Revista Română de Interacţiune Om-Calculator 7 (3) 2014, 259-272

Testarea versiunii beta a unei componente de recunoaştere dinamică a limbii – rezultate preliminare

Costin Pribeanu[1], Paul Fogarassy-Neszly[2]

[1] Institutul Național de Cercetare-Dezvoltare în Informatică – ICI Bucureşti
Bd. Mareşal Averescu, Nr. 8-10, 011455, Bucureşti
E-mail: pribeanu@ici.ro

[2] BAUM Engineering
Str. Traian Moşoiu nr. 8, 310175 Arad
E-mail: pf@baum.ro

Rezumat. Recunoaşterea automată a limbii în care este scris un text face parte din categoria generală a sarcinilor (recunoaşterea este o problemă, o sarcină, o aplicație, care se rezolvă pe baza unui algoritm), de clasificare a textelor şi are numeroase aplicații. Unele implementări recente au în vedere sinteza vocală diferențiată lingvistic. În acest caz, este necesară o recunoaşterea dinamică a limbii, care presupune implicarea unor algoritmi diferiți, capabili să identifice limba pe baza unui flux continuu de date. Această lucrare prezintă rezultatele preliminare obținute în testarea a două versiuni succesive (alpha şi beta) a unei componente de recunoaştere dinamică a limbii, cu accent pe a doua versiune. În testare au fost manipulați doi parametri: volumul de text analizat şi inerția la schimbarea limbii. Evaluarea a fost făcută din perspectiva utilizabilității aplicațiilor software, cu accent pe eficacitatea şi eficiența recunoaşterii limbii. Rezultatele testării versiunii beta sunt utile pentru determinarea valorii optime a parametrilor şi sugerează noi direcțiid e optimizare a metodei şi a algoritmilor.

Cuvinte cheie: recunoaştere automată a limbii, recunoaştere dinamică a limbii, testare beta, sinteză vocală, utilizabilitate, tehnologii asistive.

1. Introducere

Recunoaşterea automată a limbii, un domeniu de interes atât pentru lingvistica computațională cât şi pentru accesibilitatea sistemelor informatice, are numeroase aplicații, între care menționăm sinteza vocală diferențiată lingvistic şi traducerea automată (Kłosowski & Dustor, 2013; Devadoss, 2010).

În acest articol se prezintă rezultatele preliminare obținute la testarea unei componente care are ca scop recunoaşterea dinamică a limbii în care

este scris un text, în vederea sintezei vocale. Necesitatea recunoaşterii dinamice este impusă de modul de generare a fluxului audio, aproape simultan cu primirea fluxului de text (Dunning, 1994; Zissman, 1996). Deşi există aplicații software capabile să recunoască limba, precum şi aplicații capabile să redea prin voce sintetică un anumit text (TTS – Text-To-Speech), o aplicație care să îndeplinească ambele sarcini nu există.

Sinteza vocală este folosită mai ales de către persoane cu deficiențe de vedere, sau dificultăți de citire (persoane cu deficiențe cognitive, dislexici sau analfabeți), în scopul accesibilizării documentelor. Atunci când există posibilitatea ca texte scrise în limbi diferite să alterneze, este necesară discriminarea limbilor în care este scris fragmentul de text, deoarece sinteza vocală se bazează pe particularitățile fonetice ale limbii în care este scris textul, precum şi pe regulile de scriere specifice limbii respective. Cu alte cuvinte, este necesară selectarea sintezei vocale corespunzătoare limbii, pentru ca rezultatul să fie comprehensibil.

Între sistemele software care au nevoie de această componentă menționăm aplicațiile informatice pentru nevăzători (de exemplu, cititoarele de ecran), dispozitivele de citire automată pentru deficienți de vedere, aplicațiile care utilizează sisteme GPS şi aplicațiile cu redare vocală (En: self-voicing applications).

2. Metode şi algoritmi de recunoaştere a limbii

2.1 Metode de recunoaştere a limbii

Există numeroase metode de recunoaştere a limbii (Athias, 2007; Baldwin & Lui, 2010) precum si abordari care utilizează metode combinate (Gînscă et al., 2011). Întrucât într-o lucrare anterioară (Fogarassy-Neszly & Gherhes, 2013) a fost făcută o prezentare mai detaliată, în cele ce urmează le vom aminti succint împreună cu avantajele şi dezavantajele fiecăreia.

Identificarea limbii în care este scris un text poate fi făcută uneori cu uşurință prin simpla identificare a codificării caracterelor (character encoding detection) pentru textul analizat (Baldwin & Lui, 2010). Metoda nu poate fi utilizată (de exemplu) pentru discriminarea limbilor engleză şi franceză sau a limbilor rusă şi bulgară, deoarece aceste perechi de limbi folosesc caractere din acelaşi interval Unicode.

O altă metodă constă în căutarea cuvintelor (inclusiv a formelor flexionare) în dicționare specifice limbilor respective. Deşi are avantajul de a fi precisă chiar şi pentru texte scurte, metoda nu este practică datorită volumului foarte mare al dicționarelor şi faptului că implică algoritmi lenți.

O altă metodă intuitivă şi simplă, se bazează pe identificarea unor caractere speciale tipice unei limbi. Acestea sunt de obicei caracterele cu diacritice şi ligaturile. Folosirea caracterelor speciale pentru discriminarea limbilor are aplicabilitate limitată, deoarece în cazul unor texte scurte acestea pot să nu apară. De asemenea, unele dintre caracterele speciale sunt comune mai multor limbi.

O altă abordare a plecat de la identificarea unor secvențe de caractere specifice exclusiv unei singure limbi (Churcher, 1994; Churcher et al., 1994, Dunning, 1994). Deşi tentantă prin simplitate, asemenea secvențe de caractere izolate nu oferă suficientă încredere în identificarea limbii pentru care se presupune că sunt specifice, datorită unui număr mare de excepții posibile, precum şi din cauza cuvintelor preluate în alte limbi şi scrise conform regulilor din limba de origine.

Metoda secvenței de caractere specifice a condus la cea mai utilizată metodă şi anume metoda *n-gramelor* (Dunning, 1994; Cavnar & Trenckle, 1994). O n-gramă este o sub-secvență de n elemente dintr-o secvență dată; în general, secvența de elemente poate fi orice, de la caractere şi până la cuvinte. În analiza lingvistică n-gramele sunt utilizate mai mult pentru cuvinte sau pentru caractere. Atunci când este vorba de două caractere (n=2) se mai foloseşte termenul de bigramă (sau digramă), iar când este vorba de succesiuni de trei caractere (n=3) termenul consacrat este trigramă.

2.2 Aspecte specifice privind implementarea algoritmilor

Discriminarea limbilor pe baza n-gramelor (în general) pleacă de la observația că pentru fiecare limbă anumite n grame apar mai frecvent decât altele. Identificarea limbii se face cel mai simplu prin compararea frecvenței de apariție a trigramelor în textul analizat cu frecvența acestora în corpusurile limbilor care sunt avute în vedere.

În faza de „antrenare" a aplicației se construieşte spectrul de frecvențe al n-gramelor pentru fiecare limbă în parte. Acesta se bazează pe un corpus

relevant pentru limba avută în vedere şi domeniul de aplicare (dacă este cazul). Rezultatele unor studii (Dunning, 1994; Ljubesi et al., 2007) arată că un corpus de circa 50.000 de cuvinte oferă o precizie foarte bună care nu mai creşte semnificativ prin mărirea volumului.

Corpusul trebuie să fie omogen din punct de vedere al limbii caracterizate de acesta şi trebuie să fie corect gramatical şi sintactic; calitatea corpusului are o influenţă hotărâtoare asupra preciziei de identificare a limbii.

Spre deosebire de implementarea algoritmilor generici de identificare a limbii, algoritmii folosiţi la sinteza vocală diferenţiată sunt utilizaţi în condiţiile limitării la minimum a limbilor posibile; în mod normal, aplicaţiile care funcţionează în medii multiculturale trebuie să distingă între două, mai rar trei şi foarte rar patru limbi. Prima condiţie specifică de implementare constă în configurarea aplicaţiei doar pentru limbile care trebuie discriminate, conform specificaţiilor beneficiarului. Aceasta permite o viteză de răspuns mult mai mare decât a algoritmilor cu caracter general.

A doua condiţie specifică constă în selecţia metodei de discriminare în funcţie de intervalul Unicode caracteristic limbilor specificate. Dacă cele două limbi folosesc intervale Unicode diferite atunci acesta poate fi singurul criteriu de discriminare. Numai în cazul în care cel puţin două dintre limbile specificate folosesc acelaşi interval de caractere, se va folosi un algoritm bazat pe spectrul n-gramelor specific acestor limbi.

A treia condiţie specifică de implementare constă în definirea unei limbi implicite. Acest lucru este util pentru situaţiile în care rezultatul analizei este incert, mai ales datorită fragmentelor de text foarte scurte, sau datorită amestecului de cuvinte din limbi diferite în interiorul aceluiaşi fragment de text; tot rezultat incert poate să apară şi în cazul în care textul este scris într-o altă limbă decât cele pentru care algoritmul a fost configurat. În conjuncţie cu inerţia de schimbare a limbii, acest lucru permite obţinerea unui rezultat bun, chiar şi în cele mai dificile condiţii.

Ultima particularitate la implementarea algoritmilor este posibilitatea de realizare a unor liste de n-grame specifice pentru fiecare limbă analizată. În aceste liste vor fi păstrate doar n-gramele caracteristice care apar cu o pondere semnificativă în limba respectivă, de exemplu prin eliminarea n-gramelor care apar cu o frecvenţă relativă mai mică decât un anumit prag faţă de media frecvenţei tuturor n-gramelor identificate în corpusul limbii respective; valoarea acestui procent rezultă în urma testelor.

2.3 Rezultatele testării versiunii alpha

Într-o lucrare recentă (Fogarassy-Neszly & Gherheş, 2014) au fost prezentaţi succint algoritmii şi interfaţa de testare a versiunii alpha (proof-of-concept). Versiunea are un rol intermediar, scopul fiind recunoaşterea limbii în care este scris un text. Textul este scris într-o singură limbă iar recunoaşterea se face la nivelul întregului text. Testarea a fost făcută pe corpusuri de mici dimensiuni.

Procedura de testare a prevăzut mai multe sesiuni, corespunzător unor corpusuri de dimeniuni diferite, pentru fiecare din cele şase limbi utilizate: română, franceză, engleză, germană, spaniolă şi italiană. Mărirea succesivă a corpusului nu a îmbunătăţit sensibil rezultatele decât pentru primele două iteraţii (până la 6.500 de caractere).

În vederea recunoaşterii au fost pregătite trei seturi a câte cinci paragrafe de dimensiuni diferite. Primele două seturi sunt în limba română (cu diacritice şi fără diacritice) iar al treilea în limba engleză. Primele şase propoziţii sunt scurte (trei cuvinte), următoarele două de dimensiuni medii (6-11 cuvinte) şi ultima propoziţie scurtă (patru cuvinte).

Ca indicator al eficacităţii a fost utilizată rata de succes, calculată ca număr de fragmente de text pentru care limba este corect identificată raportat la numărul total de fragmente de text testate. Rezultatele au arătat că la un text de peste 6.500 de caractere, rata de succes a fost între 67.67% şi 73.33%. Rata de succes a fost mai mare pentru textul în limba română scris cu diacritice. Majoritatea eşecurilor au fost identificate în cazul propoziţiilor scurte (de trei cuvinte). În toate cazurile, limba a fost recunoscută pentru propoziţiile de 10-11 cuvinte.

3 Testarea versiunii beta

3.1 Componenta de testare

Prima versiune funcţională (beta 01) a componentei pentru recunoaşterea dinamică a limbii permite testarea algoritmilor implementaţi, precum şi analiza influenţei diverşilor parametrii şi factori asupra rezultatului obţinut pentru diferite tipuri de texte. Componenta de recunoaştere a limbii necesită

o fază de antrenare înainte de utilizare. Metoda recomandată pentru introducerea unui corpus lingvistic semnificativ este prin selectarea unui fişier de tip document. Astfel, se pot introduce corpusuri lingvistice largi care pot caracteriza suficient de exact o limbă.

Figura 1 prezintă interfața componentei de antrenare. Corpusul lingvistic poate fi introdus direct în fereastra „Analyze text", sau poate fi introdus sub forma unui fişier document de tip text (txt) sau Word (doc sau rtf). Pentru a se evita erorile triviale, datorită unor corpusuri lingvistice mult prea mici pentru a fi caracteristice, aplicația cere introducere unui text de minim 1.000 de caractere. Datorită limitării interfeței, se pot introduce direct (prin tastare sau prin copiere) maxim 100.000 de caractere.

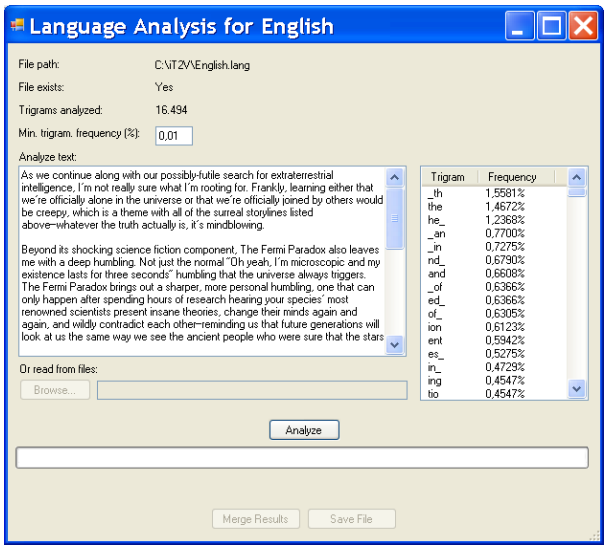

Figura 1. Interfața componentei de antrenare

Pe lângă calitatea corpusului lingvistic (volum şi reprezentativitate lingvistică) şi nivelul de ignorare a trigramelor nesemnificative, această componentă permite testarea influenței volumului de text analizat la un moment dat în timpul procesului de recunoaştere LA (*Look-ahead*), precum şi inerția la schimbarea limbii I (*Inertia*). Un alt factor (mai puțin evident, dar extrem de important) care influențează calitatea rezultatului este numărul de limbi activate.

3.2 Strategie de testare şi măsuri colectate

Fragmentul de text pentru care se face testarea cuprinde mai multe propoziţii, în două limbi: română şi engleză. Recunoaşterea se face la nivel de cuvânt. Testarea a fost făcută din perspectiva utilizabilităţii aplicaţiilor software. Din punctul de vedere al utilizabilităţii, în această etapă de dezvoltare interesează eficacitatea şi eficienţa componentei. Eficacitatea este definită de acurateţea şi completitudinea cu care este recunoscută limba. Eficienţa este definită de resursele consumate în acest scop.

Ca măsuri ale eficacităţii au fost colectate, pentru fiecare limbă, numărul de cuvinte (NC) şi numărul de propoziţii (NP) pentru care limba este corect identificată şi numărul de cazuri când schimbarea limbii este corect făcută (NS). Ca indicator al eficacităţii a fost utilizată rata de succes, calculată la nivel de propoziţie (EFP) şi la nivel de cuvânt (EFC). Rata de succes la nivel de propoziţie este importantă întrucât propoziţia este o unitate semantică având o semnificaţie pentru utilizator.

Ca măsuri ale eficienţei au fost utilizaţi cei doi parametri care pot fi modificaţi în testare, respectiv numărul de cuvinte analizate (LA) şi inerţia (I), precum şi numărul de limbi candidat.

Testarea a cuprins 3 etape, având următoarele obiective:

- Analiza influenţei mărimii corpusului de antrenare, analiza influenţei numelor proprii şi analiza variaţiei rezultatelor în funcţie de valoarea parametrilor, pentru un număr de şase limbi candidat;
- Testarea unui text din care s-au eliminat numele proprii, analiza influenţei pe care o are ordinea propoziţiilor şi influenţa pe care o are variaţia parametrilor;
- Identificarea valorilor optime a parametrilor şi confirmarea rezultatelor prin testarea cu un text preluat din lumea reală (articole din presă), având fragmente în patru limbi.

În primele două etape au fost folosite texte cu fragmente în două limbi: română şi engleză. În acest scop, pentru prima etapă au fost pregătite două seturi, a câte opt propoziţii fiecare. Primul set de propoziţii este în limba română, iar al doilea în limba engleză. Dintre cele 16 propoziţii, şase sunt scurte (patru din trei cuvinte şi două din patru cuvinte), şase sunt de 5-7 cuvinte şi numai patru propoziţii sunt mai lungi. Primele patru propoziţii

sunt în limba română şi următoarele opt în engleză, astfel încât sunt două schimbări de limbă. Pentru limba română, numărul total de cuvinte este 44. Dintre acestea, şase sunt nume proprii, iar unul este neologism. Pentru limba engleză, numărul total de cuvinte este 45.

Pentru etapa a doua, textul de recunoscut a fost adaptat corespunzător, păstrând propoziţiile şi eliminând majoritatea numelor proprii. Ordinea a două propoziţii a fost inversată.

4. Rezultate

4.1 Rezultate etapa I

Prima etapă de testare a avut un rol preliminar. Obiectivele au fost: analiza influenţei mărimii corpusului de antrenare, analiza influenţei numelor proprii şi analiza variaţiei rezultatelor în funcţie de valoarea parametrilor, pentru un număr de şase limbi candidat.

Procedura a prevăzut zece sesiuni de testare. Primele patru, corespunzător corpusurilor de diferite dimensiuni (5, 10, 20, 30 mii de cuvinte), utilizează patru limbi (română, engleză, franceză şi spaniolă). La fiecare iteraţie, trigramele din textul de antrenare au fost integrate cu cele existente în baza de date. Următoarele două sesiuni utilizează ultimul corpus (de dimensiune maximă), dar testează numai cu trei limbi candidat (română, engleză şi spaniolă), respectiv două limbi (română şi engleză). Sesiunea 7 utilizează limba italiană în locul limbii franceze, iar sesiunea 8 utilizează limba maghiară în locul limbii franceze. Următoarele două sesiuni utilizează numai trei limbi (română, engleză şi spaniolă, respectiv română, engleză şi maghiară). Toate cele patru sesiuni utilizează corpusul de dimensiune maximă disponibil (30 de mii de cuvinte).

În fiecare sesiune, a fost testată influenţa celor doi parametri, făcând câte patru teste, cu parametrii la valorile extreme. Rezultatele au fost mai bune pentru valorile mari ale parametrilor şi sugerează o testare cu variaţii mai fine. De asemenea, rezultatele au fost mai bune la testarea cu mai puţine limbi candidat.

În toate sesiunile, diferenţa între recunoaşterea limbii române şi cea a limbii engleze a fost mare. În Tabelul 1 sunt prezentate rezultatele obţinute la trei sesiuni de testare: valoarea parametrilor (LA/I), eficacitatea la nivel de propoziţie şi cuvânt (EFP, EFC) pentru fiecare limbă în care este scris

textul (ro şi en) şi eficacitatea globală, ca medie a eficacității pentru cele două limbi din textul testat (EFP, EFC)

Se observă diferența dintre rezultate în funcție de numărul de limbi candidat. De asemenea, se observă influența mai mare a inerției (I) față de volumul de text analizat (LA).

Tabelul 1 Rezultate testare (corpus de 30Kcuvinte)

LA/I	EFPro %	EFPen %	EFCro %	EFCen %	EFP %	EFC %
Sesiune 4: patru limbi candidat						
1/0	62.50	0.00	75.00	42.22	31.25	37.50
15/0	62.50	0.00	75.00	46.67	31.25	37.50
1/5	87.50	12.50	93.18	64.44	50.00	52.84
15/5	100.00	12.50	100.00	28.89	56.25	56.25
Sesiune 5: trei limbi candidat						
1/0	37.50	12.50	81.82	33.33	25.00	47.16
15/0	62.50	12.50	81.82	51.11	37.50	47.16
1/5	87.50	50.00	93.18	77.78	68.75	71.59
15/5	100.00	50.00	100.00	77.78	75.00	75.00
Sesiune 6: două limbi candidat						
1/0	75.00	25.00	93.18	57.78	50.00	59.09
15/0	62.50	25.00	86.36	55.56	43.75	55.68
1/5	100.00	50.00	100.00	80.00	75.00	75.00
15/5	100.00	62.50	100.00	77.78	81.25	81.25

4.2 Rezultate etapa II

Obiectivele au fost analiza influenței numelor proprii, ordinea propozițiilor şi variația parametrilor. Având în vedere că numele proprii ridică unele probleme, majoritatea au fost eliminate pentru a analiza schimbarea limbii. În ceea ce priveşte ordinea propozițiilor, s-a eliminat influența pe care o au unele cuvinte care sunt comune în mai multe limbi. În ceea ce priveşte parametrii, interesează valoarea cea mai scăzută pentru care se păstrează precizia.

În vederea testării au fost pregătite fişiere cu text de analizat pentru patru limbi: română, engleză, franceză şi germană. Corpusurile utilizate au fost de dimensiuni mai mari, între 45 şi 120 de mii de cuvinte. Textul de recunoscut a fost adaptat corespunzător, păstrând propozițiile şi eliminând majoritate numelor proprii. Ordinea a două propoziții a fost inversată.

Procedura de testare a prevăzut trei sesiuni a câte 24 iterații fiecare.

Prima sesiune utilizează patru limbi (ro, en, fr şi de). Următoarele două sesiuni testează numai cu trei limbi candidat (ro, en şi fr), respectiv două limbi (ro şi en). În fiecare sesiune, a fost testată influenţa celor doi parametri. Au fost făcute opt teste, mărind de fiecare dată valoarea parametrilor cu o unitate. În acest fel poate fi identificat intervalul cu valoarea cea mai scăzută a parametrului pentru care s-a păstrat precizia.

Rezultatele sunt mult mai bune decât în etapa precedentă. În Tabelul 2 sunt integrate rezultatele din primele două sesiuni, după cazurile în care parametrii au valori extreme.

Tabelul 2 Rezultate testare etapa II

LA/I	EFPr %	EFPe %	EFCr %	EFCe %	EFP %	EFC %
Sesiune 1: patru limbi candidat						
5/1	37.50	37.50	76.09	81.25	37.50	56.79
10/1	37.50	50.00	78.26	83.33	43.75	64.13
5/2	75.00	62.50	91.30	89.58	**68.75**	**76.90**
10/2	75.00	75.00	93.48	93.75	**75.00**	**84.24**
5/3	75.00	62.50	89.13	85.42	68.75	75.82
10/3	75.00	62.50	91.30	91.67	68.75	76.90
5/4	75.00	100.00	89.13	100.00	87.50	94.57
10/4	75.00	100.00	89.13	100.00	87.50	94.57
Sesiune 2: trei limbi candidat						
5/1	75.00	37.50	89.13	81.25	56.25	63.32
10/1	75.00	50.00	89.13	83.33	62.50	69.57
5/2	75.00	75.00	89.13	91.67	**75.00**	**82.07**
10/2	75.00	87.50	89.13	95.83	**81.25**	**88.32**
5/3	75.00	75.00	89.13	87.50	75.00	82.07
10/3	75.00	75.00	89.13	93.75	75.00	82.07
5/4	75.00	100.00	89.13	100.00	87.50	94.57
10/4	75.00	100.00	89.13	100.00	87.50	94.57

Rezultatele sesiunii 3 nu au fost păstrate fiind mai puţin concludente, deşi s-a observat acelaşi patern în rezultate. Pentru ambele sesiuni, cele mai bune rezultate se obţin pentru valori ale LA între 5 şi 10 şi valori ale inerţiei în jurul valorii de 2. La valori ale inerţiei de 3 se observă, în mod surprinzător, o scădere a eficacităţii. Este posibil să fie ca urmare a particularităţii cuvintelor din textul testat.

4.3 Rezultate etapa III

Obiectivele au fost identificarea valorilor cele mai scăzute ale parametrilor pentru care se păstrează precizia şi confirmarea rezultatelor pe un alt text.

Pentru testare au fost utilizate aceleaşi corpusuri.

Pentru identificarea valorii optime a parametrilor, s-a testat cu patru limbi candidat (ro, en, fr, de) în două sesiuni. În prima sesiune au fost efectuate şapte iteraţii, variind parametrul LA cu pas 1, între 5 şi 11 (parametrul I constant la valoarea 2). În cea de a doua sesiune au fost efectuate zece iteraţii variind parametrul I cu pas 0.1, între 1.5 şi 2.4 (parametrul LA constant la valoarea determinată în prima iteraţie).

Rezultatele testării sunt prezentate în Tabelul 3.

Tabelul. Rezultate testare etapa III

LA/I	EFPr %	EFPe %	EFCr %	EFCe %	EFP %	EFC %
Lb=4						
10/1.9	75.00	75.00	89.13	93.75	75.00	82.07
10/2.0	75.00	75.00	93.48	93.75	75.00	84.24
11/1.9	87.50	75.00	95.65	93.75	**81.25**	**85.33**
11/2.0	87.50	75.00	95.65	93.75	81.25	85.33
Lb=3						
10/1.9	75.00	87.50	89.13	95.83	**81.25**	**88.32**
10/2.0	75.00	87.50	89.13	95.83	81.25	88.32
11/1.9	75.00	87.50	89.13	93.75	81.25	88.32
11/2.0	75.00	87.50	89.13	93.75	81.25	88.32
Lb=3						
10/1.9	75.00	75.00	93.48	93.75	75.00	84.24
10/2.0	75.00	87.50	93.48	97.92	**81.25**	**90.49**
11/1.9	87.50	75.00	95.65	93.75	81.25	85.33
11/2.0	87.50	75.00	95.65	97.92	81.25	85.33
Lb=2						
10/1.9	75.00	87.50	89.13	95.83	81.25	88.32
10/2.0	75.00	100.00	89.13	100.00	**87.50**	**94.57**
11/1.9	75.00	87.50	89.13	95.83	81.25	88.32
11/2.0	75.00	100.00	89.13	100.00	87.50	94.57

Cele mai bune rezultate au fost obţinute pentru o valoare a parametrului LA de 11. Întrucât şi pentru valoarea lui LA de 10 rezultatele sunt apropiate, următoarea sesiune a utilizat această valoare, testând apoi şi pentru valoarea 11. În cea de a doua sesiune a fost variată inerţia. Cele mai bune rezultate s-au obţinut pentru valorile 1.9 şi 2.0. La următoarele iteraţii s-a observat o uşoară scădere a eficacităţii la recunoaşterea limbii române.

Se observă că la utilizarea a trei limbi candidat (primul caz este cu română, engleză, franceză şi al doilea cu română, engleză, germană) rezultatele cele mai bune se obţin la valori mai mici ale parametrului LA (10). De asemenea, la utilizarea a numai două limbi candidat rezultatele cele mai bune se obţin pentru LA=10 şi I=2.0

Se cuvine a fi menţionat faptul că eficacitatea la nivel de propoziţie este un indicator care are în mod inerent valori scăzute, deoarece presupune ca toate cuvintele dintr-o propoziţie să fie corect identificate. Având în vedere existenţa multor cuvinte comune pentru mai multe limbi (de exemplu „cinema") şi existenţa numelor proprii, este greu să se obţină valori ridicate pe texte din lumea reală.

Din aceleaşi motive, sunt greu de identificat în mod corect toate schimbările de limbă (la majoritatea iteraţiilor s-a identificat una singură).

Pentru confirmare pe un text din lumea reală, testarea s-a făcut cu un text format din patru paragrafe care conţin propoziţii preluate din presă (Adevărul, Times, Le Monde şi Deutche Welle). Fiecare paragraf se referă la cel puţin două domenii diferite (propoziţiile au fost preluate din rubrici diferite). Paragraful în limba română conţine opt propoziţii şi 173 cuvinte. Paragraful în limba engleză are zece propoziţii şi 203 cuvinte. Paragraful în limba franceză are nouă propoziţii şi 217 cuvinte. Paragraful în limba germană are zece propoziţii şi 161 cuvinte. Numărul de schimbări de limbă care ar trebui detectate este trei.

Tabelul 4 Rezultate test de confirmare

Limba	LA=10, I=2.0				LA=11, I=1.9			
	NP	NC	EFPr %	EFPe %	NP	NC	EFP %	EFC %
română	5	153	62.50	88.44	5	151	62.50	87.28
engleză	8	203	80.00	98.07	7	202	70.00	97.58
franceză	9	217	100.00	100.00	9	216	100.00	99.54
germană	2	126	20.00	78.26	3	137	30.00	85.09
			65.63	**91.19**			**65.63**	**92.38**

Testarea s-a făcut pentru valori ale parametrului LA de 10 şi 11 şi ale inerţiei de 1.9 şi 2.0. Rezultatele testării sunt prezentate mai jos. Aşa cum se observă din tabel, rezultatele au valori apropiate în cele două iteraţii, cu puţin mai mari pentru LA=11. În ceea ce priveşte recunoaşterea fiecărei limbi, pentru rezultatele au fost mai bune pentru limbile franceză şi engleză şi mai slabe pentru limbile română şi germană. Schimbarea limbii este corect detectată în două cazuri (en/de şi de/fr).

O comparaţie cu rezultatele anterioare arată că eficacitatea la nivel de cuvânt este mai mare (92.38% faţă de 85.33%). Eficacitatea la nivel de propoziţie este ceva mai mică (65.63% faţă de 81.25%). Pe ansamblu, rezultatele testării pe un text din lumea reală confirmă rezultatele testării pe un textul utilizat anterior.

Analiza detaliată arată că erorile provin în primul rând din identificarea unor cuvinte din română, engleză şi germană ca fiind franceze. Este posibil ca aceasta să se datoreze dimensiunilor diferite ale corpusurilor (cel în germană este mai mic) şi / sau sursei acestora.

În toate cazurile însă, numărul de cuvinte dintr-o propoziţie care sunt corect identificate este mai mare decât cel al cuvintelor care nu sunt corect identificate. Aceasta sugerează investigarea posibilităţii de a introduce în algoritmul de recunoaştere o analiză la nivel de propoziţie.

5. Concluzii şi direcţii de continuare

În cadrul acestui articol au fost prezentate rezultatele testării primei versiuni funcţionale a unei componente software de recunoaştere dinamică a limbii cu aplicaţie în sinteza vocală diferenţiată lingvistic.

O contribuţie a articolului este evaluarea performanţelor componentei din perspectiva utilizabilităţii. În acest scop, a fos propusă eficacitatea rdcunoaşterii limbii la nivel de propoziţie, eficacitatea la nivel de cuvânt şi eficacitatea globală. Ca parametri au fost utilizaţi: numărul de limbi candidat, numărul de cuvinte analizate şi inerţia la schimbarea limbii.

Rezultatele testării arată o influenţă scăzută a mărimii corpusurilor (posibil şi datorită utilizării unor corpusuri de dimensiuni relativ mici). Rezultatele sunt utile deoarece au permis identificarea unor valori optime a celor doi parametri care determină eficienţa recunoaşterii limbii (volum de text analizat şi inerţie la schimbarea limbii).

Cercetările vor continua prin investigarea posibilităţii de a introduce în algoritmul de recunoaştere o analiză la nivel de propoziţie. De asemenea, se intenţionează antrenarea componentei cu corpusuri de dimensiuni mai mari.

Confirmare

Această lucrare a fost elaborată în cadrul contractului 29DPST/13.09.2013, „*Aplicație pentru Conversia din Text în Voce Sintetică cu Recunoaşterea Automată a Limbii*”, în Programului Inovare, Dezvoltare Sisteme-Produse-Tehnologii a UEFISCDI.

Referințe

Athias, L. (2007). *Statistical Machine Translation and Automatic Speech Recognition under Uncertainty*. PhD thesis, Johns Hopkins University.

Baldwin, T., Lui, M. (2010) Language Identification: The Long and the Short of the Matter, Human Language Technologies: *2010 Annual Conference of the North American Chapter of the ACL*, 229–237.

Cavnar, W., and Trenkle, J. (1994). N-gram-based text categorization. *Proc. 3rd Symp. on Document Analysis and Information Retrieval* (SDAIR-94).

Churcher, G. (1994) *Distinctive character sequences*. Personal communication.

Devadoss, J.M. (2010) Advanced Natural Language Translation System. *Global Journal of Computer Science and Technology 9(5)*, 114-122.

Dunning, T. (1994) *Statistical Identification of Language*. Technical Report MCCS 94-273, New Mexico State University.

Fogarassy-Neszly, P., Gherhes, V. (2013) Recunoaşterea automată a limbii cu aplicație în sinteza vocală diferențiată lingvistic. *Revista Română de Interacțiune Om-Calculator 6(2)*, 155-168.

Fogarassy-Neszly, P., Gherhes, V. (2014) Aplicații pentru recunoaşterea dinamică a limbii. Popovici, D.M., Iordache D.D. (Eds.). *Proc. Conferința Națională de Interacțiune Om-Calculator – RoCHI 2014*, 51-54.

Gînscă, A. L., Boros, E., Iftene, A. (2011). Adapting Statistical Language Identification Methods for Short Queries. *Notebook Paper CLEF 2011 LABs Workshop*, 19-22 September, Amsterdam, Netherlands. Kłosowski, P., Dustor, A. (2013) Automatic Speech Segmentation for Automatic Speech Translation. *Computer Networks Communications in Computer and Information Science, 370*, 466-475.

Kłosowski, P., & Dustor, A. (2013). Automatic Speech Segmentation for Automatic Speech Translation. In *Computer Networks*, 466-475.

Ljubeši´c, N., Mikeli´c, N., Boras, D. (2007) Language identification: How to distinguish similar languages. In Lužar-Stifter, V. and Hljuz Dobri´c, V. (Eds), *Proceedings of the Intl Conference on Information Technology Interfaces*, 541–546.

Zissman, M. A. (1996) Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing 4(1)*, 31-44.