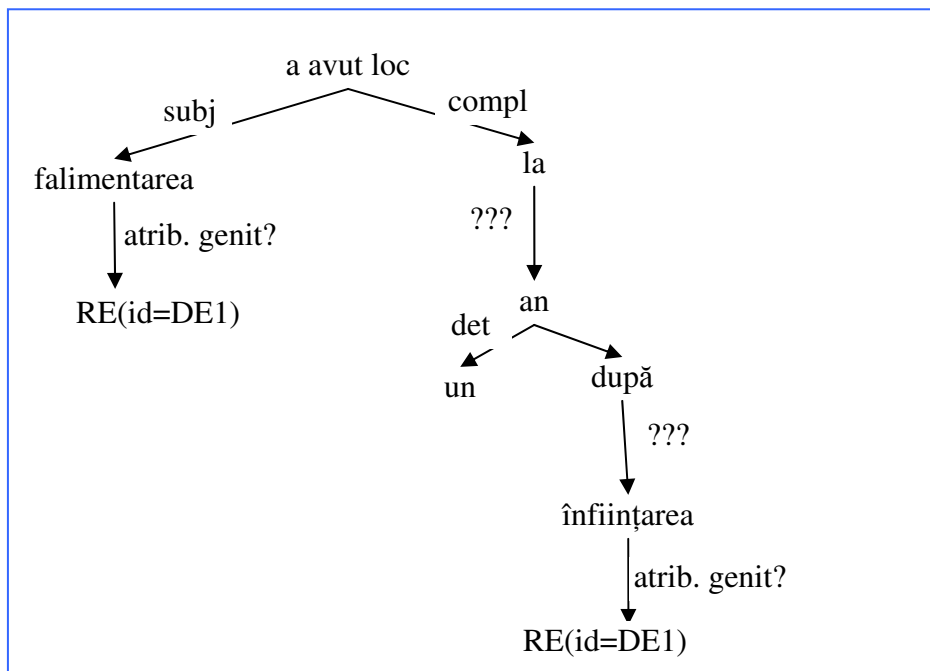
- Cineva1:** din ea se construiește o reprezentare evenimentială:

[a_lua_ființă: time=past → AG(Microsoft); → TIME(23.01.1984)]

- Alex&Justin** se parsează FDG 2:



e. **Mădălina:** RARE: găsește anaforele *firmei* → *Microsoft* și *ei* → *firmei* în 2.



f. **Cineva1:** se construiesc reprezentări evenimențiale pentru [falimentarea Microsoft] și [înființarea Microsoft] în 2. Motivul: *falimentarea* și *înființarea* sunt deverbative (lor le corespund verbe → le corespund reprezentări evenimențiale):

[falimenta:id=EV1 → AG(?); → REC(id=DE1)]

[înființa:id=EV2 → AG(?); → OBJ(id=DE1)]
[a_avea_loc: time=past → AG(id=EV1); → TIMEX(type=AFTER,
REF=EV2, VAL=1y)]

- g. **Maria&Diana&Cineva3**: aplic reguli de transformare semantico-lexicală:
a luat ființă = înființarea

din regula:

X înființează Y: [înființa → AG(X); → OBJ(Y)] ↔ Y ia ființă: [a_lua_ființă
→ AG(Y)] cu copierea celorlalte argumente
pot face transformarea:

[a_lua_ființă: time=past → AG(Microsoft:id=DE1); →
TIME(23.01.1984)] →
[înființa:id=EV2 → AG(?); → OBJ(id=DE1); → TIME(23.01.1984)]

- h. **Corina**: aplic inferențe temporale (TIMEX):

[a_avea_loc: time=past → AG(id=EV1); → TIMEX(type=AFTER,
REF=EV2, VAL=1y)] →
[a_avea_loc: time=past → AG(id=EV1); → TIME(aprox. 23.01.1985)]

- i. **Cineva2**: transfer constrângerile temporale ale evenimentelor de tip „a avea loc”
pe subiectele acestora:

[falimenta:id=EV1 → AG(?); → OBJ(id=DE1); → TIME(aprox.
23.01.1985)]

Rediscutarea exemplului d.p.d.v. a resurselor și a instrumentelor necesare:

Resurse:

Punctul g: echivalențe lexicale se găsesc și apoi se rescriu ca reguli de inferență.

establish, set up, found, launch -- (set up or found; "She set up a literacy program")

*> Somebody ----s something

be born -- (come into existence through birth; "She was born on a farm")

*> Something ----s

*> Somebody ----s

Microsoft was founded at 23 January 1984.

...

The bankruptcy of the company happened one year after its setting.

Maria&Diana: caută echivalări, reguli... pe engleză:

- o întrebare pe corpora@..

Instrumente:

Alex&Iustin: Punctele b și d: parsere FDG pe engleză (dar și română)

Cineva1: Punctele c și f: translatori de la reprezentări FDG la reprezentări evenimentiale.

Corina: Punctul h: inferențiator temporal: după notarea entităților temporale și a relațiilor se generează închiderea tranzitivă a valorilor.

Cineva2: Punctul i: injectarea constrângerii temporale a evenimentului principal pe subiect, dacă acesta este la rândul lui un eveniment. Se poate scrie o regulă.

Planificare

17 – 23 aprilie: facem să meargă motorul de indexare și căutare

26 – 30 aprilie: corpusul de întrebări – răspunsuri de anul trecut

1- 7 mai

8 – 14 mai: lanțul de prelucrări exclusiv formularea răspunsului + găsirea unei soluții teoretice pentru formularea răspunsului

15 mai – 28 mai: implementarea formulării răspunsului și teste

29 mai – 4 iunie: teste

ICIA

- Extragere de cuvinte englezești din corpusul mare care au traducere în RoWN.
- Completat RoWN cu cele mai frecvente synseturi din astea.
- Prelucrări asupra corpusului:
 - segmentare (compuși), tăguire, lematizare → dependency linking
- prelucrări asupra Q-Ro → segmentare (compuși), tăguire, lematizare → dependency linking →

19 aprilie – large talk with Dan Tufis

- prel. corpora - ..., mapare pe ROWN (~30.000 w)

- creere de sinseturi pe RO

- extragere nume proprii, clasificare (PERS, LOCATION, COMPANIES)

- sense-clustering monolingv, in WN; sensuri care se traduc prin sensuri diferite (pt identif trad. Corecte in EN)

- indexarea WSJ, LA pt dependente (linkuri obt. prin EM pe acelasi text; leg. sintactico-semantic)

- traducere: proc. Q; calculul dependentelor W din Q; pt fiecare pereche de W din Q – caut in WN; in index caut dependente deja gasite in corpora;

- TEQ va contine si hipo-hiperonime: in ce masura?
- Probl.: Indexare partiala; prop RO scurta => linkuri irelevante sau putine
- Calculul variantelor de interpretare (scoruri), incluzand linkage-urile
- Analiza (FDG) a candidatilor din Q

ICIA: (*)

Tokenizare

POS-tagging, lematizare

linkuri

Find parsers (**Alex**); Radu – la ICIA

Marius – v. parser Collins

Combinarea Alex + Collins

Combinarea cautarii (Google API) cu indexarea (fetele??)

Fetele: corpusul vechi de QA combinat si comparat cu cautarea pe Google

Gabi N.: indexator; are nevoie de input (XML), stop-words

Indexari dupa diferite criterii, acelasi output (DL: prima sapt. dupa Paste)

Pune pe wiki rezult.

Alex + Iustin – indexator Lucene (analog – wiki)

Maria: cuvinte cheie din Q, cu TEQ in EN

ICIA -> UAIC: Echivalentii pozitionali ai W din RO Q (fisier de aliniere, prin pozitii relative);

Next: expresiile logice

Participarea RO:

Diogene (sist QA pe ENG) – de folosit, cu specificatiile primite pt intrare

Servicii web ICIA – doar pt ENG deocamdata; prin SOAP, WSDL, UDDI (nlp.racai.ro IN: UTF8); de folosit pt un sist. Viitor

Fetele: vad cum e cu specificatia Diogene (mail DC)

2-3 saptamani:

1. **ICIA:** prel. (*) puse pe wiki, in format XCES
2. FDG – **Alex +Iustin**
3. NER (Gate) (**Ionut**)
4. Prelucrarea/analiza Q: 2 moduri (unul pt A intern, unul pt A cu Diogene)
D+M: pattern Q; Q type traduse direct in EN
5. expandarea Q in expr logic (**ICIA**)
6. key-words in ENG – la indexatorul lui **Gabi**
7. **Mada:** AR pe Q si pe corpusul de A, pe fiecare doc separat (dupa ICIA)
8. **Fete:** zona Google – vezi supra

9. Gabi +: adn temporale

Marți 25 aprilie

Iustin s-a jucat cu LUCENE:

- face indexare la nivel de articol
- va face retrieval la nivel de paragraf (2 mai)
- scor configurabil (tf-idf) cu un factor de boost:
 - focus, ponderi care vin din clasamentul sinonimelor, din alinierea RO-EN, din WSD – de studiat
- se va rula pe setul de întrebări din anii trecuți:
 - input: Q → se filtrează la stop words → se transmite lui LUCENE → paragrafele aduse de el se compară cu cel corect → se evaluează sistemul de retrieval → pe acest prim sistem putem jongla cu diverse ponderi care depind de nume proprii, ordinea cuvintelor etc., până se ajunge la un maxim → se obține astfel un baseline plecând de la care se pot face apoi îmbunătățiri (2 mai).

Diana propune o abordare bazată pe perechi de pattern-uri Q-A, cu adnotarea și indexarea corpusului la Nes și expresii temporale li regăsire după pattern-uri. De studiat ca o alternativă (poate împreună cu Adi?).

Diana plasează pe Wiki setul de Q-A din anii trecuți (2 mai).

Ionuț indexează corpusul mare și cel de întrebări la NEs (2 mai).

Corina și Alex testează GUTime pentru etichetarea la expresii temporale a corpusurilor (2 mai?).