(19) **OFICIUL DE STAT PENTRU INVENŢII ŞI MĂRCI**
Bucureşti

ROMÂNIA

(11) **RO 130875 B1**

(51) **Int.Cl.**
G06F 17/21 (2006.01);
G06F 17/27 (2006.01)

(12) **BREVET DE INVENŢIE**

(21) Nr. cerere: **a 2014 00346**

(22) Data de depozit: **07/05/2014**

(45) Data publicării menţiunii acordării brevetului: **30/09/2020** BOPI nr. **9/2020**

(41) Data publicării cererii:
29/01/2016 BOPI nr. **1/2016**

(73) Titular:
• **PETRICĂ LUCIAN**, *INTRAREA VÂSLEI NR. 1, BL. PM63, SC. 2, ET. 5, AP. 77, SECTOR 3, BUCUREŞTI, B, RO;*
• **CUCU HORIA**, *ALEEA POLITEHNICII NR. 6, BL. 3, SC. 4, ET. 2, AP. 42, SECTOR 6, BUCUREŞTI, B, RO;*
• **BUZO ANDI**,
BD. GENERAL VASILE MILEA NR. 8, BL. B2, SC. 3, ET. 4, AP.57, SECTOR 6, BUCUREŞTI, B, RO

(72) Inventatori:
• **PETRICĂ LUCIAN**, *INTRAREA VÂSLEI NR. 1, BL. PM63, SC. 2, ET. 5, AP. 77, SECTOR 3, BUCUREŞTI, B, RO;*
• **CUCU HORIA**, *ALEEA POLITEHNICII NR. 6, BL. 3, SC. 4, ET. 2, AP. 42, SECTOR 6, BUCUREŞTI, B, RO;*
• **BUZO ANDI**,
BD. GENERAL VASILE MILEA NR. 8, BL. B2, SC. 3, ET. 4, AP.57, SECTOR 6, BUCUREŞTI, B, RO

(74) Mandatar:
CABINET D.NICOLAESCU, *STR.TURDA, NR.102, BL.30A, ET.7, AP.28, BUCUREŞTI*

(56) Documente din stadiul tehnicii:
"ASR domain adaptation methods for low-resourced languages:Application to Romanian language" in Procedings of the 20th IEEE European Signal Processing Cnference (EUSIPCO), H. Cucu, L. Besacier, A. Buzo, pp. 1648-1652: US8543382B2: US7966173B2

(54) METODĂ PENTRU RESTAURAREA AUTOMATĂ A SEMNELOR DIACRITICE FOLOSIND TEXTE ACHIZIŢIONATE ELECTRONIC, UTILIZATĂ ÎN PROCESAREA LIMBAJULUI NATURAL

Examinator: **ing. ANCA POPESCU**

RO 130875 B1

Invenţia se referă la o metodă de restaurare automată a semnelor diacritice într-un corpus de text format din fişiere multiple, pus la dispoziţie de utilizator, destinată a fi utilizată în domeniul sistemelor de procesare a limbajului natural.

Invenţia aparţine domeniului sistemelor de procesare a limbajului natural, sub formă vorbită sau scrisă, care includ sistemele de corectare automată a resurselor text, recunoaştere automată a vorbirii şi sinteză automată a vorbirii.

Semnele diacritice sunt utilizate în multe limbi (franceză, spaniolă, română etc.) pentru extinderea alfabetului dincolo de caracterele latine. Limba română utilizează trei semne diacritice, care sunt folosite pentru construirea a cinci caractere speciale: ă, â, î, ş, ţ. Semnele diacritice modifică pronunţia celor patru caractere de bază: a, i, s, t. Literatura ştiinţifică a domeniului raportează că în textele din limba română, până la 40% din cuvinte folosesc semne diacritice.

Din diverse motive, care includ lipsa unor tastaturi dedicate pentru multe din limbile care utilizează diacritice, din majoritatea documentelor text în format electronic lipsesc semnele diacritice. De cele mai multe ori, caracterele speciale ce utilizează semne diacritice sunt înlocuite de caracterele de bază corespunzătoare. Acest fapt duce la scăderea inteligibilităţii textului scris în format electronic, dar şi la apariţia unor cuvinte greşite din punct de vedere gramatical.

Corectarea acestor texte poate fi automatizată, folosind un sistem informatic de restaurare a semnelor diacritice. Pentru cuvintele care au o singură formă diacritică, restaurarea este trivială şi se poate face pe baza unui dicţionar. Un exemplu ar fi restaurarea diacriticelor cuvântului câine, a cărui singură formă corectă este câine. În cazul în care există mai multe forme cu diacritice ale unui cuvânt, restaurarea necesită folosirea contextului pentru alegerea variantei corecte de restaurare. De exemplu, în limba română formele articulate şi nearticulate ale multor cuvinte diferă doar prin diacritice: maşină şi maşina sunt forme corecte, cu semne diacritice, ale cuvântului de bază maşina, dar alegerea uneia din cele două variante la restaurare necesită cunoaşterea măcar a cuvântului anterior celui restaurat, în propoziţie.

Un astfel de sistem de restaurare utilizează un model statistic de limbă ce specifică probabilităţile de apariţie ale fiecărei forme diacritice, dat fiind cuvintele anterioare cuvântului de bază în textul restaurat. Un model statistic de limbă este construit (antrenat) folosind cantităţi mari de text de înaltă calitate.

În acest context, un text de înaltă calitate este acel text care conţine un număr mic de erori din punctul de vedere al utilizării semnelor diacritice.

Se cunosc sistemele de restaurare a diacriticelor folosind abordări statistice.

În [D. Burileanu, C. Ungurean, "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", cerere de brevet de invenţie RO 127582 A2] se descrie un sistem de restaurare a semnelor diacritice care necesită: (i) un dicţionar al celor mai frecvent folosite cuvinte din limba română şi a formelor flexionate ale acestora, şi (ii) un corpus de text care conţine forme corecte scrise cu diacritice.

În literatura de specialitate, au fost propuse diverse metode pentru restaurarea semnelor diacritice, dintre care cele mai notabile sunt [C. Ungurean and D. Burileanu, "An advanced NLP framework for high quality Text-to-Speech synthesis", 6th IEEE Conference on Speech Technology and Human-Computer Dialogue (SpeD), pp. 1-6, 2011] care prezintă concepte similare cu cererea de brevet menţionată anterior şi raportează o performanţă de 1,4% eroare la nivel de cuvânt şi 0,4% eroare la nivel de caracter. O altă lucrare ştiinţifică relevantă este [H. Cucu, L. Besacier, C. Burileanu, and A. Buzo, 'ASR domain adaptation methods for low-resourced languages: Application to Romanian language", in Proceedings of the 20th IEEE

European Signal Processing Conference (EUSIPCO), pp. 1648-1652, 2012] care foloseşte pentru restaurarea semnelor diacritice un model statistic de limbă de tip 3-grame, antrenat folosind un corpus de text corectat manual. Autorii raportează o eroare de 1,99% la nivel de cuvânt şi 0,48% la nivel de caracter.

Prezenta invenţie se referă la o metodă de restaurare automată a semnelor diacritice într-un corpus de text format din fişiere multiple, pus la dispoziţie de utilizator, metoda cuprinzând etapele de împărţire automată a corpus-ului în secţiuni de calitate înaltă, respectiv scăzută, folosind un prag fix de frecvenţă de apariţie a semnelor diacritice ca principal criteriu de filtrare şi, utilizare a secţiunilor de calitate înaltă pentru antrenarea unui model probabilistic de limbă care este folosit pentru restaurarea semnelor diacritice în secţiunile de calitate scăzută; metoda mai poate cuprinde o etapă de determinare automată a pragului optim de frecvenţă de apariţie a semnelor diacritice, prin repetarea iterativă a primelor două etape enunţate mai sus şi ajustarea pragului de filtrare la fiecare iteraţie, prin variere între zero şi o valoare specificată de utilizator, în vederea minimizării erorii de restaurare a semnelor diacritice, calculată folosind un text de evaluare pus la dispoziţie de utilizator, pragul optim fiind considerat acela care corespunde erorii minime de restaurare şi putând fi identificat speculativ.

În continuare, se prezintă detaliat, în legătură şi cu fig. de la 1 la 4, principiile şi realizarea invenţiei şi un exemplu de aplicare a sa constând în utilizarea metodei pentru corectarea unui corpus de ştiri în limba română, achiziţionate de pe un portal online, şi generarea unui model statistic de limbă pentru restaurarea semnelor diacritice în alte texte din limba română. Fig. 1... 4 reprezintă:

- fig. 1, metoda de creştere a calităţii unui corpus de text;
- fig. 2, algoritmul de filtrare folosind un prag fix de separare;
- fig. 3, algoritmul de identificare a pragului optim de separare;
- fig. 4, algoritmul de identificare speculativă a pragului optim de separare.

Invenţia se referă la o metodă pentru curăţarea unui corpus de text (**CT**) **101**, achiziţionat din surse predispuse la erori legate de semnele diacritice (de exemplu, de pe internet), folosind un sistem de restaurare a semnelor diacritice (**SRD**) **102**, care se bazează pe un model statistic de limbă (**ML**) **103**, generat prin procesarea **CT 101**. Invenţia are la bază faptul că generarea modelului statistic de limbă **103** se face pornind de la **CT 101**, fără a avea garanţia corectitudinii textului din **CT 101** în ceea ce priveşte utilizarea semnelor diacritice. Metoda are ca rezultat un corpus augmentat de text (**CTA**) **104**, de calitate ridicată, ce poate fi folosit pentru aplicaţii legate de recunoaşterea vorbirii, sinteza vorbirii şi alte procesări de limbaj natural.

Pentru a asigura calitatea modelului statistic de limbă **103**, textul **CT 101** este procesat de un filtru de separare **105**, folosind anumite criterii de calitate, concretizate în valoarea unui prag de separare **106**. În metoda propusă, acest prag de separare poate fi calculat automat prin metoda de selecţie a pragului de separare **107**, folosind un corpus de evaluare (**CE**) **108**, sau furnizat direct de către utilizator. Filtrarea împarte textul **101** în două categorii:

- text ce conţine cu probabilitate ridicată semne diacritice utilizate corect, care formează corpusul de calitate înaltă (**CI**) **109**;
- text care, cu probabilitate ridicată, conţine greşeli din punctul de vedere al utilizării semnelor diacritice, care formează corpusul de calitate redusă (**CR**) **110**

În urma separării corpusului de text **CT 101** în corpusurile de calitate înaltă **CI 109**, respectiv redusă **CR 110**, **CI 109** este folosit pentru generarea modelului probabilistic de limbă **103**. Acest model este generat folosind un bloc de analiză statistică **111**. Apoi, modelul de limbă **103** este utilizat la restaurarea semnelor diacritice în **CR 110**, folosind **SRD 102**, bazat pe metodologia standard enunţată în literatura de specialitate şi stadiul tehnicii. Textul obţinut din restaurarea semnelor diacritice asupra **CR 110** este adăugat la **CI 109**, rezultând un corpus final, **CTA 104**, de dimensiune egală cu **CT 101**, dar de calitate superioară.

Un element esenţial al invenţiei este algoritmul de filtrare folosit de filtrul de separare **105**, care separă corpusul în două secţiuni de calitate înaltă **109**, respectiv scăzută **110**. Invenţia propune o metodă de separare bazată pe observaţii statistice asupra limbilor ce utilizează semne diacritice. În aceste limbi, frecvenţa de apariţie a semnelor diacritice în text este relativ constantă, indiferent de natura subiectului descris. Astfel, este de aşteptat ca dintr-o analiză asupra unui text achiziţionat online, de dimensiuni rezonabile (minim sute de cuvinte), să reiasă o frecvenţă de utilizare a semnelor diacritice apropiată de media generală pentru limba textului respectiv.

În practică, multe din textele achiziţionate online nu folosesc semne diacritice sau le folosesc doar parţial, având o frecvenţă a diacriticelor mult sub media limbii. Filtrul de separare **105** foloseşte această proprietate pentru identificarea fişierelor text de calitate redusă. Fig. 2 prezintă filtrul **105** în detaliu. Atât timp cât mai există fişiere în **CT 101**, blocul **201** de extragere de fişiere extrage un fişier text **202** din **CT 101** şi livrează acest fişier către blocul **203** de calculare a frecvenţei semnelor diacritice. Dacă frecvenţa calculată este mai mică decât pragul de separare **106**, fişierul de text se consideră a fi de calitate redusă şi este adăugat la corpus-ul **CR 110** de către blocul de adăugare **204**. Dacă frecvenţa depăşeşte acest prag, fişierul se consideră a fi de calitate ridicată şi este adăugat la corpus-ul **CI 109** de către blocul de adăugare **205**. Procesul se încheie atunci când nu mai există fişiere în **CT 101**.

Alegerea pragului de separare **106** determină calitatea corpus-ului final **104**. Dacă pragul de separare este ales cu o valoare prea mică, fişiere de calitate redusă vor fi considerate a fi de înaltă calitate, scăzând calitatea totală a **CI 109** şi, prin urmare, a modelului statistic de limbă **103**. Alternativ, dacă pragul de separare este ales cu o valoare prea mare, numărul fişierelor considerate a fi de calitate înaltă va fi mic, reducând dimensiunea **CI 109**, ceea ce duce la scăderea calităţii modelului statistic de limbă **103**.

În prezenta invenţie, metoda prezentată anterior este completată şi cu o etapă de selecţie automată a pragului optim de separare. Algoritmul de selecţie automată este prezentat în fig. 3. Algoritmul menţine şi ajustează iterativ un prag temporar **301**, care este iniţial setat la o valoare minimă, specificată de utilizator, în blocul **302**. Pragul temporar este folosit pentru filtrarea **CT 101** şi generarea modelului de limba **111**. Sistemului din fig. 1 i se adaugă modulul **301** de evaluare a modelului statistic de limbă folosit pentru restaurarea semnelor diacritice. Acest modul foloseşte corpus-ul de evaluare **108**, de dimensiuni relativ mici şi calitate bună pentru a determina eroarea de restaurare. Pragul de separare temporar **301** este înregistrat ca optim **304** dacă eroarea de evaluare este mai mică decât a tuturor iteraţiilor anterioare. Pragul temporar este incrementat în blocul **305** cu un increment specificat de utilizator. Pentru fiecare valoare a pragului de separare se execută paşii de filtrare, generare a modelului statistic de limbă folosind corpus-ul CI şi evaluare a erorii de restaurare, folosind acest model de limbă. La final, pragul optim identificat **304** este folosit ca prag de separare **106**.

Un potenţial dezavantaj al metodei prezentate anterior este faptul că filtrarea se execută de un număr mare de ori, pentru fiecare variantă posibilă a pragului de separare, între valoarile sale minimă, respectiv maximă. În funcţie de dimensiunea corpus-ului **101**, timpul de execuţie a acestui proces pe un calculator personal poate deveni foarte mare.

Pentru reducerea acestui timp de execuţie, metoda prezentată anterior este completată şi cu un mecanism de întrerupere speculativă a căutării pragului optim de separare. Decizia întreruperii căutării se bazează pe observaţia că eroarea de evaluare va scădea, odată cu creşterea valorii pragului de separare, către o valoare minimă, după care va creşte pe măsură ce pragul se incrementează şi valoarea sa se îndepărtează de pragul optim. Astfel, se poate considera că, odată ce eroarea de evaluare a crescut cu o anumită valoare faţă de eroarea minimă înregistrată anterior, căutarea poate fi oprită. Fig. 4 prezintă algoritmul cu întrerupere speculativă.

Metoda care face obiectul prezentei invenţii prezintă avantajul că permite restaurarea semnelor diacritice în textul de intrare, şi generarea unui model de limbă ce poate fi folosit pentru restaurarea diacriticelor în alte texte, fără a necesita resurse de text adiţionale, de înaltă calitate (ridicată). Pentru multe din limbile ce utilizează semne diacritice, nu există astfel de resurse de text de înaltă calitate, singurele resurse disponibile fiind cele ce pot fi achiziţionate online, a căror calitate nu poate fi garantată.

Metoda conform invenţiei mai are avantajul că permite dezvoltarea de aplicaţii de procesare a limbajului natural sub formă textuală şi vorbită, pentru limbi cu resurse subdezvoltate, unde majoritatea resurselor text sunt achiziţionate online. Metoda nu este specifică limbii române, ci poate fi aplicată oricărei limbi care utilizează semne diacritice.

În cele ce urmează, vom exemplifica utilizarea metodei propuse, împreună cu optimizările destinate selecţiei pragului de filtrare, pentru curăţarea greşelilor de folosire a diacriticelor în limba română, comparând rezultatele cu starea tehnicii. Metoda este aplicată asupra unui corpus de text care constă în articole de ştiri, achiziţionat online de pe www.libertatea.ro. Parametrii folosiţi pentru evaluarea metodei sunt următorii:

- TC conţine 325.000 de fişiere şi peste 70 de milioane de cuvinte în total. Din totalul fişierelor, aproximativ 153.000 nu conţin deloc diacritice, iar 6000 de fişiere conţin diacritice în mod sporadic.
- Pentru evaluarea filtrării cu prag fix furnizat de utilizator, s-a folosit un prag de separare cu valoarea de 20% (frecvenţa semnelor diacritice în text)
- Pentru evaluarea metodelor de selecţie automată a pragului de separare, s-a folosit un corpus de evaluare de 4 milioane de cuvinte, verificate şi corectate manual.
- Pentru metodele de selecţie automată a pragului de separare, s-a folosit o gamă de valori a pragului de separare între 0 şi 25%, în increment de 1%
- Pentru metoda speculativă de oprire a căutării pragului optim, s-a decis oprirea căutării atunci când eroarea creşte cu 5% faţă de minim.

Comparaţia cu stadiul tehnicii s-a făcut folosind ca referinţă sistemul de restaurare a diacriticelor din [H. Cucu, L. Besacier, C. Buriieanu, and A. Buzo, "ASR domain adaptation methods for low-resourced languages: Application to Romanian language", in Proceedings of the 20th IEEE European Signal Processing Conference (EUSIPCO), pp. 1648-1652, 2012.].

Tabelul 1 prezintă rezultatele evaluării metodelor propuse şi comparaţia cu stadiul tehnicii.

Comparaţie cu stadiul tehnicii

Tabelul 1

Metoda	Cantitate necesară de text de calitate (10^6 cuvinte)	Eroare de restaurare (%)	Prag de separare identificat (%)	Timp de găsire a pragului de separare (ore)
Stadiul tehnicii	15	1,13	-	-
Prag fix	0	1,11	20	0
Căutare exhaustivă	4	0,92	13	16
Căutare speculativă	4	0,92	13	10

Se obsevă că metoda propusă duce la restaurarea semnelor diacritice cu eroare mai mică decât stadiul tehnicii, în toate cele 3 cazuri evaluate. Metoda cu prag fix de evaluare duce la o eroare de restaurare de 1,11%, având avantajul timpului de rulare minim, datorită faptului că nu necesită găsirea pragului de separare. De asemenea, această metodă nu necesită text de calitate.

Ambele metode de căutare automată a pragului optim de separare duc la obţinerea unor rezultate mai bune din punctul de vedere al erorii de restaurare. În cazul ambelor metode, se obţine o scădere de 17% a erorii, identificându-se pragul de separare optim la 13%. Dintre cele două metode de căutare automată a pragului optim de separare, căutarea speculativă are timpul de rulare cel mai mic, 10 ore comparativ cu 16 ore în cazul căutării exhaustive.

Revendicări

1. Metodă pentru restaurarea automată a semnelor diacritice folosind texte achiziţionate electronic, utilizată în procesarea limbajului natural, implementată ca program software executabil pe un calculator sau sub formă de circuit integrat, **caracterizată prin aceea că** această metodă cuprinde etapele de:

- achiziţionarea online a unui corpus de text (**101**) format din fişiere multiple, din surse disponibile, ce pot fi afectate de erori legate de semnele diacritice,
- împărţirea automată a corpusului de text (**101**) într-o secţiune de calitate înaltă (**109**) şi o secţiune de calitate redusă (**110**), prin procesarea fiecărui fişier în ceea ce priveşte diacriticele, prin intermediul unui filtru de separare (**105**) folosind drept criteriu de filtrare un prag de separare (**106**) corespunzător frecvenţei de apariţie a semnelor diacritice, prag ce poate fi calculat automat folosind un corpus de evaluare (**108**) sau care exprimă în procente frecvenţa de apariţie a semnelor diacritice în limba respectivă,
- utilizarea secţiunii de calitate înaltă (**109**) pentru antrenarea unui model probabilistic de limbă (**103**), generat prin intermediul unui bloc de analiză statistică (**111**), şi
- utilizarea modelului probabilistic de limbă (**103**), astfel antrenat, pentru restaurarea semnelor diacritice a secţiunii de calitate redusă (**110**), obţinându-se în final un corpus de text augmentat (**104**) de calitate ridicată, ce poate fi folosit pentru aplicaţii legate de procesări de limbaj natural, în special recunoaşterea şi sinteza vorbirii.

2. Metodă conform revendicării 1 **caracterizată prin aceea că**, pentru determinarea automată a unei valori optime a pragului de separare (**106**), etapele metodei de la revendicarea 1 sunt repetate iterativ, pornind de la un prag temporar (**301**) setat la o valoare minimă şi ajustarea incrementală a acesteia la fiecare iteraţie, până la o valoare maximă, valoarea optimă a pragului de separare (**106**) corespunzând erorii minime de restaurare a semnelor diacritice, calculată folosind corpusul de evaluare (**108**).

3 .Metodă conform revendicării 2 **caracterizată prin aceea că** pentru reducerea timpului de determinare automată a valorii optime a pragului de separare (**106**), procesul de iterare este oprit atunci când eroarea de restaurare a iteraţiei curente depăşeşte cu un procent prestabilit valoarea minimă a erorii de restaurare înregistrată pentru iteraţiile anterioare, pragul optim fiind considerat cel care corespunde respectivei erori minime de restaurare.

4. Sistem informatic pentru restaurarea automată a semnelor diacritice folosind texte achiziţionate electronic, utilizat în procesarea limbajului natural, **caracterizat prin aceea că** acesta cuprinde un circuit integrat configurat pentru realizarea etapelor metodei din revendicarea 1.

(51) Int.Cl.
G06F 17/21 (2006.01);
G06F 17/27 (2006.01)

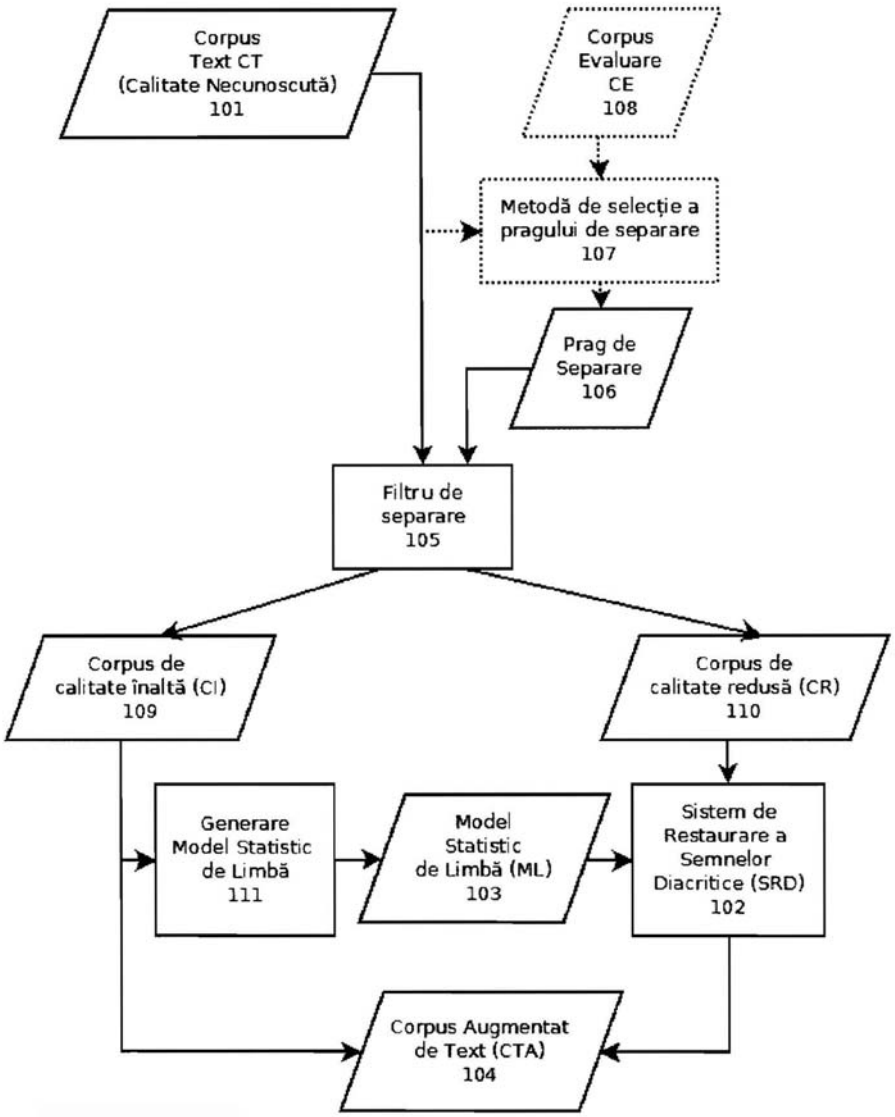

Fig. 1

(51) Int.Cl.
G06F 17/21 (2006.01);
G06F 17/27 (2006.01)

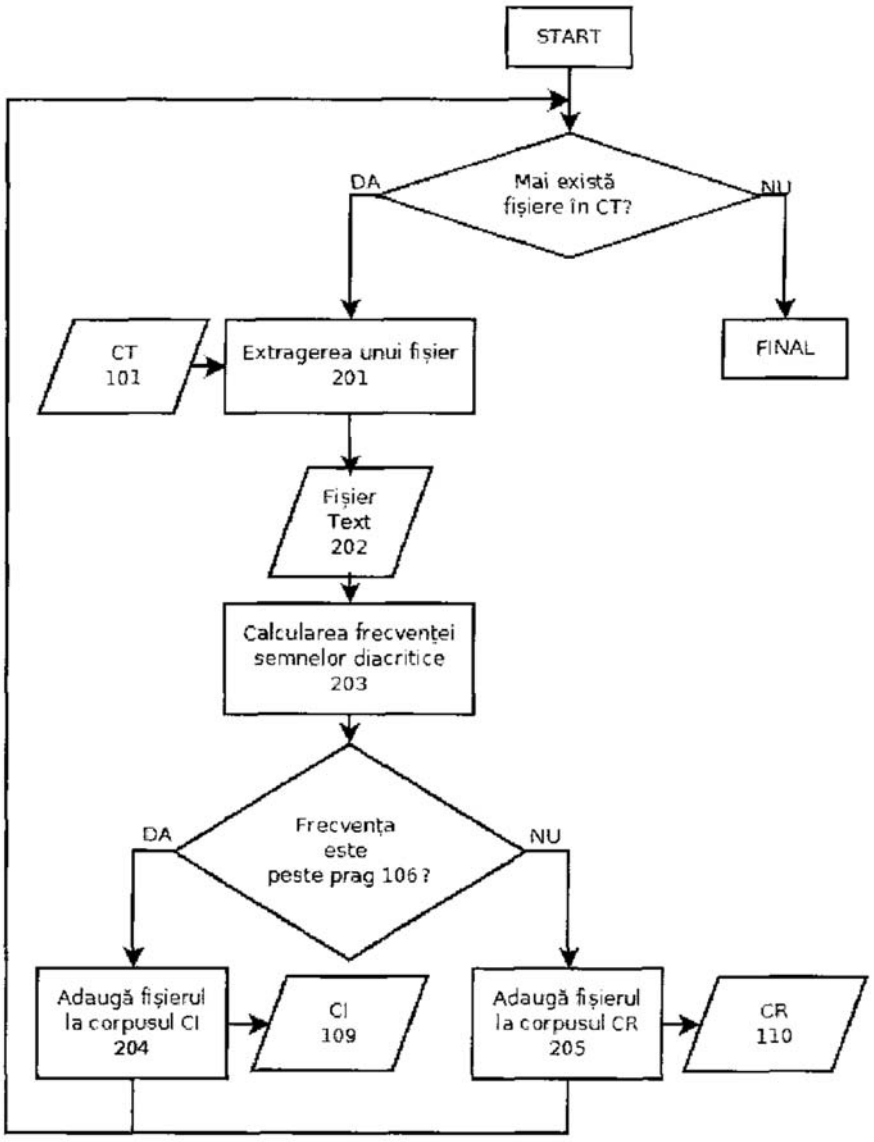

Fig. 2

(51) Int.Cl.
G06F 17/21 (2006.01);
G06F 17/27 (2006.01)

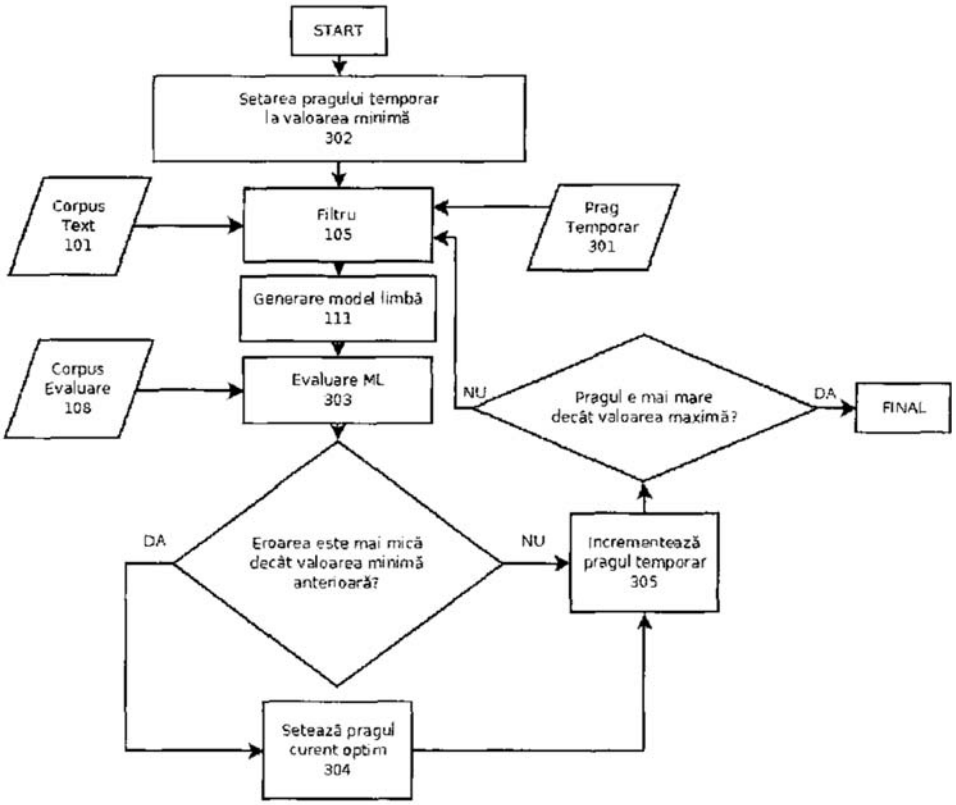

Fig. 3

(51) Int.Cl.
G06F 17/21 (2006.01);
G06F 17/27 (2006.01)

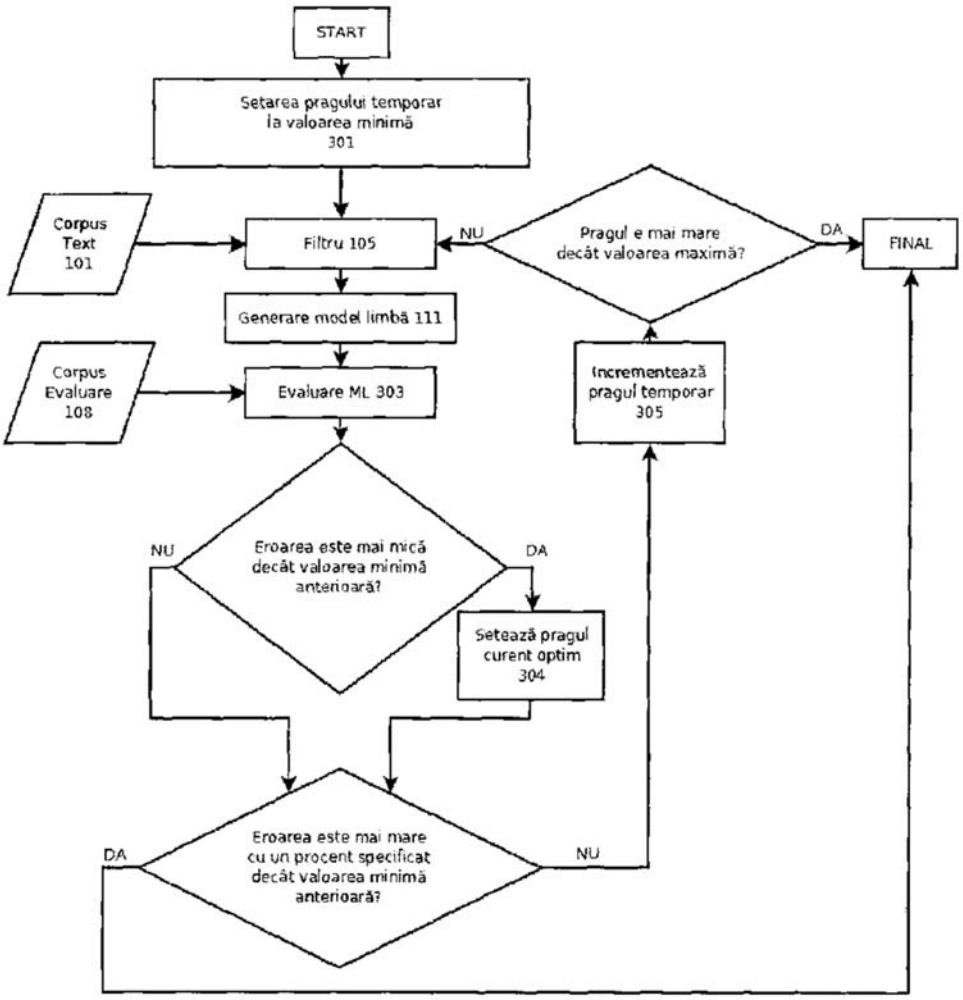

Fig. 4

Editare şi tehnoredactare computerizată - OSIM
Tipărit la: Oficiul de Stat pentru Invenţii şi Mărci
sub comanda nr. 387/2020