(19) OFICIUL DE STAT
PENTRU INVENŢII ŞI MĂRCI
Bucureşti

ROMÂNIA

(11) RO 127582 A2
(51) Int.Cl.
G06F 17/20 (2006.01)

(12) # CERERE DE BREVET DE INVENŢIE

(21) Nr. cerere: **a 2010 01247**

(22) Data de depozit: **29.11.2010**

(41) Data publicării cererii:
29.06.2012 BOPI nr. **6/2012**

(71) Solicitant:
• **UNIVERSITATEA "POLITEHNICA" DIN BUCUREŞTI**, *SPLAIUL INDEPENDENŢEI NR.313, SECTOR 6, BUCUREŞTI, B, RO*

(72) Inventatori:
• **BURILEANU DRAGOŞ**, *STR. JOHANNES KEPLER NR. 2, BL. 2, SC. 2, AP. 61, SECTOR 2, BUCUREŞTI, B, RO;*
• **UNGUREAN CĂTĂLIN**, *STR. MOLDOVIŢA NR. 2, BL. M2D9/4, AP. 44, SECTOR 4, BUCUREŞTI, B, RO*

(54) # METODĂ DE INSERARE AUTOMATĂ A SEMNELOR DIACRITICE PENTRU SINTEZA VORBIRII PORNIND DE LA TEXT ÎN LIMBA ROMÂNĂ ÎN APLICAŢII CE AU CA SUPORT REŢELE DE DATE

(57) **Rezumat:**

Invenţia se referă la o metodă de inserare automată a semnelor diacritice în texte în limba română scrise fără diacritice, destinată utilizării în sinteza vorbirii pornind de la text, pentru aplicaţii ce trebuie să prelucreze documente în format electronic şi au ca suport reţele de date. Metoda conform invenţiei are la bază o strategie statistică axată pe *n*-grame, şi constă în esenţă într-un proces de filtrare secvenţială pe mai multe niveluri, bazat pe probabilitatea de apariţie a cuvintelor scrise cu diacritice şi pe contextul în care acestea apar. Fiecare criteriu de filtrare aduce o îmbunătăţire suplimentară, fără a mări caracterul de ambiguitate. Astfel, după ce tratează iniţial cuvintele fără diacritice sau pentru care nu există ambiguităţi, metoda utilizează incremental trei etaje de filtrare: la nivel de unigrame, la nivel de bigrame şi la nivel de trigrame, informaţia contextuală fiind adăugată gradual, după fiecare etaj, simultan cu eliminarea variantelor incorecte.

Revendicări: 1

RO 127582 A2

METODĂ DE INSERARE AUTOMATĂ A SEMNELOR DIACRITICE PENTRU SINTEZA VORBIRII PORNIND DE LA TEXT ÎN LIMBA ROMÂNĂ ÎN APLICAŢII CE AU CA SUPORT REŢELE DE DATE

Invenţia se referă la o metodă de inserare automată a semnelor diacritice în texte în limba română scrise fără diacritice, destinată utilizării în sinteza vorbirii pornind de la text pentru aplicaţii ce trebuie să prelucreze documente în format electronic şi au ca suport reţele de date, cum ar fi spre exemplu ascultarea mesajelor de tip e-mail/SMS folosind reţelele telefonice publice sau citirea unor informaţii de pe pagini web.

Sunt cunoscute un număr de studii pe plan mondial în problema refacerii caracterelor scrise cu semne diacritice, realizate în special pentru limbile franceză, spaniolă, ungară şi doar parţial pentru limba română. Au fost astfel propuse câteva metode, dintre care unele se bazează pe algoritmi pur-statistici, fără alte cunoştinţe lingvistice, iar altele folosesc modele de limbaj şi diferite niveluri de procesare lingvistică.

– De exemplu, din categoria metodelor statistice se poate menţiona un algoritm de restaurare a diacriticelor care foloseşte tehnici de învăţare automată şi care operează la nivel de caracter, propus în [R. Mihalcea, V. Nastase, "Letter Level Learning for Language Independent Diacritics Restoration", in *Proc. CoNLL 2002*, Taipei, Taiwan, 2002, pp. 105-111]. Algoritmul a fost implementat pentru limbile cehă, ungară şi poloneză iar pentru limba română performanţa raportată (la nivel de caracter) a fost de aproximativ 98,3%.

Dezavantajele acestei metode sunt legate de următoarele aspecte: în primul rând, considerăm că abordarea bazată doar pe inspectarea literelor ce înconjoară caracterul analizat, nu este adecvată limbii române, unde multe ambiguităţi pot fi rezolvate doar la nivel de cuvânt (un exemplu simplu este cel legat de substantivele feminine articulate/nearticulate); în al doilea rând, rezultatele raportate sunt puţin relevante, deoarece baza de date folosită pentru testarea metodei este foarte mică (aproximativ 50.000 de cuvinte).

Lucrările au fost realizate în cadrul contractului
de cercetare UPB-CNCSIS, IDEI-782

RECTOR,
Prof. dr. ing. Ecaterina Andronescu

– Din cea de-a doua categorie de metode menţionate, în [D. Tufiş, A. Chiţu, "Automatic Diacritics Insertion in Romanian Texts", in *Proc. of COMPLEX'99*, Pecs, Hungary, June 16-19, 1999, pp. 185-194] este propus un algoritm mai complex de inserare a diacriticelor pentru textele scrise în limba română, care funcţionează pe baza informaţiei de parte de vorbire şi conduce la o acurateţe raportată de 97,4%.

Dezavantajul major al acestei metode este complexitatea mare a algoritmului, bazat pe resurse externe (dicţionare) de dimensiuni mari şi prelucrări lingvistice adiţionale, fapt ce nu este compatibil cu utilizarea sa eventuală într-un sistem de sinteză automată a vorbirii.

Problema tehnică pe care şi-a propus să o rezolve invenţia poate fi formulată pornind de la constatarea că *sinteza automată a vorbirii* reprezintă o tehnologie importantă pentru aplicaţiile care au ca suport reţelele de date. Astfel de aplicaţii, cum ar fi cele de citire a mesajelor de tip text (e-mail, SMS, pagini HTML etc.) trebuie să rezolve numeroase probleme, una dintre acestea fiind necesitatea refacerii semnelor diacritice în textul original. De obicei, majoritatea utilizatorilor nu folosesc diacritice (semne distinctive plasate deasupra, în interiorul sau dedesubtul caracterelor), chiar dacă terminalul (sau sistemul de operare) permite acest lucru; pe de altă parte, în cazul proceselor de lucru care utilizează o codare limitată sau neunitară a caracterelor (de exemplu cea pe 7 biţi), diacriticele sunt înlocuite fie cu semne indescifrabile fie cu caracterele de bază (în cel mai bun caz). Această situaţie este valabilă atât pentru limba română, cât şi pentru alte limbi vorbite care folosesc alfabetul latin (franceza, spaniola, germana, daneza, ungara, poloneza, suedeza etc.) şi care trebuie să-şi suplimenteze setul alfabetic de bază prin folosirea literelor scrise cu semne diacritice (sau *markeri* de accent) pentru a indica sunete (sau accente) diferite. Pe de altă parte, sinteza automată a unui text scris fără diacritice duce la o inteligibilitate scăzută; uneori, ambiguităţile sintactice şi semantice fac ca o întreagă frază rostită să fie aproape de neînţeles. Repoziţionarea automată a diacriticelor reprezintă însă o problemă extrem de dificilă, deoarece practic nu există reguli lingvistice care să permită rezolvarea acestei sarcini.

Metoda care face obiectul prezentei invenţii rezolvă problema tehnică menţionată anterior prin faptul că la intrarea în sistemul de sinteză a vorbirii, se verifică mai întâi dacă textul de intrare (scris la tastatura calculatorului sau provenind dintr-un fişier stocat anterior) conţine caractere scrise cu diacritice, iar dacă acest lucru nu se întâmplă, metoda se aplică prin intermediul unui algoritm eficient de inserare automată a semnelor diacritice.

Metoda este originală, în contextul preocupărilor similare pe plan naţional şi internaţional. Principalul aspect original ce poate fi menţionat constă într-un proces de *filtrare*

Lucrările au fost realizate în cadrul contractului
de cercetare UPB-CNCSIS, IDEI-782

RECTOR,
Prof. dr. ing. Ecaterina Andronescu

secvenţială, pe mai multe niveluri, bazat pe probabilitatea de apariţie a cuvintelor scrise cu diacritice şi pe contextul în care acestea apar. Fiecare criteriu de filtrare aduce o îmbunătăţire suplimentară, fără însă a mări caracterul de ambiguitate (deci formele posibile teoretic de scriere a unui cuvânt cu sau fără diacritice). Astfel, metoda utilizează practic trei etaje de filtrare: la nivel de *unigrame* (pe cuvânt), la nivel de *bigrame* (de asemenea pe cuvânt) şi la nivel de *trigrame* pentru terminaţii de cuvânt. Un al doilea aspect important este reprezentat de faptul că se realizează în fapt o filtrare minimală la nivelul bazei de date de antrenare, îndepărtând, într-o primă fază, numai formele asupra cărora se decide că sunt "sigur greşite".

Prezenta invenţie înlătură dezavantajele metodelor cunoscute, în primul rând prin faptul că se aplică ca un modul de pre-procesare înaintea unui sistem de sinteză a vorbirii pornind de la text, permiţând astfel în final sinteza corectă a unui text scris fără diacritice. În acest sens, trebuie subliniat faptul că metodele abordate pe plan mondial (pentru diferite limbi unde sunt utilizate în scriere semne diacritice), ca şi cele pentru limba română prezentate anterior, sunt destinate în marea lor majoritate doar aplicaţiilor şi studiilor strict lingvistice. În al doilea rând, metoda care face obiectul invenţiei este extrem de performantă, atât din punct de vedere al acurateţei obţinute, cât şi a timpului de calcul şi a resurselor de memorie necesare (ambele extrem de importante în situaţia în care întreg sistemul de sinteza a vorbirii trebuie să funcţioneze în timp real). Astfel, metoda utilizează o abordare statistică axată pe *n*-grame, se bazează pe cunoştinţe lingvistice limitate şi necesită un corpus de antrenare de dimensiuni medii.

Se prezintă în continuare în detaliu obiectul invenţiei, un exemplu de aplicare a sa constând din utilizarea metodei de către autorii invenţiei ca modul de pre-procesare în cadrul unui sistem complet de sinteză a vorbirii pornind de la text în limba română (dezvoltat în cadrul colectivului din care fac parte autorii), sistem folosit la rândul său într-o aplicaţie de tip "cititor de e-mail/SMS" pentru telefoane mobile.

Limba română foloseşte trei semne diacritice, şi anume *accentul circumflex*, *sedila* şi *accentul grav*, rezultând cinci litere scrise cu diacritice: **ă, â / î** (folosite pentru acelaşi sunet dar în circumstanţe diferite), **ş** şi **ţ**. Unele dintre diacritice pot face diferenţa între două forme distincte ale aceluiaşi cuvânt, de exemplu **casa** – **casă** (substantiv articulat – nearticulat), **înalta** – **înaltă** (adjectiv articulat – nearticulat), sau între două cuvinte complet diferite ca înţeles, de exemplu **fată** – **faţă**. Trebuie menţionat faptul că în textele scrise în limba română,

Lucrările au fost realizate în cadrul contractului
de cercetare UPB-CNCSIS, IDEI-782

RECTOR,
Prof. dr. ing. Ecaterina Andronescu

cuvintele scrise cu diacritice au un aport semnificativ, mai mare decât în alte limbi. Astfel, se estimează că frecvenţa lor de apariţie este între 25% şi 40%, în timp ce în limba franceză aproximativ 15% dintre cuvinte pot avea diacritice. În alte limbi, ca olandeza, ceha şi slovaca, aportul cuvintelor scrise cu diacritice este semnificativ mai mare decât în limba română.

După o analiză făcută pe un volum mare de texte, extrase din diferite domenii, am decis că putem clasifica cuvintele limbii române, din punctul de vedere al interacţiunii lor cu semnele diacritice, în următoarele cinci categorii (le vom folosi ulterior cu denumirea de "**tipul 1**", "**tipul 2**" etc.):

1) Cuvinte scrise fără diacritice; exemplu: **lemn**.
2) Cuvinte scrise întotdeauna cu diacritice: **câteva, ştiinţific**.
3) Cuvinte ambigue, cu două forme: **masa, masă**.
4) Cuvinte ambigue în care un anumit număr de diacritice sunt întotdeauna prezente: **cămaşă, cămaşa**.
5) Cuvinte ambigue în care oricare semn diacritic poate să apară sau nu: **pana, pană, pânã**.

Această împărţire pe clase distincte a cuvintelor scrise cu diacritice stă la baza realizării metodei de inserare automată a semnelor diacritice, care constă în două etape mari: *antrenare* şi *testare*.

Etapa de antrenare presupune realizarea următorilor paşi:

1. Construirea manuală a unui dicţionar *D*1 care conţine cele mai frecvent folosite cuvinte din limba română şi care conţine de asemenea cât mai multe forme flexionate ale acestora.
2. Pe baza lui *D*1 se construieşte o structură dicţionar *D*2 (*"hash table"*) care pune în corespondenţă fiecare cuvânt din *D*1, cu diacriticele înlăturate, cu toate formele sale posibile de scriere cu diacritice (extrase din *D*1).
3. Pe baza unui corpus de antrenare *A*, care conţine forme corecte scrise cu diacritice, se construiesc la nivelul cuvintelor un set de trei dicţionare, astfel: o structură *U* care conţine *unigramele* extrase din dicţionarul de antrenare împreună cu frecvenţele lor de apariţie, o structură *B* care conţine toate *bigramele* posibile extrase din corpusul *A*, şi o structură *T*, care conţine *trigramele* la nivelul sufixelor cuvintelor din *A*.

Trebuie menţionat faptul că în cazul metodei propuse, termenul de *sufix* al cuvintelor, cu care realizăm dicţionarul *T* nu se referă strict la conceptul morfologic cunoscut, ci se referă

de fapt la ultimele caractere ale fiecărui cuvânt. Acest număr de caractere finale a fost testat în intervalul 1 la 4 pe baza rezultatelor oferite în etapa de testare, în aşa fel încât erorile de inserare a diacriticelor să fie minimizate. S-a observat că iniţial are loc o scădere a erorilor de refacere a diacriticelor odată cu creşterea numărului de litere finale cu care se formează trigramele, proces urmat apoi de o creştere a numărului acestor erori simultan cu obţinerea unui dicţionar *T* de dimensiuni foarte mari.

Etapa de testare (cu alte cuvinte **de funcţionare propriu-zisă**) este compusă în principal din trei procese de filtrare în cascadă, în care datele de la ieşirea fiecărui etaj superior constituie date de intrare în etajul imediat următor. Intrarea generală constă în secvenţe de text în care semnele diacritice au fost înlăturate. Etapa de testare funcţionează astfel:

1. Pe baza dicţionarelor *D*1 şi *D*2 sunt inserate formele care nu conţin diacritice şi formele care nu introduc ambiguităţi, altfel spus, cuvintele de **tipul 1** respectiv de **tipul 2**. Etapa este una de tip determinist şi presupune o înlocuire biunivocă a formei fără diacritice cu cea corectă (cu semnele diacritice inserate).
2. Pentru fiecare pereche de cuvinte consecutive din textul test, care conţine cel puţin o formă ambiguă, se formează setul de variante posibile cu toate aceste cuvinte; dacă cel puţin o pereche din acest set este găsită în corpusul de bigrame *B*, sunt eliminate toate perechile de variante care nu apar în *B*, altfel sunt lăsate toate variantele nemodificate.
3. Din textul de intrare care conţine variantele de ieşire de la pasul 2, se formează toate tripletele de câte trei cuvinte consecutive şi din acestea se extrag sufixele; se obţin astfel perechi alternative de triplete de tip cuvinte-sufixe; dacă există cel puţin o tripletă de sufixe în dicţionarul *T*, atunci renunţă la toţi membrii din setul de cuvinte corespunzătoare tripletelor de sufixe care nu sunt în *T* şi furnizează spre ieşire doar variantele pentru care s-au găsit membri în *T*; dacă nu s-a găsit niciuna dintre variante în *T*, lasă tot setul de cuvinte neschimbat.
4. Folosind setul de unigrame *U*, păstrează pentru fiecare cuvânt doar varianta cea mai probabilă, pe baza frecvenţelor de apariţie calculate din corpusul de antrenare, dintre variantele furnizate la pasul **3**.

În acest punct al descrierii metodei ce face obiectul invenţiei, trebuie făcute câteva precizări referitoare la modul de obţinere a dicţionarelor de prelucrare şi a corpusurilor de *n*-grame.

Lucrările au fost realizate în cadrul contractului
de cercetare UPB-CNCSIS, IDEI-782

RECTOR,
Prof. dr. ing. Ecaterina Andronescu

a. Obţinerea dicţionarelor de prelucrare

– Se porneşte de la un dicţionar care conţine cuvinte scrise cu diacritice, cu cât mai multe forme flexionate. Concret, dicţionarul astfel creat are un număr de peste 330.000 de intrări.

– Se elimină semnele diacritice păstrându-se câte o singură formă din cuvintele astfel obţinute şi se compară cu dicţionarul de start. Sunt separate în acest fel cuvintele cu probleme la refacerea diacriticelor de cele fără ambiguităţi. Dicţionarul cuvintelor "cu probleme" va fi folosit şi în etapa de reducere a dimensiunilor fişierelor cu *n*-grame.

După această primă etapă, se obţin două dicţionare distincte, la care se mai adaugă un dicţionar de nume proprii şi unul de abrevieri, create *offline*, formându-se un aşa-zis *bloc al dicţionarelor de prelucrare*.

b. Obţinerea fişierelor reduse de *n-grame*

Se formează *offline* un corpus, prin juxtapunerea de texte scrise cu diacritice. Pe baza acestui corpus de start de aproximativ 10^7 intrări, se extrag *unigramele*, *bigramele* şi *trigramele*.

Studiile realizate de autorii invenţiei au arătat că în limba română cele mai mari probleme pentru refacerea diacriticelor sunt cele create de ambiguitatea **a** – **ă** la finalul cuvintelor. Acest fenomen este produs cu o frecvenţă mare de apariţie de unele construcţii lingvistice scurte precum **ca** – **că**, **sa** – **să**, **sau** – **său** etc. şi de formele articulate/nearticulate ale substantivelor feminine, de exemplu: **casa** – **casă**, **piaţa** – **piaţă** etc.

Din perspectiva ideii expuse anterior, am introdus un etaj de filtrare care foloseşte trigramele obţinute din terminaţiile cuvintelor. Câştigul adus de acest etaj este de aproximativ 0,8%, mult mai bun decât cel al trigramelor obţinute din cuvintele întregi, tocmai prin abordarea unei probleme care apare cu o incidenţă foarte mare.

Pentru reducerea dimensiunilor fişierelor cu *n*-grame (implicit pentru scăderea resurselor necesare în prelucrare), se poate folosi dicţionarul cuvintelor cu probleme din blocul dicţionarelor de prelucrare. Cu alte cuvinte, se păstrează numai *n*-gramele cuvintelor care introduc ambiguităţi la refacerea diacriticelor, celelalte forme fiind înlăturate.

Pentru exemplul concret discutat de aplicare a invenţiei vom prezenta în continuare câteva aspecte importante referitoare la metoda expusă.

În primul rând trebuie specificat faptul că în etajele implementate cu bigrame şi trigrame nu au fost folosite probabilităţile de apariţie pentru aceste *n*-grame. Deşi o

Lucrările au fost realizate în cadrul contractului
de cercetare UPB-CNCSIS, IDEI-782

RECTOR,
Prof. dr. ing. Ecaterina Andronescu

implementare realizată la nivel de cuvânt, folosind probabilităţile trigramelor obţinute pe baza unui corpus de antrenare, ar fi fost binevenită, dimensiunile corpusului de antrenare ar trebui să fie uriaşe, de ordinul a câteva sute de MB, pentru a se obţine un corpus de trigrame eficient, care să conţină cât mai multe dintre trigramele posibile (dată fiind şi dimensiunea mare a dicţionarului de start $D1$). În aceste condiţii fenomenele nedorite produse de insuficienţa datelor de antrenare (*"data sparseness"*) ar fi fost clare, lucru pus în evidenţă şi în cazul metodei propuse, în care la nivelul trigramelor s-a lucrat doar cu terminaţii de cuvinte pentru care singura condiţie impusă a fost cea de existenţă a trigramelor.

Astfel s-a putut observa că, odată cu creşterea lungimii trigramelor, după ordinul 4, etajul trigramelor introduce erori mari din cauza fenomenului amintit. Acest lucru este uşor de intuit deoarece, odată cu creşterea lungimii trigramelor, lungimea sufixelor devine comparabilă cu dimensiunea cuvintelor, iar dimensiunea vocabularului de sufixe tinde să se apropie de dimensiunea lui $D1$. Etajul trigramelor este eficient atâta vreme cât se realizează un bun echilibru între dimensiunea vocabularului de sufixe şi dimensiunea corpusului de antrenare. Chiar şi în aceste condiţii minimale impuse bigramelor şi trigramelor – care presupun numai existenţa acestora, s-a putut observa încă efectul nedorit al insuficienţei datelor de antrenare, tradus prin lipsa unor n-grame mai puţin întâlnite şi furnizarea către etajul unigramelor a unor variante incorecte. Acest etaj final va decide varianta cea mai probabilă, urmărind probabilităţile de apariţie ale fiecărei variante calculate pe baza corpusului de antrenare A. Două aspecte sunt de remarcat referitor la etajul final: în primul rând, trebuie spus că dacă în U nu există nici o variantă, cuvântul de intrare este lăsat nemodificat. În al doilea rând, deoarece acesta este un etaj final, o singură variantă – cea mai probabilă, va fi furnizată la ieşire.

De asemenea, trebuie menţionat faptul că metoda de refacere a diacriticelor ce face obiectul invenţiei a fost dezvoltată într-o manieră incrementală, tratând iniţial cuvintele fără diacritice sau pentru care nu există ambiguităţi, după care se adaugă gradual informaţie contextuală furnizată prin nivelul bigramelor de cuvinte şi/sau cel al trigramelor – simultan cu eliminarea variantelor incorecte, pentru ca în final să se filtreze variantele rămase prin folosirea probabilităţilor unigramelor calculate pe baza frecvenţelor de apariţie obţinute din corpusul de antrenare. Fiecare nivel de procesare aduce un plus în reducerea ratei de eroare la nivel de cuvânt. Deşi s-au încercat diferite variante de succesiune ale etajelor de filtrare, am constatat că cea mai bună configuraţie este cea în care etajul unigramelor este ultimul. Acest lucru poate fi demonstrat şi teoretic deoarece fiecare etaj de filtrare, cu excepţia celui de unigrame ia în calcul numai apariţia perechilor de cuvinte sau tripletelor de sufixe, fără a se

Lucrările au fost realizate în cadrul contractului
de cercetare UPB-CNCSIS, IDEI-782

RECTOR,
Prof. dr. ing. Ecaterina Andronescu

ţine cont de frecvenţele de apariţie ale acestora. În acest fel, pentru un text de intrare dat, nivelurile de filtrare realizate cu bigrame respectiv trigrame pot întoarce una, două sau chiar toate variantele ambigue, printre care se află şi varianta corectă. Fiecare set de filtrare reduce, teoretic, acest set de variante oferite către ieşire, dar acest lucru nu este garantat.

Ceea ce se urmăreşte cu primele două niveluri este ca, din punct de vedere statistic, numărul variantelor incorecte furnizate către ultimul etaj să fie sensibil diminuat. Prin urmare, etajul bigramelor şi cel al trigramelor reduce numărul de variante incorecte furnizat către etajul final, însă numai etajul unigramelor va întoarce câte o singură variantă pentru fiecare set de variante posibile ale fiecărui cuvânt, adică pe cel mai frecvent, eliminându-le pe celelalte. Toate rezultatele experimentale pe care le vom raporta în continuare vor avea ca ultim nivel de prelucrare etajul unigramelor.

În cadrul metodei ce face obiectul invenţiei, antrenarea a fost făcută pe un corpus extras din texte literare scrise în limba română, care conţine peste 10^7 cuvinte. Aceste texte au contribuit de asemenea la extinderea dicţionarului *D*1, construit iniţial manual, pe baza *Dicţionarului Explicativ al Limbii Române*. Din păcate acest dicţionar conţine numai formele de bază ale cuvintelor, adică sub 70.000, iar pentru rularea algoritmului am avut nevoie de un dicţionar care să conţină cât mai multe forme morfologice flexionate. La ora actuală dicţionarul conţine aproximativ 330.000 de forme şi include aproximativ 6.000 de nume proprii. Dicţionarele (vom folosi mai departe şi denumirea de *corpusuri*) de unigrame, bigrame şi trigrame au fost extrase din textele de antrenare şi salvate ca fişiere text separate.

Corpusul de testare a fost corectat manual şi conţine propoziţii extrase din articole, texte literare şi teze de doctorat din mai multe domenii de activitate, disponibile în format electronic în limba română. Acest corpus are următoarele caracteristici: (i) numărul de cuvinte de intrare: **1.200.000**, (ii) procentul de cuvinte scrise fără diacritice: 57,41%; procentul de cuvinte scrise întotdeauna cu diacritice: 16,33%; procentul de cuvinte ambigue la refacerea diacriticelor: 26,26%.

Pe baza celor descrise anterior, au fost calculate următoarele mărimi de performanţă standard, utilizate în testele statistice clasice:

1. *Precizia*: definită ca raportul dintre numărul total de diacritice inserate corect şi numărul total de diacritice inserate.
2. *Recall*: definit ca raportul dintre numărul total de diacritice inserate corect şi numărul de diacritice din baza de test corectată manual.
3. *F-measure*: definită ca media armonică dintre *precizie* şi *recall*.

În Tabelul 1 sunt prezentate principalele rezultate obţinute, detaliate pentru fiecare dintre cele patru clase de ambiguitate enunţate anterior. Rezultatele au fost calculate pentru trei versiuni de rulare ale metodei propuse, astfel:

(i) Ca bază de comparaţie s-a luat metoda de restaurare a diacriticelor pe baza dicţionarului *D2*, pentru cuvintele care nu prezintă ambiguităţi, fără a fi necesare date de antrenare.

(ii) A doua metodă foloseşte bigramele extrase din corpusul de antrenare şi probabilităţile unigramelor calculate pe baza aceluiaşi corpus.

(iii) Metoda completă, care constă din toate cele trei etaje de filtrare aşezate succesiv: etajul bigramelor, cel al trigramelor cu sufixe, şi cel ce utilizează unigramele.

Clasa de ambiguitate	**Precizie (%)**	**Recall (%)**	***F*-measure (%)**	***F*-measure medie (%)**
Baza de comparaţie				
a / ă / â	78.03 / 92.69 / 99.50	99.65 / 35.27 / 61.86	87.53 / 51.10 / 76.29	77.11
i / î	98.63 / 99.67	99.96 / 89.54	99.30 / 94.34	98.71
s / ş	94.83 / 97.89	99.48 / 81.48	97.10 / 88.93	95.25
t / ţ	95.30 / 91.25	98.82 / 71.63	97.03 / 80.25	94.57
Bigrame şi unigrame				
a / ă / â	97.82 / 91.58 / 99.24	96.49 / 94.66 / 99.57	97.15 / 93.10 / 99.40	96.19
i / î	99.97 / 99.57	99.94 / 99.82	99.95 / 99.69	99.93
s / ş	99.93 / 99.49	99.85 / 99.76	99.89 / 99.63	99.83
t / ţ	99.86 / 98.07	99.66 / 99.22	99.76 / 98.64	99.60
Bigrame, trigrame şi unigrame				
a / ă / â	98.22 / 93.30 / 99.26	97.22 / 95.62 / 99.55	97.72 / 94.45 / 99.41	**96.93**
i / î	99.98 / 99.56	99.94 / 99.82	99.96 / 99.7	**99.93**
s / ş	99.93 / 99.51	99.85 / 99.77	99.89 / 99.64	**99.84**
t / ţ	99.87 / 98.2	98.69 / 99.27	99.78 / 98.74	**99.63**

Tabelul 1. Măsurarea performanţelor metodei de inserare automată a diacriticelor

Lucrările au fost realizate în cadrul contractului
de cercetare UPB-CNCSIS, IDEI-782

RECTOR,
Prof. dr. ing. Ecaterina Andronescu

În Tabelul 1 sunt prezentate de asemenea valorile medii statistice ponderate ale *F-measure* pentru fiecare clasă de ambiguitate în care ponderile sunt calculate la nivel de caracter. Se observă că la nivel de caracter se obţine o valoare medie ponderată pentru *F-measure* de **99,34%**.

Din analiza datelor obţinute putem trage concluzia că etajul de filtrare cu bigrame aduce o îmbunătăţire substanţială de performanţă la nivel de caracter, pentru toate clasele de ambiguitate. În plus, etajul de filtrare cu trigrame cu sufixe aduce o îmbunătăţire evidentă numai pentru clasa de ambiguitate **a / ă / â**, ceea ce este un lucru pozitiv dacă ţinem cont de faptul că această clasă este sursa cea mai importantă de erori în procesul de refacere a diacriticelor. Pentru clasa **i / î** de ambiguitate rezoluţia metodei este cea mai bună. Explicaţia vine din faptul că, în limba română, **î** are o poziţionare quasi-regulată, putând să apară fie la început de cuvânt, fie în interiorul unor cuvinte compuse. Se poate merge chiar mai departe să se spună că pentru această clasă se poate oferi o metodă de refacere a diacriticelor de tip determinist.

După cum se observă, etajul trigramelor nu alterează rezultatele obţinute bune obţinute prin filtrarea cu bigrame pentru toate cele trei clase de ambiguitate.

Au fost efectuate şi alte experimente folosind trigramele obţinute la nivelul prefixelor, mai exact a caracterelor de început ale cuvintelor, dar această metodă nu a adus nici un plus evident de performanţă. Explicaţia vine de acolo că, prin natura inflexiunilor din limba română, prefixele apar mai degrabă independent de context.

Ca o primă concluzie se poate observa că cel mai problematic caracter, care aduce cel mai slab nivel de performanţă la refacerea diacriticelor, este reprezentat de clasa **a / ă / â**. În ciuda faptului că o mare parte dintre erori au fost corectate, rămân unele situaţii pe care algoritmul nu le tratează corect. Cele mai semnificative astfel de erori sunt prezentate în Tabelul 2. De aici se observă că cel mai mare număr de erori este adus de perechea ambiguă **ca – că** pe care o putem regăsi cu o frecvenţa mare de apariţie în corpusul de test. Această situaţie de ambiguitate este oarecum explicabilă dacă se ţine cont de faptul că există numeroase cazuri în care **ca** şi **că** apar în aproximativ aceleaşi contexte.

Totodată trebuie arătat faptul că există şi alte limitări pe care metoda nu le poate soluţiona, printre care pot fi enumerate:

1. Unele nume proprii, care conţin diacritice, sau unele cuvinte rare, care nu apar în dicţionarul *D*1. Soluţia poate fi cea de creştere a dicţionarului *D*1 prin adăugarea de noi cuvinte. Trebuie menţionat faptul că adăugarea de noi forme în *D*1 nu

Lucrările au fost realizate în cadrul contractului
de cercetare UPB-CNCSIS, IDEI-782

RECTOR,
Prof. dr. ing. Ecaterina Andronescu

poate duce la creşterea dicţionarelor de bigrame respectiv de trigrame. Aceste structuri au fost obţinute din corpusul de antrenare *A* iar pentru mărirea lor este necesar să îl extindem pe acesta şi cu alte contexte care să conţină noile forme introduse în *D*1.

2. Forme puternic dependente de context, la care nu există doar la nivelul frazei informaţii suficiente pentru a dezambiguiza varianta corectă. Ca exemplu se poate da situaţia ambiguă ***Am văzut o fată frumoasă la geam – Am văzut o faţă frumoasă la geam***. Este evident că pentru a clarifica o astfel de situaţie confuză sunt necesare informaţii suplimentare, la nivel de discurs.

Cuvânt	Bază de comparaţie	Bigrame şi unigrame	Bigrame, trigrame şi unigrame
ca / că	19,718	4,890	2,478
fată / faţă	2,199	912	495
sau / său	2,548	441	350
sa / să	30,064	395	341
lua / luă	547	330	272
uşa / uşă	1009	362	267
banca / bancă	507	248	243

Tabelul 2. Numărul de erori frecvente corectate gradual de metodă

Lucrările au fost realizate în cadrul contractului
de cercetare UPB-CNCSIS, IDEI-782

RECTOR,
Prof. dr. ing. Ecaterina Andronescu

REVENDICĂRI

Metodă de inserare automată a semnelor diacritice pentru sinteza vorbirii pornind de la text în limba română în aplicaţii ce au ca suport reţele de date, caracterizată prin aceea că la intrarea unui sistem de sinteză a vorbirii destinat unor aplicaţii cum ar fi cele de citire a mesajelor de tip text (e-mail, SMS etc.) pe telefoane mobile, se verifică mai întâi dacă textul de intrare conţine caractere scrise cu diacritice, iar dacă acest lucru nu se întâmplă, metoda se aplică prin intermediul unui algoritm eficient de inserare automată a semnelor diacritice.

Metoda are la bază o strategie statistică axată pe *n*-grame şi constă în esenţă într-un proces de *filtrare secvenţială*, pe mai multe niveluri, bazat pe probabilitatea de apariţie a cuvintelor scrise cu diacritice şi pe contextul în care acestea apar. Fiecare criteriu de filtrare aduce o îmbunătăţire suplimentară, fără însă a mări caracterul de ambiguitate (deci formele posibile teoretic de scriere a unui cuvânt cu sau fără diacritice). Astfel, după ce tratează iniţial cuvintele fără diacritice sau pentru care nu există ambiguităţi, metoda utilizează incremental trei etaje de filtrare: la nivel de *unigrame* (pe cuvânt), la nivel de *bigrame* (de asemenea pe cuvânt) şi la nivel de *trigrame* pentru terminaţii de cuvânt, informaţia contextuală fiind adăugată gradual, după fiecare etaj, simultan cu eliminarea variantelor incorecte. Acest lucru se realizează practic prin folosirea probabilităţilor *n*-gramelor, calculate pe baza frecvenţelor lor de apariţie obţinute din corpusul de antrenare.

RECTOR,
Prof. dr. ing. Ecaterina Andronescu