TENDINŢE ACTUALE ÎN PROCESAREA LIMBAJULUI NATURAL PENTRU LIMBA ROMÂNĂ

Mircea PETIC, *dr, lect. sup. univ.,*
Universitatea de Stat „Alecu Russo" din Bălți

Summary: *The article presents some trends in natural language processing research topic for Romanian language such as the use of computational linguistic resources. In addition, some useful applications are demonstrated, that use computational linguistic resources in both pre-processed texts as well as unstructured texts.*

Key-words: *natural language processing, computational linguistic resources, linguistic web services.*

În zilele noastre foarte multe domenii ale activităţii umane se informatizează. Nu este o excepţie nici limbajul natural. Mai mult, este esenţial pentru supravieţuirea unei limbi ca ea să fie folosită în sistemele de informare electronică. În acest sens, tehnologiile limbajului uman devin tot mai importante în era informaţională. Limba română începe să se contureze ca una din limbile semnificative în privinţa resurselor informatice şi a tehnologiilor aplicate ei. Totuşi, chiar şi cei mai fideli dintre susţinătorii informatizării limbajului sînt de acord cu privire la imensa cantitate de necunoscut care însoţeşte acest proces. Astfel tehnologizarea unei limbi necesită consolidarea bazei teoretice din domeniul lingvisticii.

Eforturi pentru dezvoltarea resurselor lingvistice şi instrumente de procesare a limbajului natural pentru limba română sînt depuse în puține centre academice din România (în general București, Iași și Cluj-Napoca) și Republica Moldova (Chișinău), cît și în unele centre din afara acestor regiuni. Astfel cele mai importante centre de cercetare în procesarea limbajului natural sînt:

- Institutul de Matematică şi Informatică al AŞM, Chişinău;
- Universitatea Tehnică a Republicii Moldova, Chişinău;
- Universitatea „A. I. Cuza" din Iaşi;
- Institutul de Informatică Teoretică al Academiei Române (Filiala Iași);
- Institutul de Cercetări în Inteligenţă Artificială, Bucureşti.

Una din primele aplicaţii de procesare a limbajului natural a fost cea de despărțire în silabe a cuvîntului. Automatizarea procesului de divizare în silabe nu reprezintă o problemă ordinară, deoarece, la etapa de prelucrare a cuvîntului, informaţia fonetică este inaccesibilă, deşi anume această informaţie este importantă (Demidov&Verlan, 1996). Algoritmii elaboraţi şi programele corespunzătoare, dezvoltate ulterior, se bazează pe regulile clasice de divizare a cuvintelor în silabe, avînd drept reper semnificaţia fonetică a literelor. Totuşi, caracterul specific al limbii române nu permite formalizarea completă a regulilor de divizare a cuvintelor în silabe, în mare parte din cauza vocalelor. Acestea pot fi simple şi complexe (aşa-numitele semivocale), accentuate şi neaccentuate, şi regula de divizare în silabe depinde de categoria vocalei (frea-măt, re-cre-at, ghiont, ghi-o-cel etc). În acest fel, unicul lucru care poate fi stabilit în momentul în care se cunoaşte cuvîntul este consecutivitatea vocalelor şi a consoanelor. Informaţia fonetică nu este accesibilă. În această ordine de idei, este imposibilă acoperirea completă a regulilor de împărţire în silabe. Cele mai complicate situaţii apar în procesarea consecutivităţii de vocale (au, ea, ia, ie, io, ii, oa, ua), care pot fi (sea-ră, ier-ta, iod, au-gust) sau nu diftongi. Aceeaşi problemă e şi cu

procesarea triftongilor. În acest sens, algoritmul de împărţire a cuvintelor în silabe din limba română a prevăzut procesarea a unui număr destul de mare de situaţii. Chiar dacă nu se pretinde la completitudinea algoritmului, totuşi circa 70% de cuvinte din aproximativ 70000 de cuvinte din textele din domeniul ştiinţei şi artei sînt împărţite corect în silabe (Demidov&Verlan, 1996).

Un alt tip de aplicaţii de procesare a limbajului natural sînt cele de flexionare automată. Prima încercare de acoperire a morfologiei flexionare pentru limba română a fost modelarea FAVR în mediul Mac-ELU. În urma analizei atributelor specifice fiecărei părţi de vorbire în parte, în descrierea morfologică implementată în Mac-ELU s-au utilizat 20 de categorii gramaticale. Clasificarea s-a efectuat nu numai în baza cerinţelor prelucrărilor morfo-lexicale, ci şi a granularităţii necesare analizei şi, respectiv, a generării sintactice (Tufiş, 1996).

O altă aplicaţie de flexionare automată ce merită atenţie este sistemul AnMorph, descris în (Cristea&Forăscu, 2006). În prezent baza de date a aplicaţiei acoperă doar parţial morfologia limbii române. În sistemul morfologic se compară formele introduse de utilizator cu acele care pot fi generate, pornind de la o paradigmă deja existentă în baza de date a sistemului şi, dacă asemănarea este confirmată, se generează restul formelor. Dacă se întîmplă acest lucru, utilizatorul doar verifică şi validează partea tabelului generată automat.

O altă abordare poate fi găsită în metodele de flexionare propuse de S. Cojocaru, care ţine cont de clasificarea descrisă de A. Lombard şi C. Gâdei în (Lombard& Gâdei, 1981). Algoritmul utilizează o gramatică funcţională de flexionare, care formalizează procesul de realizare a alternanţelor vocalice şi consonantice şi de formare a flexiunilor (prin concatenarea temei cuvîntului de bază cu seturile de terminaţii corespunzătoare). Pentru limba română această gramatică include 866 de reguli şi 320 de seturi de terminaţii. Această metodă, numită metodă statică de flexionare, a contribuit substanţial la acumularea resurselor lingvistice, fiind aplicată pentru flexionarea a circa 30.000 de cuvinte pentru care era cunoscut grupul de flexionare din (Lombard& Gâdei, 1981). Metoda statică diferă de cea dinamică, în cadrul căreia se încearcă să se calculeze modelul de flexionare pentru fiecare cuvînt în parte. Algoritmul a fost testat pe 30.000 de cuvinte, care nu au fost incluse în clasificarea statică. De asemenea, au fost depistate unele iregularităţi (3% din numărul de cuvinte flexionate). Dincolo de aceasta s-a elaborat un set de programe care în regim semiautomat, generează toate formele flexionate. În anumite cazuri pentru obţinerea corectă a flexiunilor dorite, se solicită implicarea utilizatorului pentru selectarea modelului potrivit (Ciubotaru et al, 2007).

Aplicaţiile morfologiei flexionare descrise mai sus îmbogăţesc resursele lingvistice, dar numai cu formele acestor cuvinte deja existente, fără a obţine cuvinte cu sensuri noi, această funcţie revenindu-i procesului de derivare pentru care în baza regulilor, constrîngerilor, schimbului de afixe, precum şi a metodelor de proiectare a derivatelor a fost elaborat un algoritm de derivare automată care, fiind aplicat asupra a trei resurse lingvistice computaţionale preprocesate, a produs 11191 cuvinte derivate noi. Aceste cuvinte, fiind ulterior flexionate, au completat dicţionarul RRTLN (*Resurse reutilizabile ale tehnologiei limbajului natural*) cu 123106 intrări noi (Fig. 1) (Petic, 2011).

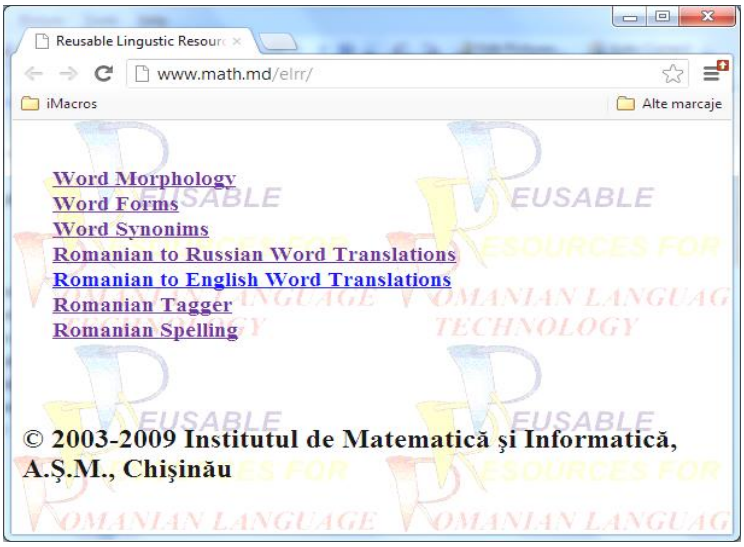

Figura 1. *Interfaţa Web a RRTLN*

RRTLN – conţine o bază de date cu informaţie lingvistică la nivel de cuvînt şi un set de programe de gestionare (Boian et al, 2005). Astfel, lexiconul conţine nu doar reprezentarea grafică a cuvîntului, dar şi informaţia despre categoriile morfologice ale lui şi posibilele funcţii sintactice. RRTLN are aproximativ 100.000 de cuvinte leme şi circa 1.000.000 de flexiuni. De menţionat, că un cuvînt poate avea mai multe intrări pentru diferite părţi de vorbire, astfel avînd o altă semantică, de exemplu, *bun* ca adjectiv în sens de binevoitor, amabil, *bun* ca adverb reprezentînd o aprobare şi *bun* ca substantiv în sens de bunic sau de proprietate. Alături de informaţia lingvistică pentru limba română, complexul RRTLN conţine şi informaţii cu privire la sinonime şi traduceri (engleză, rusă). Un set de programe de gestionare a bazei de date (Colesnicov, 2010) asigură efectuarea operaţiilor necesare. Popularea BD cu informaţie morfologică se efectuează cu ajutorul unui program, care în baza lemei indicate, permite analiza şi generarea automată a circa 90% din numărul total al cuvintelor lemă (Boian et al, 2010), celelalte 10% prezentînd iregularităţi.

Rezultatele de mai sus au permis elaborarea corectorului ortografic pentru limba română care a fost elaborat şi integrat în editorul de text Microsoft Word. În plus, în colaborare cu grupul editorial LITERA un CD cu corectorul ortografic cu o baza lingvistică de 1 milion de cuvinte a fost publicată.

O altă tendinţă actual de cercetare în procesare a limbajului natural constă în procesarea corpusurilor lingvistice computaţionale. În lucrarea V. Bobicev a fost descris modul de procesare a mai multor corpusurilor cu ajutorul unor metode statistice. În experimentele cu restabilirea semnelor diacritice în text au fost obţinute rezultate foarte bune de restabilire, aproximativ 98%. În experimentele de clasificare a documentelor aproximativ 97% de documente au fost clasificate corect (68 din 70), ceea ce poate fi considerat un rezultat destul de reuşit. În experimentul de dezambiguizare morfologică programul a dezambiguizat absolut corect 95,3% de cuvinte. Din cele 4,7% de dezambiguizare greşită, numai în 2,5% a fost comisă o eroare în partea de vorbire. În restul cazurilor, 2,2%, a fost greşit detectat numai o caracteristică morfologică din cod. În concluzie experimentele efectuate demonstrează abilita-

tea metodelor statistice de a rezolva o clasă mare de probleme practice de prelucrare a textului (Bobicev, 2007).

În teza de doctorat a Victoriei Maxim este tratat un alt subiect promițător pentru cercetare ce ține de analiza computaţională a corespondenţei de afaceri. Drept rezultat au fost create următoarele resurse lingvistice computaţionale: corpusul paralel bilingv englez-român cu texte ale scrisorilor de afaceri, adnotarea lui automată şi dezambiguizarea manuală și o ontologie sumară a conceptelor de business. În plus s-a reuşit formalizarea sintactică a Grupurilor Nominale ce conţin termeni și extragerea din corpus a unităţilor terminologice, validarea lor şi organizarea într-un dicţionar englez-român de termeni de afaceri. Ca rezultat a fost elaborat un program de traducere asistată de calculator a textelor comerciale din limba engleză în cea română (Maxim, 2008).

Un alt domeniu care se dezvoltă actualmente în procesarea limbajului natural ține de dezvoltarea serviciilor Web lingvistice. În continuare vor fi descrise succint posibilitățile http://www.racai.ro/ webservices/. De menționat că toate aplicațiile care vor fi pomenite există ca aplicații de sine stătătoare (Tufiș et al, 2007).

Ceea ce ține de identificarea limbii se poate de spus că acest serviciu asigură identificare a limbii unui text scris într-una dintre cele 22 de limbi ale Uniunii Europene. Textul ar trebui să conțină un număr minim de 10-15 cuvinte.

Adnotarea morfo-lexicală se face cu o mare acuratețe (în jur de 98%) atît pentru limba română cît și pentru limba engleză. Există două implementări diferite pentru sarcina de adnotare morfo-lexicală: una se bazează pe paradigm HMM (Hidden Markov Models – Modele Markov Ascunse) iar cealaltă folosește abordarea Maximum Entropy (Entropie Maximă).

Un alt serviciu garantează codificarea de corpusuri paralele în format XCES pornind de la texte neprelucrate. O altă posibilitate a serviciilor este de a vizualiza cu ajutorul grafurilor hiporbolice la elementele celei mai mari ontologii lexical pentru limba română Ro-WordNet.

DIAC$^+$ este un serviciu care permite recuperarea automată a diacriticilor în texte în limba română scrise fără – sau scrise doar parțial – cu caractere diacritice. DIAC$^+$ este de asemenea disponibil și ca o aplicație de sine stătătoare, în forma unui DLL pentru MSOffice.

Unul din cele mai importante centre universitare care se ocupă cu succes de cercetarea în domeniul de procesare a limbajului natural este Universitatea „Alexandru Ioan Cuza” din Iași. Unele din direcţiile prioritare sînt:

- Parsarea articolelor din DEX pentru obţinerea arborelui lexico-semantic al unei intrări de dicţionar într-un fişier XML care reprezintă codificarea definiţiilor din intrarea respectivă;
- Realizarea inferenţelor textuale pe limba română presupune ca în baza a două fragmente de text se cere precizarea dacă înțelesul unuia din texte poate fi dedus din celălalt.
- Extragerea automată a definiţiilor din texte este o parte a unui sistem complex de întrebare răspuns, pentru regăsirea răspunsurilor la întrebări de tip definiție.

Elaborarea și dezvoltarea unui sistem de întrebare răspuns este una din cele mai tentante sarcini în cercetarea din domeniul procesării limbajului natural. Primul sistem de întrebare răspuns românesc a fost dezvoltat în anii 80 și era reprezentat de o

interfață ce facilita comunicarea cu o rețea semantic (care codifica cunoaşterea). Astăzi astfel de sisteme folosesc documentele text ca baza de cunoaştere şi integrează tehnici de prelucrare a limbajului natural pentru a găsi răspunsul la o întrebare pusă în limbajul natural. Arhitectura general a unui sistem întrebare răspuns constă din: o colecţiei de texte din Wikipedia (XML, HTML), componentele de filtrarea documentului (html→txt), pre-procesarea lingvistică (tokenizare, lematizare, etc.), analiza întrebării, crearea indexului şi căutarea documentară, extragerea răspunsului şi analiza rezultatului (Iftene et al, 2007).

Din cele menţionate în articol constatăm că limba română începe să se contureze ca una din limbile semnificative în privinţa resurselor informatice şi a tehnologiilor aplicate ei. Totuşi, chiar şi cei mai fideli dintre susţinătorii informatizării limbajului natural sînt de acord că acest proces este încă supus hazardului. De aceea sînt deosebit de importante cercetările şi elaborările, care permit să înaintăm cu anumiţi paşi concreţi în acest domeniu.

Referinţe bibliografice

1. V. Demidov T. Verlan, An approach to the word division into syllables for Romanian lan-guage, Computer Science Journal of Moldova, v.4, n.1 (10), 1996, pp. 59-68.
2. Tufiş D. ş. a. Morfologia limbii române, o resursă lingvistică reversibilă şi reutilizabilă. În: Limbaj şi Tehnologie. Bucureşti: Editura Academiei, 1996, p. 59-65.
3. Cristea D., Forăscu C. Linguistic Resources and Technologies for Romanian Language. In: Computer Science Journal of Moldova, Vol. 14, nr. 1 (40), 2006, p. 34-73.
4. Lombard A., Gâdei C. (1981) *Dictionnaire morphologique de la langue roumain*, Bucureşti, Editura Academiei, 232 p.
5. Ciubotaru C. ş. a. Contribuții la proiectul RoLTech: Platforma pentru tehnologia limbii române: resurse, instrumente, interfețe. În: Lucrările atelierului „Resurse Lingvistice şi Instrumente Pentru Prelucrarea Limbii Române. Iaşi: Editura Universităţii, 2007, p. 171-179.
6. Petic M. Lexical derivation approaches for functional extention of computational linguistic resources. In: Proceedings of the 8th International Conference "Linguistic Resources and Tools for processing of the Romanian language" 8-9 december 2011, 26-27 april 2012. Bucharest, Editura Universitatii "Alexandru Ioan Cuza" Iasi, pp. 29-38
7. Boian E. ş. a. Technologization of Romanian: linguistic resources, applications, tools. In: Proceedings of the 4th International Conference on Microelectronics and Computer Science. Chisinau, Vol. II, 2005, p. 519-522.
8. Colesnicov A. Selection of processing tools in digital research environments. In: Proceedings of „Actual problems of mathematics and informatics", UST – 80 ani. Chişinău, 2010, p. 88-89.
9. Boian E. ş. a. Application based on Reusable Linguistic Resources. In: Multilinguality and Interoperability in Language Processing with Emphasis on Romanian. Bucureşti: Editura Academiei, 2010, p. 461-475.
10. Bobicev V. Metode și algoritmi statistici de procesare a textelor (în baza textelor în limba română). Teză de dr. în informatică, Chişinău, 2007. 229 p.
11. Maxim V. Analiza computaţională a corespondenţei de afaceri. Teză de dr. în informatică, Chişinău, 2008. 220 p.
12. Tufiş D. ş. a. Servicii web lingvistice ale ICIA. În: Lucrările atelierului „Resurse Lingvistice şi Instrumente Pentru Prelucrarea Limbii Române. Iaşi: Editura Universităţii, 2007, p. 61-68.
13. Iftene A. ş. a. Construirea unui sistem de Întrebare Răspuns pentru limba română. În: Lucrările atelierului „Resurse Lingvistice şi Instrumente Pentru Prelucrarea Limbii Române. Iaşi: Editura Universităţii, 2007, p. 109-118.