Teme de licenţă 2016
Dan Cristea

1. Parametrizarea Portalului COROLA
în cotuletă cu lector dr. Alex Moruz (alex.moruz@gmail.com), drd. Andrei Scutelnicu (andreiscutelnicu@gmail.com)

COROLA este un acronim de la *COrpus of ROmanian LAnguage*. Numele complet al proiectului este Corpus Computaţional Reprezentativ al Limbii Române Contemporane, un proiect prioritar al Academiei Române, în curs de realizare la Institutul de Inteligenţă Artificială Bucureşti şi Institutul de Informatică Teoretică din Iaşi. Portalul proiectului poate fi accesat la adresa: http://78.96.45.245/. Actualmente Portalul găzduieşte tehnologii care realizează următoarele operaţiuni:

- curăţirea textelor primite de la colaboratorii noştri (edituri, autori etc.) de formate de tipărire - proces semiautomat;
- completarea metadatelor ce trebuie să însoţească documentele (autor, editură, an apariţie etc.) - proces semiautomat;
- lansarea unui lanţ de prelucrări lingvistice care urmăresc adnotarea automată cu informaţii de segmentare şi de natură morfologică: graniţe de propoziţii şi cuvinte, părţi de vorbire şi leme. Ulterior, după perfecţionarea instrumentelor corespunzătoare, se urmăreşte adăugarea şi a altor niveluri de analiză: sintactică, semantică etc.

Activităţile manuale se derulează prin intermediul unei interfeţe web.

Proiectul de licenţă urmăreşte parametrizarea activităţilor Portalului pentru a separa procedurile aplicabile oricărei limbi de cele specifice unei limbi anume. În felul acesta, adaptarea lui pentru o anumită limbă ar implica particularizări şi accesul la resurse specifice (lanţuri specifice de prelucrare etc.).

În esenţă se va dezvolta un *framework* de creare a corpusurilor, care poate interesa foarte mulţi cercetători preocupaţi de construirea de corpusuri pentru limbile care nu dispun încă de astfel de resurse.

2. Monitorizarea activităţilor în Portalul COROLA

Se urmăreşte monitorizarea dinamicii corpusului COROLA. În particular, complexitatea deosebită a acestui corpus, datorată sutelor de mii ori milioanelor de fişiere de text şi voce care vor trebui să-l compună şi a adnotărilor adăugate lor, parțial automat, parțial manual, precum şi dinamica achiziționării lui, care se va desfăşura pe parcursul a mai multor ani, impune o monitorizare atentă a evoluției lui. IIT a elaborat deja (Cristea et al., 2014) un graf de dependențe între schemele de adnotare care să facă posibilă asignarea unui unic identificator (corespunzător nodului din graf ce caracterizează adnotarea pe care o include) fiecărui fişier aflat într-o anumită fază de adnotare. Simultan, se urmărește ca fişierele să aibă asociate în

partea lor de metadate, între altele, şi schema de codificare, tipul adnotărilor (manuală sau automată), instrumentele de adnotare folosite, versiunile lor etc. Existența grafului de dependențe între diferitele scheme de adnotare permite regăsirea şi complementarea informației din metadate referitoare la adnotarea unui anumit text. Va exista o legătură directă între metadate, graful de dependențe şi schemele de adnotare. Această viziune va face posibilă:

1) identificarea adnotărilor contribuite de experți manual față de echivalentul lor inclus automat de tehnologie;

2) urmărirea lanțului de adnotări adăugate unui fişier, pe măsura rulării acestuia şi continuarea în cazul întreruperilor;

3) calcularea minimului de elemente *stand-off* necesar de refăcut în cazul înlocuirii unui modul cu altul perfecționat;

4) elaborarea unor indicatori globali şi unor statistici care să oglindească stadiul evoluției corpusului,

5) definirea restricțiilor asupra interogărilor KWIC (*key word in context*) posibil de adresat fiecărui fişier în funcție de adnotările pe care le include etc.

3. Formularea interogărilor GGS în limbaj natural

În cotutelă cu drd. Radu Simionescu (radsimu@gmail.com)

GGS-3 (*Graphical Grammar Studio*) este un sistem cadru de dezvoltare a grafurilor de tip reţele tranzitive recursive (*recursive transitive networks* - RTN), prin care se pot descrie reguli simbolice de prelucrare a limbajului natural. Puterea lor de prelucrare depăşeşte cu mult cea a expresiilor regulate, pentru că limbajul GGS-urilor acceptă definirea de variabile pentru memorarea temporară de valori, parsarea în avans sau retroactivă fără consumarea tokenilor de intrare, definirea de macro-uri etc. GGS-3 a fost aplicat până acum cu succes la definirea de reguli pentru corectarea erorilor adnotărilor la parte de vorbire lăsate de un POS-tagger statistic, la antrenarea de parsere sintactice şi la generarea de gramatici sintactice (Simionescu, 2014, 2015, 2016).

Datorită flexibilităţii şi puterii lui de exprimare, se speră ca GGS-3 să poată fi utilizat la formularea de constrângeri pentru regăsirea de contexte de apariţie ale cuvintelor în corpusul COROLA (v. mai sus).

Proiectul urmăreşte realizarea unei interfeţe în limbaj natural care să preia exprimări simple de condiţii şi să genereze secvenţe GGS-3, care, adresate corpusului COROLA (sau oricărui altui corpus care respectă acelaşi format), să genereze liste de contexte (co-ocurenţe) care respectă condiţiile.

Exemple de astfel de interogări KWIC (*key word in context*):

- *Vreau ocurențele lemei 'mişca' (vb).* => se va genera o listă de ocurenţe centrate pe formele verbului 'a mişca', cu contexte stânga-dreapta care includ, implicit, doar fraza în care apare verbul;

- *'sufla' urmat de 'în' (contexte dreapta în lungime de 5 cuvinte)* => se va genera o lista de ocurenţe care încep cu o formă a verbului 'a sufla' şi continuă de prepoziția 'în' şi cel mult încă 3 cuvinte;
- *exemple ale verbului 'a duce' în care apare un complement circumstanţial instrumental* => se va genera o listă de contexte frazale în care verbul 'a duce' este urmat de un complement circumstanţial instrumental (ex: "Ieri, Ion a fost atât de amabil să mă ducă cu maşina la gară.").

4. Inferarea gramaticii unei limbi
Proiect propus de drd. Radu Simionescu (radsimu@gmail.com) - luați legătura direct cu dânsul.

5. Recunoaşterea relaţiilor semantice în texte
Ariton Andrei - aryton_andrey19@yahoo.com, andrei.ariton@info.uaic.ro - pt relaţii de rudenie
Andreia Băiceanu - andreia_baiceanu@yahoo.com, mihaela.baiceanu@info.uaic.ro - pt relaţii afective şi sociale

Tema urmăreşte antrenarea unui program care să recunoască diferite tipuri de relaţii semantice care sunt exprimate într-un text liber între menţiuni de entităţi. O entitate poate fi: **persoană** sau **parte fizică a unei persoane**, **instituţie**, **locaţie geografică** etc. Exemple de menţiuni de entităţi de tip **persoană**: *Ion*, *el*, *bărbatul cu pălărie de soare*, *sa*; **părţi fizice de persoane**: *mâna sa dreaptă*, *ochii*, *piciorului* etc.; **instituţii**: *Universitatea "Alexandru Ioan Cuza"*, *Primărie*, *PNL* etc.; locații geografice: *Iași*, *Bulevardul Copou*, *Palas Mall* etc.

Iată şi câteva exemple de relaţii:
- **referenţialitate** de tip **part-of**: X part-of Y dacă X este o parte fizică a persoanei Y
1:[*Ion*] *şi-a acoperit* 2:[*ochii*] *cu* 3:[*mâna*]. => [2] part-of [1], [3] part-of [1]
- diferite tipuri de rudenie (**kinship**): X **daughter** Y dacă X este fiica lui Y, X **parent** Y dacă X este părinte pentru Y etc.
1:[*fiica lui* 2:[*Zamolxis*]] => [1] daughter-of [2]
Când 1:[*ți*]*-ai vizitat* 2:[*părinții*] *ultima oară?* => [2] parent [1]
- relaţii spaţiale:
Când urci 1:[*Bulevardul Copou*]*, pe stânga ai să întâlnești clădirea* 2:[*Universității*]*.* => stabiliţi voi ce relaţie se poate afirma între entităţile [1] şi [2]?...

În antrenare se vor folosi două corpusuri care conţin notaţii ale entităţilor şi relaţiilor, construite manual: QuoVadis şi MappingBooks, ambele amintite la curs. Sarcina voastră este de a crea patternuri de recunoaştere a relaţiilor, de a găsi trăsăturile cele mai relevante pentru recunoașterea relaţiilor şi de a antrena pachete de statistică (Weka etc.) în acest scop.

6. Deducerea abilităţilor de manevră în jocuri interactive
proiect propus de studentul Petru Manea