Revista Română de Interacţiune Om-Calculator 6 (2) 2013, 169-192

Identificarea entităţilor, citatelor şi evenimentelor în ştiri şi texte din Web-ul social în limba română

Adrian-Nicolae Zamfirescu[1,2], Traian Eugen Rebedea[1,2]

[1]Universitatea Politehnica din Bucureşti, Facultatea de Automatică şi Calculatoare, Splaiul Independenţei, Nr. 313, 060042, Bucureşti, România

[2]TeamNet International, Splaiul Independenţei, Nr. 319, 060044, Bucureşti, România
E-mail: zamfirescuan@yahoo.com, traian.rebedea@cs.pub.ro

Rezumat. În cadrul prelucrării limbajului natural, detectarea automată a entităţilor cu nume a reprezentat una dintre cele mai importante provocări, care încă nu a fost rezolvată perfect până în acest moment nici pentru toate tipurile de texte scrise în limba engleză. Mai mult, identificarea entităţilor a deschis calea rezolvării unor alte probleme în care sunt implicate aceste construcţii lingvistice, precum identificarea citatelor şi a declaraţiilor făcute în general de persoane, dar şi de către companii sau alte tipuri de organizaţii, sau a extragerii evenimentelor. Problema identificării şi a clasificării automate a entităţilor a apărut din necesitatea de a putea observa o evoluţie a statisticilor în contextul diverselor informaţii din textele scrise, raportate la anumite persoane publice, organizaţii sau alte tipuri de entităţi cu nume care prezintă interes în diverse domenii. În articolul de faţă vom încerca rezolvarea problemelor menţionate anterior pentru texte scrise în limba română şi provenind din diverse surse online, precum ştiri, articole de pe bloguri sau comentarii din reţele sociale. Întâi, vom face o scurtă trecere prin fundamentele teoretice pentru rezolvarea acestor probleme, iar apoi vom prezenta metodele de rezolvare ale problemelor bazate pe aplicarea unor algoritmi de clasificare automată, combinaţi cu euristici bazate pe reguli şi expresii regulate. În final, vom prezenta rezultatele şi eficienţa diverselor metode utilizate, o comparaţie între acestea, precum şi concluziile referitoare la problemele abordate.

Cuvinte cheie: prelucrarea limbajului natural, extragerea informaţiilor, clasificare automată, entităţi cu nume, citate, evenimente

1. Introducere

În multe aplicaţii de analiză automată a textelor, identificarea entităţilor cu nume reprezintă un prim pas esenţial în cadrul procesului de analiză întrucât extrage o parte din elementele reprezentative din cadrul textului analizat (Jurafsky şi Martin, 2008). Acest articol descrie realizarea unor componente pentru detectarea automată a următoarelor elemente din cadrul articolelor

de ştiri, a textelor publicate pe bloguri, extrase din rețele sociale sau a comentariilor publicate pe diverse situri web în limba română:

- **Entitățile cu nume** care apar în aceste texte, prin identificarea numelor acestora, dar şi prin clasificarea acestora în câteva categorii predefinite.
- **Evenimentele** descrise de texte în care aceste entități sunt implicate, cu preponderență persoanele.
- **Citate** extrase din declarațiile persoanelor, atât exprimate ca citate, interviuri sau alte forme de vorbire directă în texte, cât şi cele relatate sub formă de vorbire indirectă.

Problemele menționate anterior se încadrează în domeniul Prelucrării Limbajului Natural (PLN) şi au fost rezolvate cu tehnici specifice PLN şi extragerii de informații (*information extraction*) din texte, care vor fi detaliate în cadrul articolului. PLN este un domeniu de cercetare multi-disciplinar, care combină elemente din: inteligență artificială, învățare automată, lingvistică, statistică, antropologie şi altele. Rezolvarea problemelor principale studiate în cadrul PLN, care au drept obiectiv final înțelegerea limbajului natural, pot influența multe aspecte din interacțiunea om-calculator. În mod ideal, aceasta presupune înțelegerea textului de către calculator aşa cum este perceput de către mintea umană (sau într-un mod asemănător celui în care îl înțelege aceasta), preprocesarea lui şi translatarea sa într-un spațiu de reprezentare care să permită unor algoritmi şi euristici specifici să extragă informația dorită pentru a fi folosită de către alte aplicații.

Problema identificării şi a clasificării entităților cu nume are ca punct de pornire tocmai această punte de trecere între limbajul natural, inteligibil pentru om şi o formă a sa modificată, propice pentru observarea anumitor şabloane care să intuiască sau să învețe tehnici de identificare sau clasificare în cadrul textelor. În cadrul lucrării de față, metodele de clasificare a entităților se axează pe distribuirea acestora în trei categorii: **persoane**, **organizații** şi **teritorii**. Acestea sunt cele mai importante categorii folosite de către majoritatea sistemelor de clasificare a entităților cu nume (Finkel, Grenager şi Manning, 2005; Bird, 2006), dar la care se pot adăuga şi alte entități generale (precum data calendaristică, valori monetare etc.) sau specifice (de exemplu, nume de gene sau proteine, nume de proiecte de cercetare etc.).

Odată extrase şi clasificate entitățile dintr-un text, putem să descoperim alte elemente utile care pot fi aflate despre entitatea respectivă. Cum o entitate cu nume apare într-un anumit context, acesta poate fi expresia unui fapt important consumat în timp şi spațiu, adică un **eveniment**. Aşadar, o entitate sau un grup de mai multe entități identificate pot conduce în continuare la detectarea unui eveniment în care acestea sunt implicate. În acest articol, ne vom baza preponderent pe evenimentele în care sunt implicate persoane, atunci când acestea sunt localizate prin cel puțin o referință temporală. Definiția aceasta este un pic diferită față de abordarea uzuală a detectării evenimentelor în sistemele de extragere a informațiilor, însă în contextul aplicației dezvoltate sunt importante, în special, evenimentele în care sunt implicate entitățile cu nume. Mai mult, şi în sistemele uzuale pentru detectării eventimentelor se poate observa că majoritatea acestora sunt legate de cel puțin o entitate cu nume (Yang, Pierce şi Carbonell, 1998).

Cum scopul final al aplicației din care fac parte modulele prezentate în cadrul acestei lucrări îl reprezintă monitorizarea evoluției unei entități cu nume în textele publicate online în România, probabil categoria cea mai importantă de entități este aceea a persoanelor, întrucât acestea sunt cele mai frecvente în cadrul textelor analizate. Cum unul dintre scopurile lucrării este de a identifica mențiunile persoanelor în cadrul articolelor de ziare online sau al postărilor pe bloguri sau situri de socializare precum Facebook sau Twitter, un element interesant în informațiile prezentate despre o entitate îl pot reprezenta **citatele**, precum şi alte tipuri de declarații sau de vorbire indirectă prezentă în textele analizate şi care, momentan, pot fi legate automat de către o entitate de tip persoană.

Astfel, un al treilea obiectiv al lucrării, pe lângă identificarea entităților şi a evenimentelor în care sunt implicate acestea, îl reprezintă extragerea declarațiilor şi citatelor. Ele se pot găsi la nivel textual atât în vorbirea directă – prin citate din declarațiile entităților respective sau din interviuri, cât şi în cea indirectă – prin relatarea autorului de articol cu privire la ce a spus o anumită persoană publică.

Un aspect important care trebuie menționat este faptul că pentru rezolvarea acestor probleme au fost folosite anumite instrumente (biblioteci şi date) disponibile liber (*open-source*, respectiv *open-data*). Pentru

modulele de învățare statistică pentru clasificarea entităților cu nume au fost folosite două biblioteci (care vin cu aplicații suport) create pentru învățare automată – Weka (Hall et al., 2009) sau PLN – Mallet (McCallum, 2002). Pentru elementele de analiză lexicală şi morfologică a textelor în limba română s-a pornit de la baza de date cu toate cuvintele şi formele flexionate existente în principalele dicționare publicate pentru română şi care este pusă la dispoziție de către DEX Online (iar datele pot fi descărcate gratuit de la http://www.dexonline.ro).

În continuare, acest articol va continua cu o scurtă prezentare a cercetărilor anterioare în problematica abordată, prezentând rezultate obținute pentru limba engleză. În cadrul secțiunii 3, este descrisă arhitectura aplicației care rezolvă aceste probleme şi sunt detaliate tehnologiile şi algoritmii folosiți pentru rezolvarea acestora. Secțiunea 4 continuă cu evidențierea rezultatelor obținute pentru fiecare problemă în parte, fiind prezentată şi o comparație între abordările utilizate pentru clasificarea entităților cu nume. Lucrarea se termină cu nişte observații relevante despre problematica abordată şi cu concluzii despre proiectul realizat.

2. Cercetări anterioare

O entitate cu nume (sau entitate numită) reprezintă o secvență de cuvinte care exprimă elemente din lumea reală şi care de obicei pot fi organizate în categorii bine definite (Aggarwal şi Zhai, 2012). Problema recunoaşterii entităților cu nume (REN) este parte din domeniul care se numeşte extragerea informațiilor (*information extraction*), care foloseşte atât elemente de PLN, dar şi din regăsirea informațiilor. În cadrul REN, se doreşte localizarea în texte a unor expresii specifice şi clasificarea acestora în categorii predefinite, precum nume de persoane, organizații, teritorii, expresii ale timpului, ale cantităților, valori monetare, procente ş.a. (Aggarwal şi Zhai, 2012).

Deşi studiul REN în texte a început dinainte de 1990, problema a fost conceptualizată formal în cadrul celei de-a şasea conferințe de înțelegere a mesajelor (*Message Understanding Conference*, *MUC-6*, http://cs.nyu.edu/faculty/grishman/muc6.html), din 1995, ca o subproblemă din cadrul domeniului extragerii informațiilor.

Primele soluții aduse pentru problema REN s-au bazat pe aplicarea unor şabloane, reguli sau automate finite, în general create manual (Grishman,

1995). Din cauză că această abordare presupunea expertiză umană pentru elaborarea şabloanelor, precum şi din cauza faptului că şabloanele create manual nu puteau acoperi toate cazurile de entități prezente în corpusurile mai mari, sistemele ulterioare au încercat să învețe automat aceste şabloane din corpusuri adnotate, folosind diverse tipuri de reguli, transductoare sau automate finite (Mikheev, Moens şi Grover, 1999). Totuşi, cele mai recente studii şi aplicații în domeniul REN se bazează pe metode statistice de învățare automată. Una dintre primele astfel de aplicații s-a numit Nymble şi descoperă entitățile cu nume folosind modele Markov ascunse (HMM) (Bikel et al., 1997). Alte modele pentru clasificarea automată utilizate cu succes în cadrul REN sunt Entropie Maximă, SVM (*Support Vector Machine*) sau CRF (*Conditional Random Fields*) (Aggarwal şi Zhai, 2012).

Interesul în detectarea automată a citatelor şi a vorbirii indirecte din textele scrise a urmat ca o consecință logică a descoperirii entităților cu nume. Similar tehnicilor de recunoaştere a entităților, şi această problemă se reduce la două abordări:

- Folosirea expresiilor regulate şi a altor tipuri de modele pentru reprezentarea şabloanelor
- Învățarea automată folosind diverse modele statistice

Prima abordare este una mai simplă. Dezvoltatorul trebuie să definească un set mai mult sau mai puțin complex de expresii regulate, care, aplicate asupra textului procesat, să realizeze o potrivire (*matching*) prin care să se extragă structurile reprezentative (vorbitor, text citat sau declarat, data etc.). Ca şi în cazul REN, există şi aici alternativa identificării automate a acestor şabloane (Krestel, Bergler şi Witte, 2008).

Alternativa la utilizarea expresiilor regulate este antrenarea unui clasificator pe un corpus robust, adnotat manual cu declarații şi vorbire indirectă. În SUA, în timpul alegerilor din 2012 a fost dezvoltat proiectul *Politics Verbatim* pentru detectarea automată a declarațiilor politicienilor în texte, iar modelul Entropie Maximă a fost tehnologia aleasă pentru a rezolva această problemă (Davies, 2012). Acest model presupune definirea unor proprietăți (*feature*s) care să permită calculatorului definirea unor elemente cheie pentru a determina dacă o structură de text este citat sau nu.

Un alt domeniu de interes din extragerea informației din texte îl reprezintă detectarea evenimentelor. Ceea ce defineşte un eveniment

depinde în mod direct de contextul în care acesta se află şi din care se doreşte a fi extras. Declerck (2005) afirmă că existența unui eveniment presupune inițial existența unor entități şi a unei relații între ele, evenimentul fiind declanşat de o schimbare de stare survenită în contextul curent. Pornind de la aceste idei, în cadrul lucrării am considerat că un eveniment relevant în ştiri sau texte din rețele sociale pentru o entitate trebuie să fie legat de către o dată când acesta a avut loc şi, eventual, de către o locație. Aplicațiile practice realizate pentru recunoaşterea evenimentelor sunt direcționate de regulă înspre un anumit domeniu specific. Totuși, majoritatea au fost dezvoltate pentru texte din rețele sociale, în special Twitter, şi folosesc metode specifice regăsirii informației, precum extragerea termenilor cei mai relevanți şi identificarea entităților în texte, împreună cu clustering sau învățare supervizată pentru a detecta şi grupa evenimentele în sine (Sayyadi, Hurst şi Maykov, 2009).

3. Descrierea implementării – algoritmi şi euristici pentru rezolvarea problemelor

Sistemul dezvoltat pentru detecția şi clasificarea entităților numite a fost conceput pentru o aplicație de monitorizare a publicațiilor online în limba română, care analizează texte provenite dintr-o gamă variată de surse. Spre deosebire de majoritatea sistemelor de REN dezvoltate şi evaluate în studiile anterioare, în cadrul acestui lucrări am folosit o abordare mai pragmatică folosind o metodă semi-supervizată, în locul celei supervizate clasice.

Astfel, analiza porneşte de la un corpus mare de texte care este folosit pentru a detecta o primă listă de entități cu nume potențiale (candidate). Dintre acestea, sunt păstrate inițial doar acele entități potențiale care pot fi găsite în versiunea în limba română a enciclopediei Wikipedia. În acest al doilea pas, fiecare entitate candidat este automat clasificată într-una din cele trei categorii de interes. Abia după aceea, putem construi un clasificator care va fi folosit pentru etichetarea automată a restului entităților care nu sunt regăsite în Wikipedia. În plus, după ce sunt cunoscute anumite entități, acestea sunt folosite pentru extragerea citatelor şi a declarațiilor rostite de către aceste entități sau a evenimentelor în care sunt implicate. Întregul proces este reprezentat grafic în Figura 1.

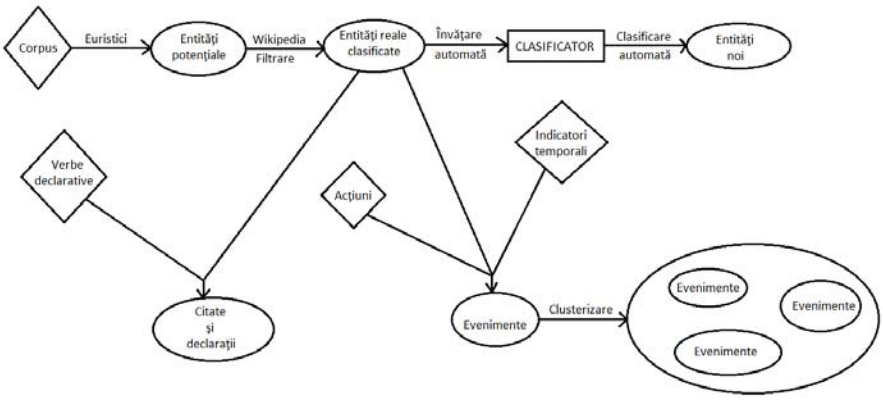

Figura 1. Diagrama funcțională a sistemului implementat pentru REN şi extragerea citatelor şi a evenimentelor

3.1 Detectarea şi clasificarea entităților

În secțiunea curentă vom descrie paşii necesari pentru a identifica şi clasifica entitățile cu nume extrase din textele apărute în articole online, în textele de bloguri sau în conținutul siturilor de socializare precum Facebook sau Twitter în limba română. În total, corpusul analizat cuprinde peste 200.000 de texte provenite din aceste surse, publicate în special în perioada iulie-decembrie 2012.

Aşa cum am precizat anterior, pentru învățarea automată a şabloanelor sau a regulilor folosite pentru REN este necesar un corpus bogat, din care să se poată extrage suficiente informații. În domeniul prelucrării informației textuale, acest corpus este reprezentat de o colecție de texte similare cu cele care urmează a fi procesate pentru se extrage structurile dorite.

Lucrarea curentă urmăreşte detectarea şi clasificarea entităților cu nume din articolele online de ştiri şi clasificarea lor în 3 categorii: persoane, organizații şi teritorii. După aceea, se construiesc liste cu aceste entități pentru a putea fi identificate şi în alte tipuri de texte, precum cele preluate din rețele sociale.

Ideea premergătoare extragerii listei potențialelor entități a fost aceea a obținerii unui corpus suficient de robust astfel încât majoritatea entităților

cu nume care apar în acest corpus să aibă o frecvență de apariție suficient de mare, astfel încât identificarea acestora să poată fi făcută cu un nivel ridicat de încredere.

Colectarea corpusului a fost realizată cu ajutorul crawler-ului Apache Nutch (http://nutch.apache.org/), care a preluat articole de pe câteva sute de situri de ştiri şi bloguri din România. După ce articolele au fost parcurse şi analizate (parsate) şi informațiile utile (precum titlul, data şi textul articolului) au fost extrase, acestea au fost salvate şi indexate folosind Apache Solr (http://lucene.apache.org/solr/) pentru a putea fi accesate rapid, inclusiv în cadrul procesului de determinare a entităților. Acest proces trece prin următoarele etape, descrise în continuare în cadrul acestei secțiuni:

1. Extragerea potențialelor entități folosind euristici
2. Filtrarea potențialelor entităților cu ajutorul Wikipedia
3. Clasificarea automată a entităților
4. Utilizarea clasificatorului pentru entitățile neconfirmate

Extragerea potențialelor entități folosind euristici

În cadrul primei etape în procesul de REN are loc detectarea şi memorarea tuturor cuvintelor care încep cu majusculă. Folosind un corpus alcătuit din articole de pe situri de ştiri şi bloguri, au fost de aşteptat imperfecțiuni în redactarea lor. Deşi majoritatea articolelor au şi secțiuni de comentarii, iar identificarea se doreşte atât la nivel de articol, cât şi la nivel de utilizator care comentează articolul, extragerea inițială este realizată doar din corpul efectiv al articolului. Motivul principal al acestei alegeri este că siturile de ştiri şi blogurile respectă, de obicei, normele de scriere a numelor proprii cu majuscule, spre deosebire de textele din comentarii sau din rețele sociale.

Aşadar, pentru fiecare text de articol analizat, acesta este supus unui proces de *token*-izare, prin împărțirea sa în cuvinte şi având ca delimitatori spațiile şi semnele de punctuație dintre ele. Dacă un cuvânt începe cu literă mare, este reținut pentru a fi inclus în grupul potențialelor entități. Dacă următorul cuvânt care urmează după acesta începe tot cu literă mare şi între cele două nu se află niciun semn de punctuație precum “.”, “,”, “?”, “!”, “;”, atunci al doilea se alătură primului şi tot aşa până când se încheie fraza sau a fost întâlnit un cuvânt scris cu literă mică care să închidă grupul curent

format. În acest fel se construieşte o listă de *potențiale entități cu nume* din articolele analizate.

Astfel, fiecare grup de cuvinte consecutive, începând cu majuscule şi despărțite exclusiv prin spații, formează un nou candidat pentru a fi o entitate cu nume. Există totuşi o excepție în regula formării acestor grupuri: dacă între 2 cuvinte consecutive începând cu majusculă apar anumite cuvinte speciale (precum unele articole, prepoziții, conjuncții etc.), atunci se vor include şi acestea în structura formată. Un exemplu elocvent pentru acest caz este entitatea cu nume *Cetatea de Scaun a Sucevei*.

Tot în cadrul acestei etape, este realizată prima filtrare care să ajute la restrângerea ariei de cuvinte care denotă în mod eronat o entitate cu nume: toate cuvintele începând cu majusculă, care apar la început de frază, sunt supuse unui verificări a frecvențelor de apariție în 3 contexte diferite:

- Număr de apariții ale acestor cuvinte la început de frază (Na_îf)
- Număr de apariții ale acestor cuvinte, scrise cu majusculă, în interiorul unei fraze (Na_ifM)
- Număr de apariții ale acestor cuvinte, dar începând cu literă mică, în interiorul unei fraze (Na_ifm)

Fiecare astfel de frecvență oferă o estimare asupra încrederii că un cuvânt sau grup de cuvinte reprezintă într-adevăr un nume de entitate. Spre exemplu, un nume de entitate nu ar trebui să apară, în mod normal, în mijlocul unei fraze, scris cu minusculă. Există totuşi situații când anumite cuvinte crează ambiguitate datorită polisemiei. Un bun exemplu ar fi cuvântul *Marin*, care pe de o parte denotă un nume de persoană, iar pe de altă parte, scris cu literă mică, devine un adjectiv care sugerează o trăsătură specifică mării.

Prin urmare, sunt folosite următoarele euristici pentru o filtrare preliminară a listei de potențiale entități cu nume:

- Na_ifM = 0 şi Na_ifm > 0
- Na_ifM > 0, dar Na_ifm / Na_ifM > 10
- Na_îf + Na_ifM + Na_ifm < 0.05% * Ndocs, unde Ndocs reprezintă numărul total de documente din corpus.

Pe baza procesării rezultatelor numerice obținute se formează o primă tranşă de cuvinte detectate ca necorespunzătoare pentru a fi candidate în

continuare ca entități cu nume. Astfel sunt eliminate multe dintre apariții, care constituie în majoritatea cazurilor conjuncții, prepoziții sau substantive comune la început de frază.

Pentru restul de expresii potențiale se declanşează procesul de grupare a lor, având drept principale obiective realizarea unei clasament al celor mai des întâlnite entități în corpusul analizat şi gruparea tuturor formelor lexicale folosite pentru a desemna aceeaşi entitate. În cadrul acestei etape este folosită baza de date DEX Online pentru a aduce la forma de bază cuvintele declinate (de ex. *Băncii Centrale a României* va ajunge *Banca Centrală a României*).

Filtrarea potențialelor entități cu ajutorul Wikipedia

Cu toate ajustările făcute, lista de entități, împreună cu formele lor lexicale şi frecvențele de apariție, conține în continuare reziduuri, ca urmare a unei structuri imperfecte a textelor din articolele de ştiri.

Primul pas efectiv în detectarea unei entități reale din lista de entități potențiale (şi totodată a clasificării ei) este făcut cu ajutorul variantei în limba română a enciclopediei Wikipedia. Datorită baze de date foarte mari pe care o pune la dispoziție gratuit, Wikipedia a fost folosită ca sursă pentru căutarea entităților cu nume identificate. De altfel, una dintre noile metode din domeniul REN de identificare a entităților este legarea (*cross-linking*) la Wikipedia, proces denumit şi *wikification*. În acest caz, o entitate nu este neapărat descoperită, cât mai degrabă asociată cu o referință existentă deja.

În cazul de față, entitățile căutate pe Wikipedia şi care sunt regăsite în conținutul acesteia devin confirmate ca entități reale. Mai mult decât atât, arhitectura Wikipedia face posibilă şi clasificarea lor, încă din această etapă, facilitând totodată formarea corpusului adnotat ce va fi folosit în pasul următor, pentru învățarea supervizată.

Clasificarea presupune distribuirea entităților într-una din cele trei categorii: persoane, organizații sau teritorii. Pentru a clasifica entitățile care au pagini asociate pe Wikipedia, sunt folosite două metode:

- În primă fază s-a încercat găsirea casetei de informații tipice pentru fiecare tip de entitate în parte, dar care există doar pentru unele pagini. Astfel, pentru fiecare din cele trei clase, au fost asociați vectori de cuvinte-cheie (de ex. data naşterii pentru persoane) care să indice categoria. În cazul unei potriviri de şablon, entității îi este asociată clasa respectivă.

- În lipsa casetei de informații, este analizată prima frază din articol. Cum Wikipedia are un format specific al primului paragraf care descrie succint entitatea, şi în acest caz este folosită o euristică ce foloseşte şabloane şi expresii regulate. Spre exemplu, în cazul persoanelor, multe dintre descrieri încep cu secvența de caractere "(n. ", imediat după numele persoanei. Alt caz apare când după numele entității urmează un predicat nominal, al cărui nume predicativ este un cuvânt care indică meseria persoanei; în acest caz, cuvântul este căutat pentru potrivire în lista completă de meserii din România, extrasă din alte surse online.

Clasificarea automată a entităților

În urma "wikificării", au fost găsite referințe pentru majoritatea entităților candidat – în special, pentru cele foarte frecvente, care au fost imediat şi clasificate. Rămâne totuşi un număr considerabil de potențiale entități neconfirmate. În acest moment, datorită adnotării cu clase a unui număr semnificativ de entități, este posibilă crearea unui corpus de antrenare pentru un clasificator automat. Acesta poate fi folosit pentru a identifica şi clasifica restul entităților candidat, precum şi a entităților noi.

Dintr-o perspectivă practică, un aspect important din spatele algoritmilor de învățare supervizată care vor fi testați este identificarea atributelor (*feature*-uri) semnificative, care descriu documentele în punctele esențiale pentru a putea permite procesul de inferență şi generalizare. Practic, combinația de atribute cea mai potrivită pentru clasificare se determină prin aplicarea unei funcții-indicator asupra setului de documente – funcția de învățare. Aşa cum a fost menționat în pasul precedent, clasificarea primelor entități identificate a fost făcută cu ajutorul articolelor corespunzătoare paginilor de pe Wikipedia. Aşadar, aceste texte pot constitui şi corpusul de antrenare utilizat pentru a extrage atributele necesare.

Ne propunem deci să determinăm acele atribute care să fie relevante pentru categoria din care face parte fiecare entitate în parte. La prima privire, întregul articol referitor la o entitate ar trebui să fie util pentru discriminare între clase, întrucât se concentrează pe entitatea respectivă. Deci o primă abordare presupune folosirea întregului text al articolului pentru definirea setului de antrenare.

Din punctul de vedere al acestei abordări, atributele care vor fi puse la dispoziție algoritmului de clasificare sunt chiar cuvintele din fiecare articol, eliminând cuvintele de stop şi aplicând lemantizare. Această abordare ar trebui să fie utilă deoarece fiecărei entități îi sunt asociate o multitudine de cuvinte care vor dobândi anumite ponderi în raport cu clasa din care face parte entitatea. Totuşi, în cadrul etapei de testare prin crosvalidare în 10 runde a clasificatorului rezultat, după ce acesta a fost antrenat cu 800 de entități distribuite aproximativ egal în cele trei categorii, s-a constatat obținerea unei acurateți de doar 55% folosind modelul Entropie Maximă din cadrul Mallet.

După analiza acestor rezultate, am ajuns la concluzia că deşi fiecare categorie a avut asociate multe atribute provenite din cuvintele extrase din articolele de pe Wikipedia, acest lucru a avut totuşi un dezavantaj major prin faptul că multe cuvinte s-au dovedit a fi irelevante pentru clasele analizate. De fapt, studiind mai atent un astfel de articol, se observă că multe dintre fraze se referă la lucruri, fapte, persoane aferente entității în cauză, iar acest lucru crează confuzie în cadrul procesului de învățare.

Reconsiderând modul de alegere al setului de antrenare, informațiile păstrate pentru fiecare entitate sunt reduse pentru a fi mai relevante pentru entitatea respectivă (şi, implicit, pentru clasa din care aceasta face parte). Cum aproape întotdeauna prima frază dintr-un articol Wikipedia se referă la entitatea respectivă, este normal ca şi cuvintele constituente ale frazei să fie relevante şi descriptive pentru acea entitate.

Ca o abordare intermediară între cele două strategii, am decis să considerăm doar primul paragraf din fiecare articol. Antrenând din nou corpusul, observăm o creştere a acurateței până la 73%. Cu toate acestea, rezultatele obținute nu sunt acceptabile pentru a putea fi folosite în practică, deci trebuie căutate alte îmbunătățiri.

În mod ideal, ar trebui să extragem acei termeni care să se raporteze sau să determine exact persoana, instituția sau locația referită de către fiecare articol. La nivel de text, acest lucru poate fi făcut analizând morfo-sintactic propozițiile constituente ale articolelor. Multe cuvinte nu descriu în mod direct entitatea. Astfel, această abordare se axează pe extragerea strict a atributelor care se leagă de entitate. Din fiecare articol asociat unei entități sunt identificate adjectivele din vecinătatea entității, verbele predicative ce reflectă o acțiune săvârşită de către entitate şi numele predicative din predicatele nominale asociate entității.

În acest fel, numărul de atribute se reduce considerabil, însă de această dată devine mult mai specific, mai orientat spre descrierea entității. Folosind aceste noi atribute se obține o îmbunătățire substanțială, obținând o acuratețe de 85% folosind acelaşi clasificator Entropie Maximă implementat în Mallet. Această abordare a fost testată cu mai mulți algoritmi de învățare supervizată: SMO, Bayes Naiv şi Entropie Maximă (Mohri, Rostamizadeh şi Talwalkar, 2012) – primii doi implementați în Weka, ultimul în Mallet, iar acuratețea medie pentru cele trei clase variază în intervalul 80-90%. Rezultatele comparației sunt prezentate în secțiunea următoare.

Utilizarea clasificatorului pentru entitățile neconfirmate

În urma etapei de antrenare a clasificatorului, au fost testate mai multe abordări de alegere a atributelor. Mai departe, clasificatorul rulat pe corpusul de antrenare poate fi folosit pentru a clasifica entitățile care nu au fost găsite pe Wikipedia sau pe cele găsite, dar neclasificate din cauză că nu s-au încadrat în şabloanele definite anterior.

Pentru această etapă, este important să poată fi colectat un corpus suficient de mare de fraze ce includ entitățile din care vor fi extrase atributele. O problemă în această abordare este lipsa unui corpus consistent de texte pentru entitățile care apar foarte rar. Tocmai de aceea, pentru entitățile care apar rar, este stringentă obținerea tuturor caracteristicilor posibile existente în text, raportate la respectivele entități. Acest lucru presupune metode avansate de analiză sintactică a textelor (pentru extragerea co-referințelor, arbori de parsare etc.). Dacă în etapa de antrenare, obținerea atributelor se realiza extrăgând din text cuvinte apropiate, care sunt evident legate de către entitate (un adjectiv alăturat entității, un substantiv care apare ca nume predicativ al unui predicat nominal reprezentând o acțiune săvârşită de entitate etc.), de această dată este necesară o analiză aprofundată a întregului context în care apar menționate entitățile ce se doresc clasificate.

Pentru acest lucru, se folosesc parsere sintactice special implementate şi dedicate procesului de stabilire a relațiilor dintre cuvintele unui text. Analizând relațiile ce includ entitatea observată, se extrag cuvintele care fac referire la aceasta şi se adaugă ca atribute pentru setul de test, în scopul

clasificării. Această etapă este în curs de implementare şi nu vor fi raportate rezultate obținute în cadrul acestui articol.

3.2 Extragerea citatelor şi a declarațiilor

Declarațiile unei persoane reprezintă o sursă permanentă de analiză şi de generare a noi subiecte de discuție. În acest moment, în cadrul aplicației dezvoltate, acestea sunt detectate folosind şabloane şi expresii regulate. Pe de o parte, această soluție este foarte rapidă şi conduce la un procent de identificare ridicat. Pe de altă parte, în acest moment nu există un corpus adnotat manual de declarații ca să poată fi aplicată clasificarea automată. Rezultatele satisfăcătoare apar în urma observării folosirii unor şabloane specifice exprimării jurnalistice. În continuare, vor fi prezentate câteva tipuri de exprimări ce anticipează (marchează) declarația unei persoane:

- "Doamnă Ministru, aşa cum ați declarat...", a declarat Victor Ponta.
- Boc a declarat că pentru el este important să termine cursa.
- Preda a mai spus că nimeni nu trebuie să țină de scaun…
- "Voi depune plângere penală împotriva lui Ponta..." a spus Dan Diaconescu.
- Ioan Oltean a declarat vineri, într-o conferință de presă, că România…

După cum se observă şi din exemplele anterioare, o declarație poate însemna, la nivel de text, un extras direct din cuvintele persoanei (un citat) sau poate fi expresia vorbirii indirecte.

Indiferent de tipul declarației, un aspect esențial pe baza căruia se începe căutarea este identificarea unor cuvinte-cheie, care să indice existența unei astfel de declarații. În urma analizei modului în care sunt scrise articolele şi a limbajului folosit de către redactori, se poate observa că aceste cuvinte-cheie sunt reprezentate de anumite verbe care indică faptul că urmează o afirmație sau o declarație a unei terțe persoane. Acestea se numesc verbe de zicere (Bălăşoiu, 2004), iar printre acestea se află: *a spune*, *a declara*, *a zice*, *a afirma*, *a preciza*, *a anunța* ş.a.

Aşadar, ne vom folosi de existența acestor verbe declarative pentru a localiza paragraful din care se poate extrage un citat/o declarație. Abuzând uşor de terminologie, numim un citat spusele unei persoane afişate ca atare în articol (de obicei încadrate în ghilimele) – vorbirea directă, iar o

declarație va fi exprimată prin relatarea la persoana a treia de către autor a afirmației – vorbirea indirectă.

În cazul citatelor, se observă că există, de obicei, simboluri specifice care delimitează vorbirea directă (ghilimele, un paragraf etc.), precum şi faptul că entitatea cu nume apare după verbul declarativ în textele de ştire. Identificând caracterele delimitatoare ce se află în apropierea verbului declarativ, citatul poate fi extras cu uşurință.

Şi al doilea caz, cel al declarațiilor sau al vorbirii indirecte, este caracterizat de anumite cuvinte-cheie. O declarație este precedată într-o majoritate covârşitoare de cazuri de construcții specifice (de exemplu, conjucția "*că*") care urmează verbului declarativ. De cele mai multe ori, aceste construcții specifice sunt situate la nivelul textului imediat după verbul care anunță declarația.

Există însă şi cazuri particulare care trebuie tratate. Spre exemplu, între conjuncția "că" şi verbul declarativ poate apărea o sintagmă care să ofere un plus de informații asupra declarației. Pentru a stabili validitatea unor astfel de situații, grupurile necunoscute de cuvinte sunt tratate morfo-sintactic, pentru a nu depista verbe care să anuleze detectarea citatului sau a declarației.

3.3 Detectarea evenimentelor

Un alt aspect care poate fi exploatat în domeniul prelucrării textului atunci când sunt implicate entitățile cu nume este extragerea evenimentelor la care acestea au participat. Domeniul semantic al perceperii unui eveniment este unul vast şi poate fi mult detaliat şi extrapolat. În cadrul abordării curente, am considerat că un eveniment implică săvârşirea unei acțiuni. La nivelul analizei textuale, o acțiune este exprimată efectiv printr-un predicat, adică un cuvânt sau un grup de cuvinte care să reprezinte un verb sau un grup nominal.

Până a ajunge la verb, trebuie precizată condiția primordială pusă pentru a considera existența unui eveniment, şi anume plasarea sa în timp. Cu alte cuvinte, în textul analizat trebuie localizate coordonate temporale care să precizeze într-o anumită măsură momentul la care evenimentul are (sau a avut) loc. Această idee este folosită în special pentru a colecta informații cât

mai specifice despre eveniment. Tocmai de aceea, precizarea timpului în care se declanşează o anumită acţiune creşte probabilitatea ca aceasta să fie nucleul unui eveniment.

Problema localizării expresiilor temporale în texte ar putea fi din nou un subiect de dezbătut cu privire la ce fel de metode pot fi folosite pentru detectarea lor. Cum acestea sunt un tip specific de entităţi, se pot aplica aceleaşi metode ca cele folosite pentru REN. Însă datorită specificului expresiilor temporale, vom apela la expresiile regulate pentru identificarea lor, pentru că în acest context sunt rapide şi foarte eficiente.

Apariţia verbului împreună cu marcarea timpului nu sunt însă suficiente. Pentru a putea face o conexiune cu entităţile implicate în eveniment, acestea trebuie să fie iniţial detectate în fraza respectivă, iar predicatul exprimat prin verbul găsit să refere o astfel de entitate în arborele de parsare.

După ce avem o listă de astfel de evenimente extrase din texte, este util să descoperim evenimentele care formează un centru de interes mai mare în opinia publică. Deci, ne dorim să grupăm aceste evenimente în funcţie de anumite criterii pentru a putea observa în ce zone se concentrează atenţia oamenilor din mass-media cu privire la persoane publice, locuri, instituţii.

Pentru aceasta am apelat la folosirea algoritmilor de clustering pentru texte (Manning şi Schutze, 1999). Clustering-ul presupune gruparea unui set de obiecte în aşa fel încât obiectele din acelaşi grup au un grad de similaritate mai mare între ele decât au cu cele ce formează alte grupuri.

Ceea ce rămâne de definit este similaritatea între două documente. În cazul textelor, abordarea uzuală implică extragerea cuvintelor şi calcularea ponderilor TF-IDF (*term frequency-inverse document frequency*) ale acestora pentru întregul set de documente. Aceste ponderi vor fi folosite pentru reprezentarea în spaţiul vectorial. Mai exact, fiecare eveniment (frază) din spaţiul limbajului natural va avea asociat un vector de astfel de ponderi, care va fi reprezentarea sa în spaţiul vectorial asociat.

Spaţiul TF-IDF produce o statistică numerică ce reflectă importanţa unui cuvânt într-un document care face parte dintr-un corpus. Valoarea TF-IDF creşte proporţional cu numărul de apariţii ale unui cuvânt într-un document, dar este invers proporţională cu frecvenţa cuvântului în tot corpusul, lucru care ajută la penalizarea cuvintelor care sunt mai comune.

Trecând paragrafele din spaţiul limbajului natural în cel vectorial, în continuare aceste numere trebuie folosite pentru a determina gradul de similaritate dintre evenimente. Măsura folosită pentru acest lucru este

similaritatea cosinus (Manning şi Schutze, 1999). Aceasta presupune determinarea cosinusului unghiului dintre cei doi vectori reprezentând două documente diferite. Valorea obținută este pozitivă, subunitară şi are următoarele semnificații:

- O valoare egală cu 0 implică faptul că cele 2 documente sunt complet diferite (nu au nici un cuvânt în comun).
- O valoarea egală cu 1 semnifică faptul că documentele sunt identice (conțin aceleaşi cuvinte, chiar dacă de un număr diferit de ori).

Prima observație care s-a putut face este faptul că această abordare inițială, de a utiliza fiecare cuvânt din frază, nu oferă scoruri satisfăcătoare.

Având în vedere faptul că un eveniment are la bază un fapt săvârşit, o acțiune, ne dorim ca această acțiune să fie un punct de echivalență între două evenimente. De asemenea, atunci când am determinat un eveniment, la baza căutării sale a stat existența cel puțin a unei entități în fraza ce descrie evenimentul. Deci un alt coeficient comun este prezența aceloraşi entități în cele două evenimente comparate. Pe baza acestei logici, se aplică înainte o "curățare" a corpusului, prin filtrarea cuvintelor irelevante. Practic, se vor păstra doar cuvintele începând cu majusculă care descriu componente de entități cu nume şi cele care sunt verbe (descriind practic acțiunile) în frazele care sunt extrase pentru fiecare eveniment in parte.

4. Rezultate

Rezultatele obținute pentru fiecare dintre cele trei probleme tratate în cadrul acestui articol vor fi evidențiate prin comparații, statistici sau exemple în cadrul acestei secțiuni.

4.1 Detectarea şi clasificarea entităților

După cum am menționat şi în secțiunea anterioară, am folosit trei metode diferite de învățare supervizată pentru clasificarea entităților cu nume, pentru a observa diferențele de performanță. Cele trei metode sunt reprezentate de către: SMO (*Sequential Minimal Optimization*, o variantă de SVM) şi Bayes naiv, implementate în cadrul Weka, precum şi modelul Entropie Maximă din Mallet. Valorile acurateței obținute în urma validării

clasificatoarelor automate prin crosvalidare folosind aceşti algoritmi sunt prezentate în Tabelul 1.

Tabelul 1. Comparație între metodele de învățare supervizată pentru clasificarea entităților cu nume

Algoritmi / Categorii	SMO	Bayes Naiv	Entropie Maximă
Persoane	0.789	0.888	0.784
Organizații	0.804	0.750	0.888
Teritorii	0.939	0.777	0.965

După cum se observă, rezultatele sunt similare în cazul algoritmilor SMO şi Entropie Maximă. Acuratețea uşor mai mică apare în cazul algoritmului Naïve Bayes, generând o valoare uşor mai scăzută decât a celorlalți doi, însă are o viteză de antrenare şi de clasificare mai mare.

Crt.	Persoane	Apariţii	Organizaţii	Apariţii	Teritorii	Apariţii
1	Traian Băsescu	11870	Oltchim	4274	România	17307
2	Victor Ponta	3623	Apple	637	Bucureşti	5531
3	Dan Diaconescu	2465	Google	398	Londra	1805
4	Crin Antonescu	1799	Samsung	215	Germania	1216
5	Adrian Năstase	1117	Renault	129	Franţa	1148
6	Emil Boc	822	Academia Română	100	Cluj	1122
7	Vasile Blaga	537	Microsoft	97	Uniunea Europeană	1115
8	Mona Pivniceru	533	Petrom	93	Statele Unite	1044
9	Barack Obama	531	Rompetrol	90	Marea Britanie	1013
10	Ioan Rus	486	Nokia	84	Bruxelles	962

Figura 2. Topul celor mai frecvente entități cu nume într-un corpus format din 25.000 de ştiri

În toate situațiile de mai sus, procentele au fost stabilite printr-un proces de validare cunoscut sub numele de *10-fold cross-validation* – semnificând faptul că se împarte setul de documente în 10% set de testare, prin care se verifică corectitudinea clasificatorului format, şi 90% set de antrenare.

Pentru a reflecta identificarea entităților extrase într-un corpus format din 25.000 de ştiri colectate în perioada iulie-decembrie 2012, în Figura 2 este reprezentat topul frecvenței entităților cu nume. Sunt afişate cele mai frecvente 10 entități, grupate în cele trei categorii, împreună cu numărul lor de apariții şi sortate descrescător după apariții.

4.2 Extragerea citatelor şi a declarațiilor

Citatele şi declarațiile au fost extrase dintr-un set de aproximativ 50.000 de documente text, în special de pe siturile de ştiri şi bloguri, din aceeaşi perioadă analizată. După cum se observă în Figura 3, predomină în top persoanele politice sau cele cu funcții publice. Deci în opinia publică subiectul politic predomină, precum şi declarațiile făcute de persoanele din acest peisaj.

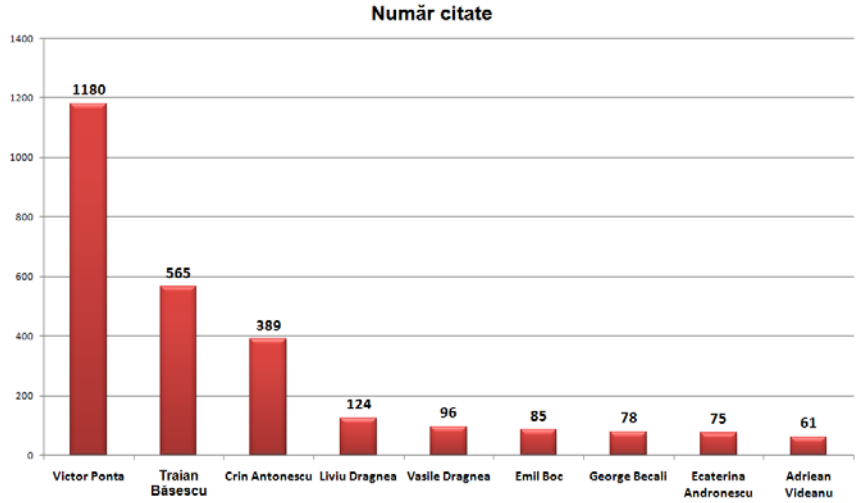

Figura 3. Entitățile cu cele mai multe declarații şi citate extrase din 50.000 de articole de pe siturile de ştiri şi de pe bloguri

4.3 Detectarea evenimentelor

Un eveniment în sine este reprezentat textual printr-o frază. Având în vedere că dintr-un set de 5000 de documente au fost detectate 2235 de evenimente,

pentru a putea oferi statistici relevante care să se limiteze în spațiul articolului, vom afișa în continuare cele mai relevante clustere de câte două evenimente, împreună cu scorul similarității aferent, extrase dintr-un subset de 100 de documente:

- "Duckadam a câştigat cu Steaua Cupa Campionilor Europeni în 1986, la Sevilla, când a apărat 4penalty-uri, intrând în Cartea Recordurilor."
- "Helmut Duckadam are această suferință încă din 1986, la doar două luni după ce a câştigat cu Steaua Cupa Campionilor Europeni la Sevilla, intrând în Cartea Recordurilor pentru că a apărat patru penalty-uri."

  ➔ Scor: **0.7**
- "TVR Info şi-a suspendat emisia pe 15 august, la ora 23."
- "TVR Cultural şi-a suspendat emisia pe 15 septembrie, de la ora 23."

  ➔ Scor: **0.63**
- În rechizitoriul întocmit de procurorii DNA se arată că, în perioada mai-septembrie 2009, Dan Diaconescu l-ar fi amenințat în mod repetat, atât în mod direct, în cadrul emisiunii Dan Diaconescu Direct din 21 iulie 2009, difuzată de postul de televiziune OTV, cât şi indirect, prin intermediul lui Doru Pârv, pe primarul unei comune din județul Arad, pentru a-l determina să le dea suma totală de 200.000 de euro.
- De asemenea, în cursul lunii aprilie 2005, Dan Diaconescu l-ar fi amenințat de mai multe ori, atât în mod direct, în cadrul emisiunii Dan Diaconescu Direct din seara zilei de 20 aprilie 2005, cât şi indirect, prin intermediul lui Ghezea Mitruş, realizatorul emisiunii Semnal de alarmă difuzată pe acelaşi post de televiziune, pe omul de afaceri Paul Petru Țârdea, pentru a-l determina să-i dea suma totală de 100.000 euro.

  ➔ Scor: **0.4**

4. Descrierea aplicației de monitorizare

Sistemul de recunoaştere şi clasificare a entităților numite descris în această lucrare este integrat şi folosit în cadrul unei aplicații de monitorizare a publicațiilor online în limba română dezvoltat de către compania TeamNet

International. În acest moment, sistemul indexează aproape un milion de elemente text diverse incluzând ştiri, articole de pe bloguri, comentarii şi statusuri din rețele sociale, precum şi texte din forumuri de discuții. Sistemul de etichetare cu entități numite este folosit pentru a descoperi aparițiile, citatele şi evenimentele în care sunt implicate entitățile menționate în textele monitorizate.

În prezent, sunt monitorizate peste 10.000 de entități cu nume diferite, însă sperăm ca această listă să fie extinsă la o iterație ulterioară a sistemului descris în această lucrare, folosind un corpus mai mare pentru etapa de învățare semi-supervizată. Mai mult, pentru fiecare entitate cu nume, aplicația reține mai multe forme lexicale (de exemplu, "CCR", "Curtea Constituțională a României", "CC a României"), iar acestea sunt folosite pentru REN în texte. După cum am menționat şi în capitolele anterioare, fiecare formă lexicală este, de asemenea, identificată dacă apare cu sau fără diacritice, în forma normală sau flexionată.

Aplicația permite şi vizualizarea interactivă a informațiilor despre entități, putând compara numărul de apariții în textele monitorizate ale mai multor entități numite diferite, precum şi numărul de citatelor şi ale evenimentelor în care acestea sunt implicate. Această aplicație este încă în curs de dezvoltare, iar la finalizarea sa rezultatele vor fi publicate într-un articol ulterior.

5. Concluzii

Prelucrarea limbajului natural constituie una dintre cele mai complexe şi interesante ramuri ale inteligenței artificiale, întrucât presupune un proces de tranziție între perceperea limbajului scris de către creierul uman şi învățarea conotațiilor acestui limbaj de către calculator. Problema identificării şi a clasificării entităților a reprezentat de la început o provocare pentru cercetătorii care au lansat ideea detectării automate a acestora în texte.

Această provocare a condus mai departe la noi centre de interes. Extragerea citatelor din text, detectarea şi extragerea evenimentelor reprezintă o parte din activitățile în continuă dezvoltare din universul PLN şi al extragerii informațiilor din texte.

Articolul de față a prezentat o abordare a problemei recunoaşterii şi a clasificării entităților, a detectării automate a citatelor şi a extragerii evenimentelor din text. Combinând atât modalități clasice de folosire a regulilor şi a expresiilor regulate, cât şi tehnici moderne de învățare automată şi statistică, s-a putut observa pe parcurs cum evoluția rezultatelor a depins de alegerea inteligentă şi contextual potrivită a euristicilor şi a metodelor de clasificare folosite. În plus, probabil acesta este cel mai complex studiu pentru REN în limba română pentru articole de ştiri, din bloguri sau texte din rețelele sociale. Deşi rezultatele obținute sunt rezonabile pentru a fi integrate în aplicații practice (85% pentru clasificarea entităților cu nume), considerăm că o rafinare a lor mai poate fi făcută în viitorul apropiat pentru a obține performanțe şi mai bune.

Studiile anterioare ale REN în limba română au atins rezultate mai bune, însă pentru corpusuri de validare de dimensiuni mult mai reduse şi pentru categorii de texte mai puțin variate, în special pentru texte jurnalistice. De exemplu, Pastra et al. (2002) raportează o precizie de 88% pentru persoane, 92% pentru locații şi 95% pentru organizații, însă considerăm că este folosit un corpus destul de mic şi insuficient variat: "1MB de texte" din ziarul "Amprenta" (Pastra et al.). Pe de altă parte, un alt studiu efectuat pentru REN în limba română nu specifică detalii despre rezultatele obținute sau corpusul folosit pentru antrenare si validare (Tufis et al. 2008). Din acest motiv, considerăm că rezultatele din această lucrare aduc o valoare față de toate studiile anterioare, în special datorită volumului mare de texte folosite, precum şi abordării semi-supervizate pentru rezolvarea problemei recunoaşterii şi clasificării entităților numite.

Referințe

Aggarwal, C.C., Zhai, C. *Mining Text Data*. Springer, 2012.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. Fast Discovery of Association Rules. *Advances in knowledge discovery and data mining*, American Association for Artificial Intelligence, Menlo Park, CA, 1996.

Bălăşoiu, C. *Discursul raportat în textele dialectale româneşti*, Ed. Univ. Bucureşti, Bucureşti, 2004.

Bikel, D.M., Miller, S., Schwartz, R., Weischedel, R. Nymble: a high-performance learning name-finder. *Proceedings of the fifth conference on Applied natural language processing (ANLC '97)*, Association for Computational Linguistics, Stroudsburg, USA, pp. 194-201, 1997.

Bird, S. NLTK: the Natural Language Toolkit. *Proceedings of the COLING/ACL on Interactive presentation sessions* (COLING-ACL '06). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 69-72, 2006.

Brodely, C.E., Friedl, M.A. Identifying and Eliminating Mislabeled Training Instances. *Proceedings of the thirteenth national conference on Artificial intelligence - Volume 1 (AAAI'96)*, Vol. 1. AAAI Press, pp. 799-805, 1996.

Davis, C. *Using machine learning to extract quotes from text*. CIR Labs, available online at http://cironline.org/blog/post/using-machine-learning-extract-quotes-text-3687, 2012.

Declerck, T. *Automatic event extraction from text on the base of linguistics and semantic annotation*. Language Technology Lab, 2005.

Finkel, J.R., Grenager, T., Manning, C. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* pp. 363-370, 2005.

Foroutan, I., Sklansky, J. Feature Selection for Automatic Classification of Non-Gaussian Data. *IEEE Transactions on Systems, Man and Cybernetics 17 (2)*, 1987.

Grishman, R. The NYU system for MUC-6 or where's the syntax?. *Proceedings of the 6th conference on Message understanding (MUC6 '95)*, Association for Computational Linguistics, Stroudsburg, USA, pp. 167-175, 1995.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. The WEKA data mining software: an update. *SIGKDD Explor. Newsl., 11(1)*, 2009.

Jurafsky, D., Martin, J.H. *Speech and Language Processing*. Pearson Prentice Hall, 2008.

Krestel, R., Bergler, S., Witte, R. Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, European Language Resources Association, Marrakech, Morocco, 2008.

Manning, C.D., Schutze, H. Foundation of Statistical Natural Language Processing. MIT Press, 1999.

Mikheev, A., Moens, M., Grover, C. Named Entity recognition without gazetteers. *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics (EACL '99)*, Association for Computational Linguistics, Stroudsburg, USA, pp. 1-8, 1999.

McCallum, A.K. *MALLET: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu, 2002.

Mohri, M., Rostamizadeh, A., Talwalkar, A. *Foundations of Machine Learning*. MIT Press, 2012.

Pastra, K., Maynard, D., Hamza, O., Cunningham, H., Wilks, Y. How feasible is the reuse of grammars for Named Entity Recognition. *Proceedings of the 3rd Conference on Language Resources and Evaluation - LREC 2002*, ELRA - European Language

Ressources Association, 2008.

Sayyadi, H., Hurst, M., Maykov, A. Event Detection and Tracking in Social Streams. *Proceedings of International Conference on Weblogs and Social Media (ICWSM 2009)*, AAAI, 2009.

Smith, M.R., Martinez, T. Improving Classification Accuracy by Identifying and Removing Instances that Should Be Misclassified. *Proceeding of International Joint Conference on Neural Networks*, 2011.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. RACAI's Linguistic Web Services. *Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008*, Marrakech, Morocco, ELRA - European Language Ressources Association, 2008.

Yang, Y., Pierce, T., Carbonell, J.. A study of retrospective and on-line event detection. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '98), ACM, New York, USA, pp. 28-36, 1998.