Nucleu de bancă de arbori sintactici pentru limba română

Proiectul „Nucleu de bancă de arbori sintactici pentru limba română” se înscrie în amplul proces de informatizare a limbii române, în contextul unui decalaj important între calitatea și anvergura tehnologiilor digitale lingvistice dedicate acestei limbi și cele ale limbilor europene care dispun de sprijin avansat în acest domeniu (cea mai avantajată în acest moment fiind limba engleză).

Nivelul de analiză sintactică automată a limbii române este una dintre direcțiile de cercetare și dezvoltare cu rezultate modeste și resurse și instrumente insuficiente. Proiectul de față își propune să contribuie la umplerea acestui gol prin crearea unei bănci de arbori sintactici, folosind ca formalism gramaticile de dependență. O bancă de arbori sintactici (engl. *treebank*) este un corpus (i.e. o colecție de texte electronice) adnotat la nivel sintactic. Fiecare propoziție analizată sintactic poate fi reprezentată grafic sub forma unui arbore: în noduri sunt cuvintele propoziției, iar arcele reprezintă relațiile sintactice dintre cuvinte.

Deoarece resursele de timp sunt limitate, se va dezvolta un treebank ale cărui dimensiuni modeste trebuie să fie compensate de calitatea acestuia. Pentru a capta cât mai multe fenomene sintactice din limba română, treebank-ul trebuie să includă propoziţii din domenii şi stiluri funcţionale diverse. De aceea, propoziţiile de adnotat au fost selectate din **ROMBAC**, un corpus românesc balansat[1] dezvoltat la ICIA [1], care cuprinde cinci secţiuni corespunzătoare la cinci domenii distincte: jurnalistic (ştiri şi editoriale), medical (scurte texte farmaceutice), juridic (texte extrase din Acquis-ul Comunitar), academic (biografii și recenzii critice ale unor autori literari), ficțiune (romane atât românești cât și traduse). Criteriul de selecție din ROMBAC a propozițiilor de adnotat a fost frecvența în corpus a verbelor ce apar în aceste propoziții: cele mai des întâlnite structuri sintactice verbale din corpusul balansat sunt implicit des întâlnite în limba română. S-au ales astfel spre adnotare 5000 de propoziții, câte 1000 din fiecare secțiune a corpusului ROMBAC.

Tot datorită constrângerilor temporale, adnotarea sintactică de la zero nu a fost fezabilă. De aceea s-a folosit o metodologie deja experimentată cu rezultate pozitive [2, 3]: adnotarea automată folosind instrumente disponibile (statistice sau bazate pe reguli) și corectarea manuală ulterioară a soluțiilor furnizate de acestea. Cum pentru limba română nu exista în momentul inițierii proiectului nici un analizor sintactic (en. *parser*) performant, am fructificat ocazia oferită de mobilitatea internațională postdoctorală și am primit sprijinul unei echipe de la Institutul de Cercetare pentru Lingvistică Aplicată

[1] Un corpus balansat (general sau specializat) acoperă un spectru larg de categorii de texte, fiecare categorie fiind reprezentată aproximativ prin același număr de cuvinte.

(IULA) din cadrul Universității Pompeu Fabra, Barcelona. Aceștia au oferit un model statistic antrenat pe un treebank de limbă spaniolă, **IULA LSP**[2] [4], pe care l-am utilizat împreună cu parserul statistic **MaltParser**[3] pentru a adnota automat resursa în dezvoltare. Anterior, echipa IULA folosise acelaşi model pentru a dezvolta un treebank pentru limba catalană, bazându-se pe similitudinea structurală a limbilor implicate. Am intuit că experimentul poate fi reprodus cu succes, din moment ce limbile română, catalană şi spaniolă, aparținând familiei limbilor romanice, împart trăsături şi şabloane sintactice. Pentru că modelul statistic este destinat unei alte limbi, era de aşteptat ca munca de corectare să fie substanțială, dar totuşi preferabilă unei adnotări exclusiv manuale. Tot la IULA am deprins utilizarea **instrumentului graphic** y**Ed**[4], care permite vizualizarea și corectarea propozițiilor adnotate într-un format arborescent, prietenos. Acest instrument lucrează cu fișiere într-un format specific, GRAPHML, de tip XML. Pentru transpunerea fișierelor adnotate din formatul produs de MaltParser (tabular, pe coloane) în formatul GRAPHML am utilizat un script Perl furnizat tot de echipa IULA. Astfel, datorită acestei colaborări, realizarea unora dintre scopurile proiectului a fost posibilă cu mai puține eforturi și într-un interval de timp redus. Întâlnirea a fost utilă și pentru echipa IULA, schimbul de experiență – în special cu privire la tipurile de erori produse de adnotatorul automat și posibilitățile de a le preveni – conducând la perspectiva publicării unui articol semnat în colaborare, care să relateze în paralel experimentele pe cele două limbi, dintr-o perspectivă a repetabilității și reproductibilității multilinguale a metodologiilor și rezultatelor cercetării.

Setul de etichete de relații de dependențe pe care l-am utilizat a fost obținut prin îmbinarea a două seturi pe care le-am avut la dispoziție: 1) setul folosit de echipa IULA, codificat în modelul statistic spaniol, a cărui prezervare este necesară pentru a putea valorifica adnotarea automată și 2) setul folosit de inițiativa de standardizare Universal Dependencies[5], a cărui integrare în resursa noastră facilitează utilizarea ei în proiecte internaționale viitoare. La acestea am adăugat și câteva relații noi, corespunzând unor fenomene sintactice specifice limbii române.

Pentru a menține consistența adnotării, s-a pornit, într-o primă etapă, cu prima jumătate (2500 de propoziții) a corpusului de adnotat în care au fost incluse în special propozițiile de lungime cuprinsă între 10 și 30 de cuvinte, iar propozițiile mai lungi, și implicit mai complexe sintactic, au fost lăsate pentru adnotare și corectare într-o etapă viitoare.

Am început adnotarea automată cu un set de 500 de propoziții din sub-corpusul jurnalistic folosind modelul statistic de-lexicalizat de limbă spaniolă, dar după corectarea

[2] http://www.iula.upf.edu/recurs01_tbk_uk.htm
[3] http://www.maltparser.org/
[4] http://www.yworks.com/en/products/yfiles/yed/
[5] http://universaldependencies.github.io/docs/

acestora am decis folosirea lor pentru re-antrenarea unui model lexicalizat pe limba română, intuind că modelul obținut va avea performanțe mai bune decât cel spaniol, chiar dacă este antrenat pe incomparabil mai puține propoziții: 500 versus 40.000. Am repetat procedura de re-antrenare după corectura fiecărui nou set de 500 de propoziții, adăugând de fiecare dată la corpusul de antrenare ultimele propoziții corectate. Ciclul de lucru este: 1) adnotare cu modelul statistic cel mai performant la dispoziție; 2) corectura setului de propoziții adnotat la pasul 1; 3) adăugarea setului corectat la corpusul de antrenare și re-antrenarea unui model extins, mai performant decât precedentul.

Folosind o metrică consacrată în domeniu, scorul LAS (eng. Label Atacchment Score, Scorul de Atașare Etichetată), care reprezintă raportul dintre numărul de cuvinte cu centre și etichete corect idenificate și numărul total de cuvinte din propoziție, au fost evaluate performanțele fiecărui nou model statistic antrenat, luând ca referință (engl. *gold standard*) varianta corectată a fiecărui set de propoziții și ca mulțime de test varianta ne-corectată. Creșterea performanței la trecerea de la modelul spaniol la cel românesc a fost substanțială, de la un scor LAS de 0,243 la 0,580, și a continuat până la 0,773 pentru modelul folosit în acest moment. Experiența noastră în munca de corectare manuală, vizibil ușurată pe măsură ce modelul statistic creștea, confirmă și ea evoluția performanțelor adnotatorului automat.

Metodologia aleasă pentru dezvoltarea treebank-ului de limbă română s-a dovedit inspirată, ajutând la obținerea unui număr substanțial de propoziții corectate (3.500 în acest moment) și a unui model statistic de calitate satisfăcătoare, care să garanteze că resursa va fi finalizată la timp. De asemenea, ne așteptăm ca scorul LAS să continue să crească în etapele de re-antrenare succesive viitoare, chiar dacă într-un ritm tot mai lent: performanţele oricărui instrument statistic sunt tot mai greu de îmbunătăţit când valorile măsurilor de evaluare se apropie de 1.

Această lucrare a fost realizată în cadrul proiectului "Cultura română şi modele culturale europene: cercetare, sincronizare, durabilitate", cofinanţat de Uniunea Europeană şi Guvernul României din Fondul Social European prin Programul Operaţional Sectorial Dezvoltarea Resurselor Umane 2007-2013, contractul de finanţare nr.POSDRU/159/1.5/S/136077.

Referinţe

1. ION, R., IRIMIA, E., ȘTEFĂNESCU, D., Tufiș, D., *ROMBAC: The Romanian Balanced Annotated Corpus*, Procedings of LREC 2012, Istanbul, Turkey.
2. ARIAS, B., BEL, N., FOMICHEVA, M., LARREA, I., LORENTE, M., MARIMON, M., MILA, A., VIVALDI, J., PADRO, M., *Boosting the creation of a treebank*, Proceedings of LREC 2014, Reykjavik, Iceland

3. FLOREA, I.M., REBEDEA, T., CHIRU, C.G. *Parser de dependenţe pentru limba română realizat pe baza parserelor pentru alte limbi romanice*. Revista Romana de Interactiune Om-Calculator 7(1), pp. 1-20, 2014.
4. MARIMON, M., BEL, N. *Dependency structure annotation in the IULA Spanish LSP Treebank*. Language Resources and Evaluation. Amsterdam: Springer Netherlands, 2014.

Cercetător postdoctorat,

Irimia Elena