ROMÂNIA

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

(11) **RO 130875 A2**

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

(12) # CERERE DE BREVET DE INVENȚIE

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

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

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

(71) Solicitant:
- **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*

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

(57) **Rezumat:**

Invenția se referă la o metodă de restaurare automată a semnelor diacritice într-un corpus de text format din fişiere multiple, pus la dispoziție de către utilizator, destinată a fi utilizată în domeniul sistemelor de procesare a limbajului natural. Metoda conform invenției constă în împărțirea automată a corpusului de text în secțiuni de calitate înaltă, respectiv, scăzută, folosind un prag fix de frecvență de apariție a semnelor diacritice, şi în utilizarea secțiunilor de calitate înaltă, pentru antrenarea unui model probabilistic de limbă, folosit pentru restaurarea semnelor diacritice în secțiunile de calitate scăzută.

Revendicări: 3
Figuri: 4

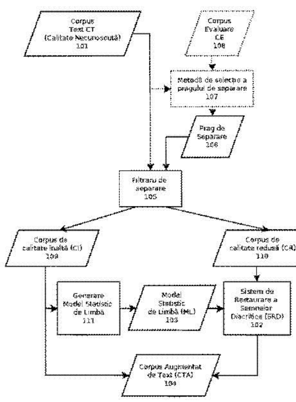

Fig. 1

RO 130875 A2

39

METODĂ PENTRU RESTAURAREA AUTOMATĂ A SEMNELOR DIACRITICE, FOLOSIND TEXTE ACHIZIȚIONATE ELECTRONIC, UTILIZATĂ ÎN PROCESAREA LIMBAJULUI NATURAL

Invenția aparține 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 pronunția celor patru caractere de bază: a, i, s, t. Literatura ştiințifică 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 inteligibilității textului scris în format electronic, dar şi la apariția 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 dicționar. Un exemplu ar fi restaurarea diacriticelor cuvântului **caine,** 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 **masina**, dar alegerea uneia din cele două variante la restaurare necesită cunoaşterea măcar a cuvântului anterior celui restaurat, în propoziție.

Un astfel de sistem de restaurare utilizează un model statistic de limbă ce specifică probabilitățile de apariție ale fiecărei forme diacritice, dat fiind cuvintele anterioare cuvântului de bază în textul restaurat. Un model statistic de limbă este construit (antrenat) folosind cantități mari de text de înaltă calitate. În acest context, un text de înaltă calitate este acel text care conține 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 aplicații ce au ca suport rețele de date", cerere de brevet de invenție RO 127582 A2] se descrie un sistem de restaurare a semnelor diacritice care necesită: (i) un dicționar al celor mai frecvent folosite cuvinte din limba română şi a formelor flexionate ale acestora, şi (ii) un corpus de text care conține 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 menționată anterior şi raportează o performanță de 1.4% eroare la nivel de cuvânt şi 0.4% eroare la nivel de caracter. O altă lucrare ştiințifică 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 invenție se referă la o metodă de restaurare automată a semnelor diacritice într-un corpus de text format din fişiere multiple, pus la dispoziție de utilizator, metoda cuprinzând etapele de împărțire automată a corpus-ului în secțiuni de calitate înaltă, respectiv scăzută, folosind un prag fix de frecvență de apariție a semnelor diacritice ca principal criteriu de filtrare şi, utilizare a secțiunilor de calitate înaltă pentru antrenarea unui model probabilistic de limbă care este folosit pentru restaurarea semnelor diacritice în secțiunile de calitate scăzută; metoda mai poate cuprinde o etapă de determinare automată a pragului optim de frecvență de apariție a semnelor diacritice, prin repetarea iterativă a primelor două etape enunțate mai sus şi ajustarea pragului de filtrare la fiecare iterație, 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 dispoziție 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 figurile de la 1 la 4, principiile şi realizarea invenției şi un exemplu de aplicare a sa constând în utilizarea metodei pentru corectarea unui corpus de ştiri în limba română, achiziționate de pe un portal online, şi generarea unui model statistic de limbă pentru restaurarea semnelor diacritice în alte texte din limba română. Figurile 1 – 4 reprezintă:

Fig. 1 - metoda de creştere a calității 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

Invenția se referă la o metodă pentru curățarea unui corpus de text (CT) 101, achiziționat 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. Invenția are la bază faptul că generarea modelului statistic de limbă 103 se face pornind de la CT 101, fără a avea garanția 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 aplicații 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 selecție 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 conține cu probabilitate ridicată semne diacritice utilizate corect, care formează corpusul de calitate înaltă (CI) 109
- text care, cu probabilitate ridicată, conține 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 enunțată în literatura de specialitate şi stadiul tehnicii. Textul obținut 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 esențial al invenției este algoritmul de filtrare folosit de filtrul de separare 105, care separă corpusul în două secțiuni de calitate înaltă 109, respectiv scăzută 110. Invenția propune o metodă de separare bazată pe observații statistice asupra limbilor ce utilizează semne diacritice. În aceste limbi, frecvența de apariție a semnelor diacritice în text este relativ constantă, indiferent de natura subiectului descris. Astfel, este de aşteptat ca dintr-o analiză asupra unui text achiziționat online, de dimensiuni rezonabile (minim sute de cuvinte), să reiasă o frecvență de utilizare a semnelor diacritice apropiată de media generală pentru limba textului respectiv.

În practică, multe din textele achiziționate online nu folosesc semne diacritice sau le folosesc doar parțial, având o frecvență a diacriticelor mult sub media limbii. Filtrul de separare 105 foloseşte această proprietate pentru identificarea fişierelor text de calitate redusă. Figura 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 frecvenței semnelor diacritice. Dacă frecvența 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 frecvența 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 calității modelului statistic de limbă 103.

În prezenta invenție, metoda prezentată anterior este completată şi cu o etapă de selecție automată a pragului optim de separare. Algoritmul de selecție automată este prezentat în Figura 3. Algoritmul menține şi ajustează iterativ un prag temporar 301, care este inițial 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

Figura 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 valorile 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 inregistrată anterior, căutarea poate fi oprită. Figura 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 sub-dezvoltate, 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, achiziționat online de pe www.libertatea.ro. Parametrii folosiți pentru evaluarea metodei sunt următorii:

- TC conține 325.000 de fişiere şi peste 70 de milioane de cuvinte în total. Din totalul fişierelor, aproximativ 153.000 nu conțin deloc diacritice, iar 6000 de fişiere conțin diacritice în mod sporadic.
- Pentru evaluarea filtrării cu prag fix furnizat de utilizator, s-a folosit un prag de separare cu valoarea de 20% (frecvența semnelor diacritice în text)
- Pentru evaluarea metodelor de selecție automată a pragului de separare, s-a folosit un corpus de evaluare de 4 milioane de cuvinte, verificate şi corectate manual.
- Pentru metodele de selecție 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% față de minim.

Comparația cu stadiul tehnicii s-a făcut folosind ca referință sistemul de restaurare a diacriticelor din [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.].

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

Tabel 1: Comparație cu stadiul tehnicii

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 obținerea unor rezultate mai bune din punctul de vedere al erorii de restaurare. În cazul ambelor metode, se obține 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 de restaurare automată a semnelor diacritice într-un corpus de text format din fişiere multiple, pus la dispoziţie de utilizator, metoda fiind **caracterizată prin** etapele de:

   împărţirea automată a corpus-ului în secţiuni de calitate ridicată, respectiv scăzută, folosind un prag fix de frecvenţă de apariţie a semnelor diacritice ca principal criteriu de filtrare, şi

   utilizarea secţiunilor de calitate ridicată pentru antrenarea unui model probabilistic de limbă care este folosit pentru restaurarea semnelor diacritice în secţiunile de calitate scăzută.

2. Metodă conform revendicării 1**, caracterizată prin aceea că** mai cuprinde şi o etapă de determinare automată a pragului optim de frecvenţă de apariţie a semnelor diacritice, prin repetarea iterativă a etapelor ~~de~~ metodei din revendicarea 1 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.

3. Metodă conform revendicării 2, **caracterizată prin aceea că** mai cuprinde şi o etapă de identificare speculativă a unui prag optimal de filtrare, prin oprirea prematură a procesului de iterare atunci când eroarea de restaurare a iteraţiei curente depăşeşte, cu un procent setat de utilizator, valoarea minimă a erorii de restaurare înregistrată pentru iteraţiile anterioare, pragul optimal fiind considerat cel care corespunde respectivei erori minime de restaurare.

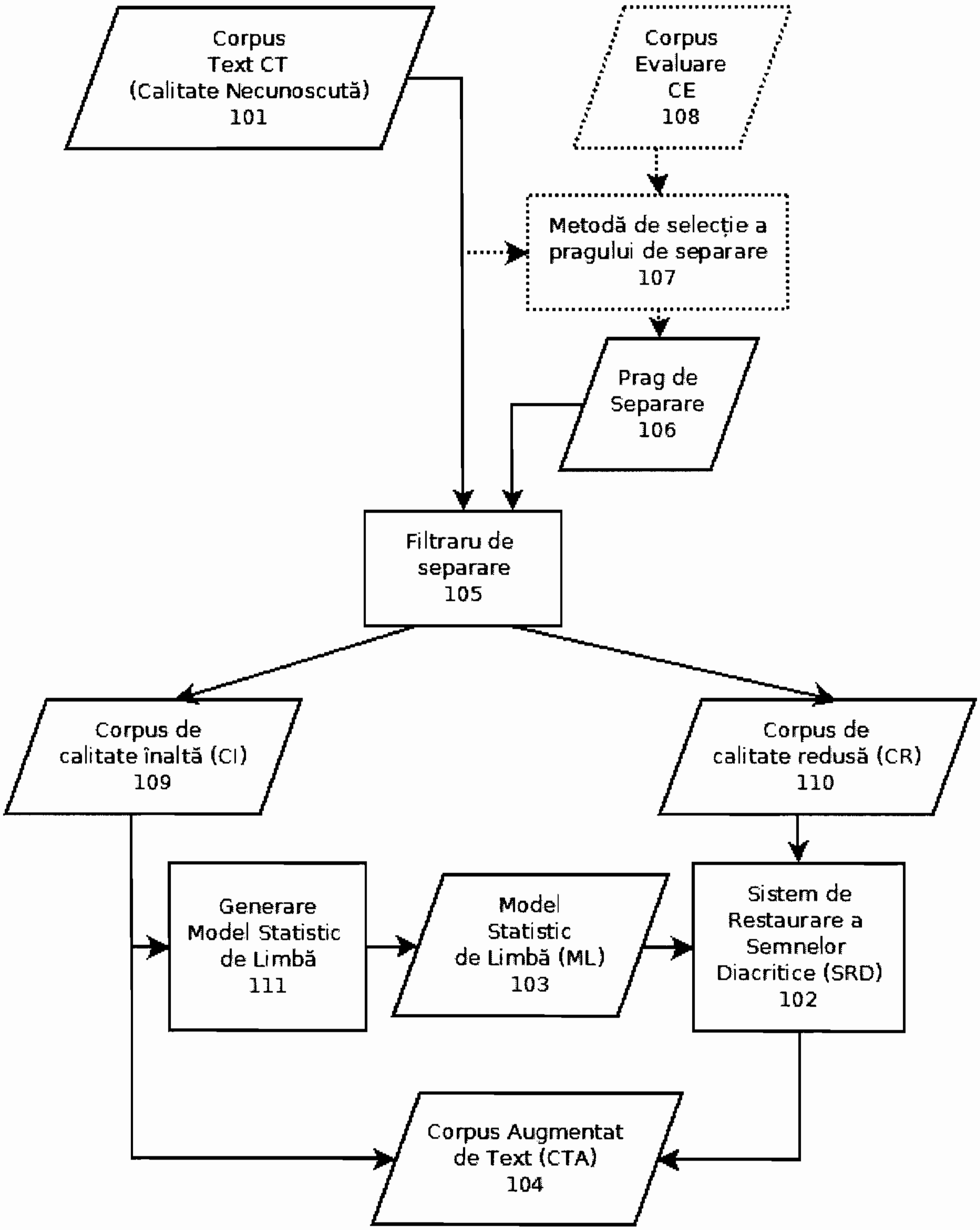

Figura 1: Metodologia de creştere a calității unui corpus de text

30

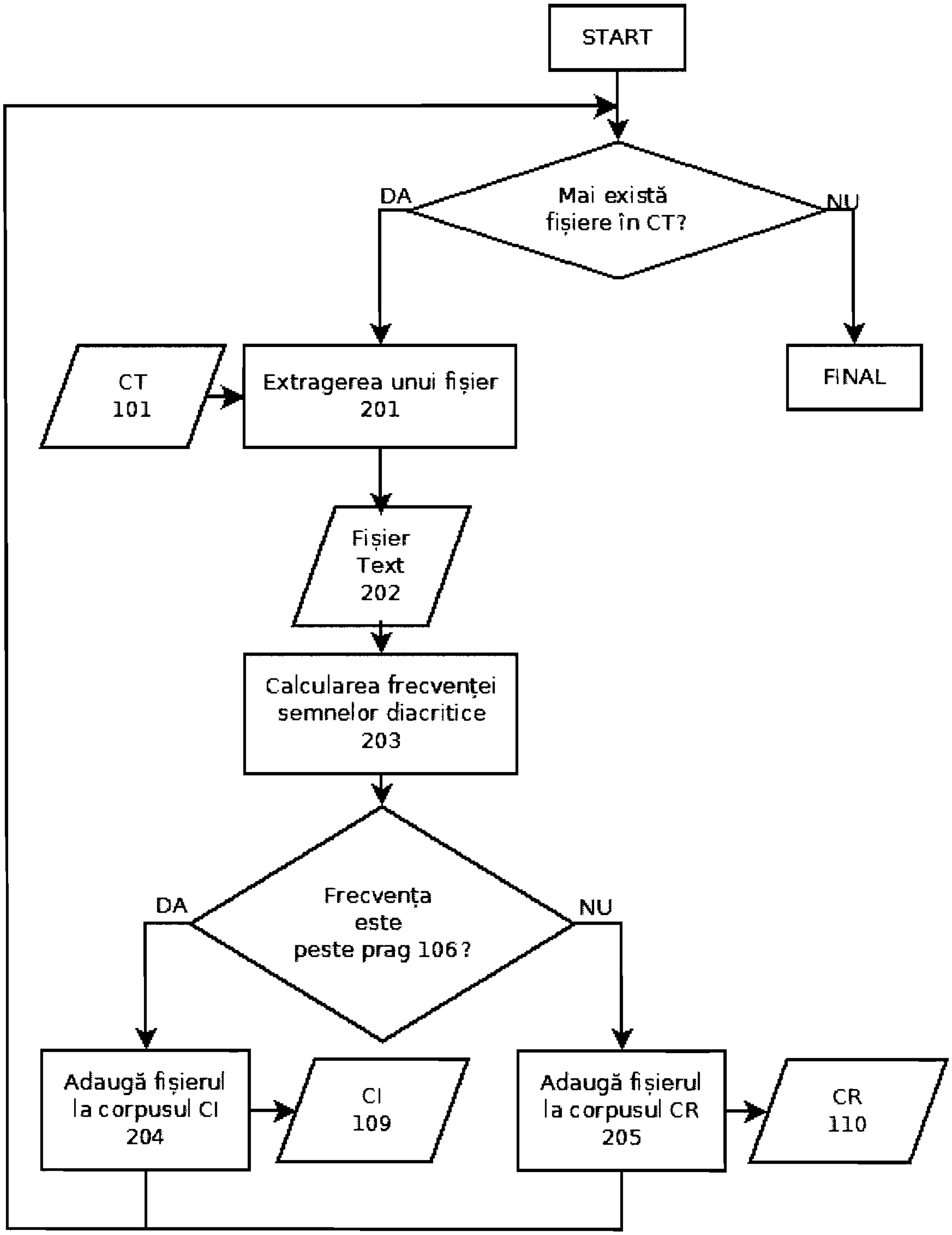

Figura 2: Filtrare folosind prag fix de separare

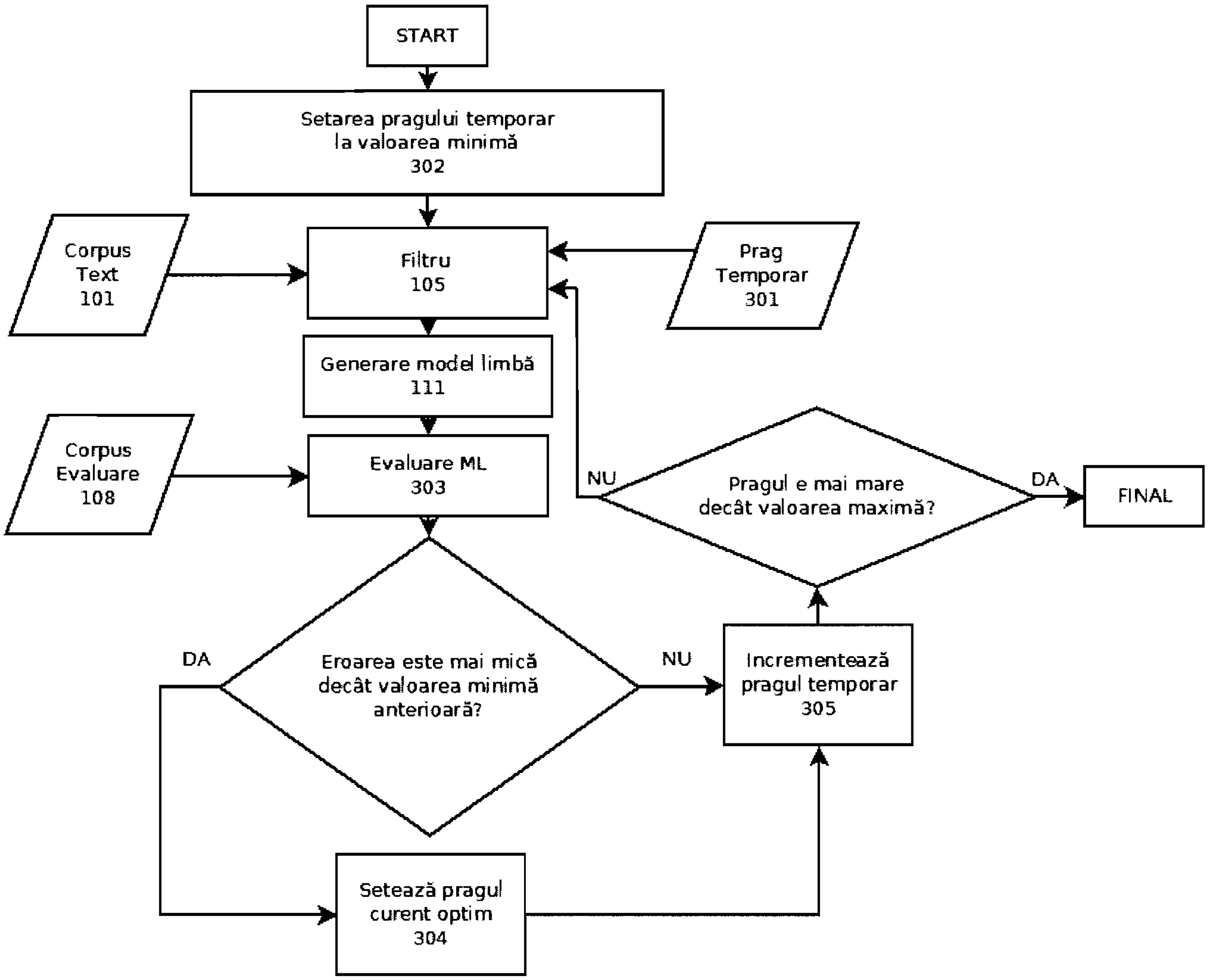

Figura 3: Identificarea pragului optim de separare

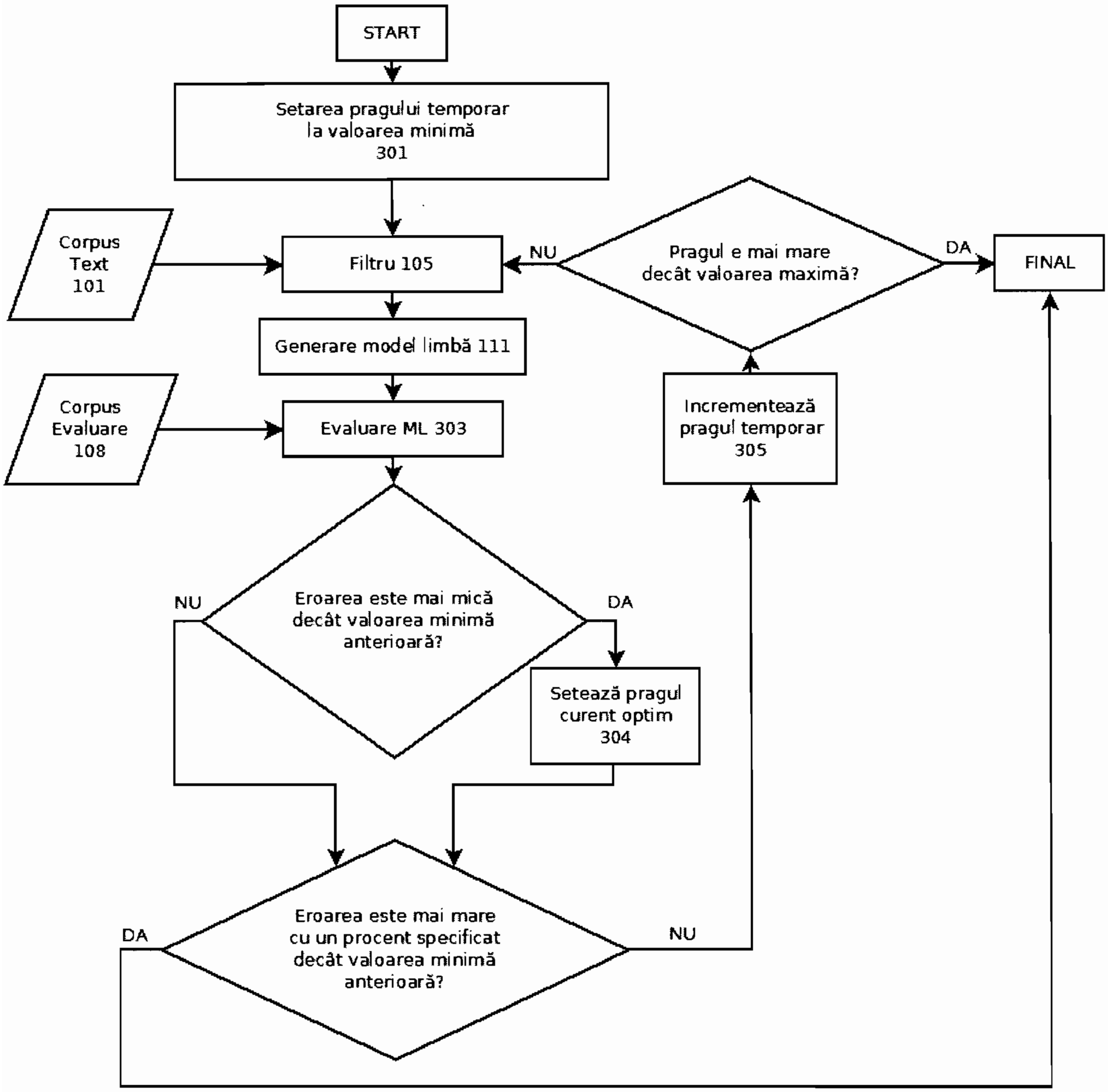

Figura 4: Identificarea speculativă a pragului optim de separare