Revista Română de Interacţiune Om-Calculator 5 (3) 2012, 15-34

POS tagger bazat pe modelul HMM de ordinul doi

Dumitru-Clementin Cercel[1], Ştefan Trăuşan-Matu[2,3]

[1]Universitatea Politehnica Bucureşti
Bd. Splaiul Independenţei, Nr. 313, 060042, Bucureşti
E-mail: clementin.cercel@gmail.com

[2] Universitatea Politehnica Bucureşti
Bd. Splaiul Independenţei, Nr. 313, 060042, Bucureşti
E-mail: stefan.trausan@cs.pub.ro

[3] Institutul de Cercetări pentru Inteligenţă Artificială
Calea 13 Septembrie, Nr. 13, 050711, Bucureşti

Rezumat. Adnotarea cu etichete morfo-sintactice („Part-of-speech tagging - POS tagging") este procesul de etichetare gramaticală a fiecărui cuvânt dintr-o propoziţie, frază sau paragraf cu partea de vorbire corespunzătoare. Acest proces este o componentă a altor aplicaţii din prelucrarea limbajului natural şi, prin urmare, rezultatele trebuie să fie cât mai precise. Odată ce o parte de vorbire a fost identificată, aceasta oferă informaţii suplimentare despre părţile de vorbire care pot apărea în aceeaşi propoziţie. În cazul etichetării cu părţi de vorbire a cuvintelor, apar ambiguităţi ca urmare a faptului că un cuvânt poate avea, în funcţie de context, valori morfologice multiple. Articolul tratează dintr-o perspectivă experimentală un POS tagger bazat pe modelul Markov ascuns ("Hidden Markov Model - HMM") de ordinul doi, folosind corpusul Brown. Testele au fost realizate pentru obţinerea rezultatelor în funcţie de diverşi parametri. Vom arăta cum evoluează precizia POS tagger-ului pentru limba engleză atunci când se modifică, pe de o parte, dimensiunea setului de antrenare, iar, pe de alta parte, domeniul setului de testare faţă de cel al setului de antrenare. Am identificat categoriile de texte din corpusul Brown folosite pentru corpusul de antrenare atunci când precizia POS tagger-ului este mai mare, respectiv mai scăzută.

Cuvinte cheie: NLP, POS Tagging, trigrame, HMM, algoritmul Viterbi, corpus Brown.

1. Introducere

Part-of-speech tagging este procesul de etichetare gramaticală a fiecărui cuvânt dintr-o propoziţie, frază sau paragraf cu partea de vorbire corespunzătoare. Etichetele unui POS tagger pot reprezenta părţi de vorbire în general (de exemplu, substantive, verbe, prepoziţii, interjecţii), dar pot să conţină informaţii suplimentare legate de caracteristicile morfologice ale

limbii respective, precum număr, gen, persoană, timpul sau aspectul verbului.

POS tagging-ul este adesea o componentă a altor aplicații din prelucrarea limbajului natural şi, ca urmare, rezultatele acestui proces trebuie să fie cât mai precise. Atât Manning & Schutze (1999), cât şi Jurafsky and Martin (2000) prezintă pe larg exemple de astfel de aplicații. O aplicare a acestui pas de preprocesare este întâlnită într-un model pentru recunoaşterea vorbirii. În regăsirea informației („information retrieval") în procesul de reducere a cuvintelor la radăcina lor („stemming"), unui cuvânt i se poate elimina afixul, cunoscându-se ce parte de vorbire reprezintă. De asemenea, POS tagging-ul se poate folosi în algoritmii de dezambiguizare a sensului cuvintelor.

În limba engleză, setul de etichete are o dimensiune în medie între 50 şi 150, spre deosebire de limbile puternic flexionare, pentru care numărul de etichete este mult mai mare. Pentru limba engleză, unul dintre cele mai utilizate corpusuri pentru antrenarea unui model probabilistic în adnotarea cu etichete morfo-sintactice a unui text este corpusul Brown (Francis & Kucera, 1979), care foloseşte 87 de etichete. Alte seturi de etichete care se folosesc deseori sunt: setul Penn Treebank (Marcus et al., 1993), ce conține 45 de etichete, setul C7 (Leech et al., 1994), alcătuit din 147 de etichete. Setul C5 (Garside et al., 1997) este format din 61 de etichete şi este folosit de corpusul BNC (British National Corpus).

În cazul etichetării cu părți de vorbire a cuvintelor, apar ambiguități ca urmare a faptului că un cuvânt poate avea, în funcție de context, valori morfologice multiple. Această problemă, de alegere a etichetei adecvate, se poate rezolva luând în considerare trăsăturile cuvântului. Aşa cum se arată în Jurafsky & Martin (2000), corpusul Brown este format din 44.019 cuvinte neambigue şi 5.490 cuvinte ambigue, care conțin de la 2 până la 7 părți de vorbire. În comparație cu acesta, corpusul WSJ (Wall Street Journal - Marcus et al, 1993) conține mai multă ambiguitate, deşi are mai puține etichete, el având 38.857 cuvinte neambigue şi 8.844 cuvinte ambigue, care utilizează între 2 şi 9 etichete.

În această lucrare vom descrie implementarea unui POS tagger ce foloseşte modelul Markov ascuns bazat pe trigrame, conform lui Jurafsky & Martin (2000) şi Brants (2000), abordarea noastră fiind una experimentală. Testele au fost realizate pentru obținerea diverselor tipuri de rezultate în funcție de diverşi parametri. Pentru testare vom folosi corpusul Brown, în

care textele componente sunt clasificate pe categorii. Această clasificare ne ajută să arătăm cum se modifică precizia când antrenăm şi testăm POS tagger-ul pe date din domenii sursă diferite.

În secțiunea următoare vom prezenta aspecte teoretice ale modelelor Markov ascunse. Secțiunea 3 se axează pe abordările curente pentru POS tagging şi factorii ce influențează performanța unui model de etichetare. În continuare vom descrie implementarea unui POS tagging ce foloseşte modelul Markov ascuns de ordinul 2. Secțiunea 4 cuprinde rezultatele testelor şi analiza acestora. Ultima secțiune prezintă concluziile şi identifică direcții viitoare de cercetare.

2. Modele Markov ascunse

Un proces Markov (Manning & Schutze, 1999) este descris de o mulțime de N stări $S = \{S_1, S_2, ..., S_n\}$, de o matrice de probabilități de tranziție $A = \{a_{ij}\}$, $1 \le i,j \le n$, $\sum_{j=1}^{n} a_{ij} = 1, \forall i$ unde $a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i)$ este probabilitatea de a trece din starea i în starea j şi de un vector de probabilități inițiale $\pi_i = P(q_1 = S_i),\ 1 \le i \le n,\ \sum_{i=1}^{n} \pi_i = 1$ ce semnifică probabilitatea ca procesul Markov să pornească din starea i, unde q_t este starea sistemului la momentul de timp t. Un proces Markov poate fi modelat ca un automat finit de stări, în care fiecare arc este etichetat cu o probabilitate de tranziție.

Modelele Markov au două proprietăți. Conform primei proprietăți, starea curentă depinde de un număr de stări din trecut. În cazul modelului Markov de ordin I, putem face predicția pentru variabila curentă folosind doar valoarea variabilei precedente, fără a mai ține cont de celelalte variabile anterioare din secvență (este condițional independentă de elementele din trecut). Pentru modelul Markov de ordin II, probabilitatea unei anumite stări este dependentă doar de ultimele două stări. Potrivit celei de a doua proprietăți, probabilitatea ca variabila curentă să fie precedată de variabila anterioară nu se schimbă în timp. Analizând cele două proprietăți în cazul POS tagging, rezultă că eticheta unui cuvânt depinde doar de eticheta anterioară şi această dependență nu se schimbă în timp.

Un model Markov ascuns este definit prin cvintuplul (N, K, A, B, π), unde K este mulțimea simbolurilor de ieşire din HMM, $B = b_{jk}$, $j \in S$, $k \in K$

reprezintă probabilitatea ca modelul să emită în starea j simbolul de ieşire k. Celelalte componente ale modelului au semnificaţia prezentată mai sus.

Într-un model Markov ascuns, secvența de stări pe care procesul o generează nu este cunoscută (este ascunsă), deoarece se cunoaşte doar secvența de observații (cuvintele în cazul nostru).

În cazul POS tagging, etichetele se reprezintă ca stări în automatul finit HMM. Astfel, N este numărul de etichete folosite de model, dimensiunea lui K reprezintă numărul de cuvinte distincte din vocabularul modelului, a_{ij} este probabilitatea ca tagul t_j să fie precedat de eticheta t_i, b_{jk} este probabilitatea ca sistemul fiind în stare t_j să emită cuvântul w_k, altfel spus, cuvântului w_k să îi corespundă partea de vorbire t_j, iar π_i este probabilitatea ca primul cuvânt al secvenței de cuvinte să fie etichetat cu t_i.

3. POS tagging

3.1 Starea actuală a domeniului

Într-un studiu realizat de Association for Computational Linguistics (http://aclweb.org/aclwiki/index.php?title=POS_Tagging_(State_of_the_art) pentru problema POS tagging este făcută o analiză a rezultatelor mai multor implementări de metode de învățare automată, folosind pentru antrenare şi testare corpusul WSJ.

Rezultatele obținute pentru cuvintele cunoscute din setul de date de antrenare, care sunt incluse în corpusul de antrenare, au o precizie ridicată. Folosind POS tagger-ul TnT bazat pe modelul Markov ascuns, precizia a fost de 96.46%. POS tagger-ul SVMTool, introdus de Giménez & Márquez (2004) şi bazat pe maşina cu vector suport („Support Vector Machine – SVM") a avut precizia de 97.16%. Precizia pentru Stanford Tagger 2.0 (http://nlp.stanford.edu/software/tagger.shtml), ce foloseşte modelul entropiei maximale, a fost de 97.32%. POS tagger-ul LTAG-spinal (Shen et al., 2007) şi care utilizează un algoritm având la bază perceptronul bidirecțional de învățare („bidirecțional perceptron learning") are precizia de 97.33%. POS tagger-ul Morče/COMPOST (Spoustová et al, 2009) are precizia de 97.44% folosind o metodă bazată pe perceptronul de medie („averaged perceptron"). Cel mai bun rezultat, cu precizia de 97.50, a fost obținut de POS tagger-ul SCCN propus de către Søgaard (2011) şi care se

bazează pe un model ce foloseşte cel mai apropiat vecin condensat („condensed nearest neighbor").

În cazul cuvintelor necunoscute, care nu se găsesc în corpusul de antrenare, cea mai mare precizie, de 91.29%, s-a obținut folosind POS tagger-ul MElt (Denis & Sagot, 2009), iar cel mai nesatisfăcător rezultat a fost obținut de POS tagger-ul TnT, de 85.86%.

Au fost realizate POS tagger-e de o precizie ridicată şi pentru alte limbi, ca de exemplu, pentru limba română (Simionescu, 2011; Tufiş et al, 2008). Precizia POS tagger-ului hibrid implementat de Simionescu (2011) a fost pentru cuvintele necunoscute de 93.31%, iar pentru întreg corpusul de testare, de 97.03%.

3.2 Modele de rezolvare pentru POS Tagging

Jurafsky & Martin (2000) fac o clasificare a metodelor de rezolvare pentru POS tagging în: metode stocastice (probabilistice), metode bazate pe reguli şi abordări hibride. În cazul metodelor stocastice, se foloseşte un corpus de antrenare şi se calculează pentru fiecare cuvânt probabilitatea să-i corespundă o anumită etichetă într-un context anume. POS tagger-ele probabilistice includ abordări care se bazează pe n-grame şi necesită o învățare supervizată dintr-o cantitate mare de date de antrenare pentru a se obține o precizie ridicată.

La metodele bazate pe reguli, într-o prima etapă, pe baza unui dicționar fiecărui cuvânt îi este asociată o posibilă listă de părți de vorbire, apoi un set de reguli referitoare la secvența de părți de vorbire permise este aplicat pentru a rezolva ambiguitățile. Un exemplu este POS tagger-ul numit English Constraint Grammar Parser (Voutilainen, 1995; Voutilainen et al, 1995), ce a avut o precizie de 99,7%, deşi unele rezultate obținute de acest tagger conțineau ambiguități.

Implementările hibride îmbină cele două soluții anterioare. O astfel de abordare este POS tagger-ul propus de Brill (1995), cu o precizie între 96-97% şi care foloseşte reguli deduse din date folosind un corpus de antrenare etichetat, pentru a stabili când un cuvânt ambiguu poate să aibă o etichetă dată.

Manning & Schutze (1999) identifică factorii care influențează procesul de etichetare a cuvintelor unui text, cu părțile de vorbire corespunzătoare:

- Dimensiunea setului de date de antrenare. Cu cât datele de antrenare sunt mai numeroase, cu atât precizia POS tagger-ului este mai mare.
- Setul de etichete folosit. Pe de o parte, un set mare de etichete permite obținerea, în contextul dat, a mai multor trăsături morfologice ale cuvântului, iar pe de altă parte, introduce mai multă ambiguitate.
- Corpusul de antrenare şi corpusul de testare. Precizia rezultatelor va fi scazută dacă aceste corpusuri nu au fost scrise de-a lungul aceleaşi perioade de timp şi nu sunt din acelaşi domeniu sursă (de exemplu, textele ştiinţifice şi textele din ziare).
- Cuvintele necunoscute („unknown words") reprezintă cuvintele pe care POS tagger-ul trebuie sa le eticheteze, dar care nu se întâlnesc în corpusul de date de antrenare. Numărul de cuvinte necunoscute poate fi şi mai mare atunci când etichetăm un text dintr-un domeniu tehnic sau texte care prezintă particularități, cum ar fi chat-urile.

Au fost dezvoltați diferiți algoritmi pentru a îmbunătăți precizia unui POS tagger în cazul cuvintelor necunoscute, folosind diferite trăsături ale cuvântului sau contextul său. Jurafsky & Martin (2000) prezintă o parte din aceşti algoritmi. Un algoritm propus de Baayen & Sproat (1996) consideră că distribuirea probabilă de etichete pentru cuvintele necunoscute este similară distribuirii de etichete pentru cuvintele care apar o singură dată într-un set de antrenare (cunoscute sub denumirea de "hapax legomena"). Weischedel et al. (1993) au descris un algoritm mai eficient pentru găsirea etichetei unui cuvânt necunoscut, ce se bazează pe regulile ortografice, luând în considerare patru trăsături: morfemele flexionare, morfemele derivative, scrierea cu majuscule şi despărțirea în silabe.

3.3 POS tagger bazat pe modelul HMM de ordinul doi

Fiind dată o secvență de cuvinte $W=w_1w_2,...,w_n$, procesul de POS tagging presupune determinarea celei mai probabile secvențe de stări pe care modelul o parcurge, adică a celei mai probabile secvențe de etichete $\hat{T}$ care maximizează P(T|W). Cum probabilitatea unei secvențe de cuvinte este

constantă, folosind regula lui Bayes (P(W)P(T|W) = P(T)P(W|T)) rezultă:

$$\hat{T} = \arg\max_{T} P(T)P(W \mid T) \quad (1)$$

Din prima proprietate Markov cunoaştem că probabilitatea unei etichete depinde de două etichete anterioare şi aplicând formula lui Bayes putem scrie că:

$$P(T) = P(t_{n+1} \mid t_1, t_2, ..., t_n)P(t_1, t_2, ..., t_n) = P(t_{n+1} \mid t_n)\prod_{i=1}^{n} P(t_i \mid t_{i-2}, t_{i-1}) \quad (2)$$

unde t_0, t_1 sunt etichete auxiliare, adăugate la începutul unei propoziţii, iar t_n este o etichetă auxiliară, adăugată la sfârşitul propoziţiei. Relaţia între un cuvânt şi eticheta sa fiind independentă de context, obţinem:

$$P(w_1, w_2, ..., w_n \mid t_1, t_2, ..., t_n) \sim= \prod_{i=1}^{n} P(w_i \mid t_i) \quad (3)$$

Probabilitatea ca un cuvânt sa fie etichetat cu o anumită parte de vorbire se calculează împărţind numărul de apariţii în text ale cuvântului w_i cu eticheta corespunzătoare t_i la frecvenţa etichetei ($P(w_i \mid t_i) = \frac{C(w_i, t_i)}{C(t_i)}$). Rezultă că modelul trigram pentru POS tagging, aşa cum se arată în Jurafsky & Martin (2000) este:

$$\hat{T} = \arg\max_{t_1 \ldots t_n}[\prod_{i=1}^{n} P(t_i \mid t_{i-1}, t_{i-2})P(w_i \mid t_i)]P(t_{n+1} \mid t_n) \quad (4)$$

Dată fiind o secvenţă de observaţii (în cazul nostru, de cuvinte), cea mai probabilă secvenţă de etichete poate fi găsită printr-o căutare în forţă brută, evaluând probabilitatea fiecărei secvenţe posibile de etichete pentru secvenţa de cuvinte de intrare, dar aceasta necesită un timp mare de execuţie. Algoritmul lui Viterbi (1967) este cel mai cunoscut algoritm pentru acest task şi se bazează pe programarea dinamică. Noi am folosit acest algoritm în implementare.

Din cauza datelor insuficiente din corpusul de antrenare, se poate întâmpla ca o secvenţă oarecare de trigrame să apară în corpusul de testare, dar nu şi în cel de antrenare şi, astfel, să stabilim incorect probabilitatea secvenţei $P(t_i \mid t_{i-2}, t_{i-1})$ ca fiind zero. Chiar şi atunci când secvenţa de trigrame apare de prea puţine ori în corpusul de antrenare, probabilitatea calculată pentru secvenţa respectivă nu ar fi o estimare exactă. De aceea,

probabilitatea unei secvenţe de trigrame se calculează ca fiind o interpolare liniară între estimarea probabilităţii maxime (”maximum likelihood”) pentru probabilităţile secvenţelor de trigrame, bigrame şi unigramă (Brants, 2000):

$$P(t_i \mid t_{i-2}, t_{i-1}) = \lambda_3 \hat{P}(t_i \mid t_{i-2}, t_{i-1}) + \lambda_2 \hat{P}(t_i \mid t_{i-1}) + \lambda_1 \hat{P}(t_i) \qquad (5)$$

$\lambda_1 + \lambda_2 + \lambda_3 = 1$, $0 <= \lambda_i <= 1$.

Pentru estimarea probabilităţii maxime, pentru fiecare din aceste probabilităţi se folosesc relaţiile următoare: $\hat{P}(t_i) = \frac{C(t_i)}{N}$,

$$\hat{P}(t_i \mid t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})} \quad \hat{P}(t_i \mid t_{i-2}, t_{i-1}) = \frac{C(t_{i-2}, t_{i-1}, t_i)}{C(t_{i-2}, t_{i-1})} \qquad (6)$$

unde N este numărul total de cuvinte, $C(t_i)$ este frecvenţa etichetei t_i în corpusul de antrenare, iar $C(t_{i-1},t_i)$ reprezintă numărul de apariţii ale secvenţei de etichete (t_{i-1},t_i) în corpusul de antrenare.

Coeficienţii lambda au fost estimaţi cu ajutorul datelor de antrenare, aplicând metoda „deleted interpolation” propusă de Brants (2000).

Pentru cuvintele care nu se găsesc în corpusul de antrenare, probabilitatea $P(w_i|t_i)$ se obţine folosind analiza sufixelor. Un sufix este definit ca fiind ultimele i litere dintr-un cuvânt. Numărul optim de litere i ale sufixelor va fi determinat experimental. Sufixele pot oferi un bun indiciu despre partea de vorbire asociată unui cuvânt. Folosind un model propus de Samuelsson (1993) şi Brants (2000), vom calcula probabilitatea unei anumite etichete, fiind cunoscute ultimele i litere dintr-un cuvânt de L litere:

$$P(t \mid l_{L-i+1}, ..., l_L) = \frac{\hat{P}(t \mid l_{L-i+1}, ..., l_L) + \Theta_i P(t \mid l_{L-i}, ..., l_L)}{1 + \Theta_i} \qquad (7)$$

Sufixele cuvintelor cu o frecvenţă mai mică sau egală faţă de o valoare de prag sunt folosite pentru construcţia unei structuri de date şi anume arborele de sufixe. Astfel, algoritmul construieşte un arbore de sufixe care conţine sufixele pentru cuvintele care încep cu literă mică, un alt arbore pentru cuvintele care încep cu literă mare şi de asemenea un arbore pentru cuvintele care încep cu cifre. Pentru estimarea probabilităţii maxime pentru un sufix al unui cuvânt, se foloseşte următoarea relaţie:

$$\hat{P}(t \mid l_{L-i+1}, ..., l_L) = \frac{C(t \mid l_{L-i+1}, ..., l_L)}{C(l_{L-i+1}, ..., l_L)} \qquad (8)$$

Parametrul Θ_i reprezintă abaterea standard a mulțimii de etichete întâlnite în corpusul de antrenare, adică este rădăcina medie pătrată a abaterilor estimărilor probabilităţii maxime pentru fiecare etichetă de la media probabilităţii maxime a mulţimii de etichete.

$$\Theta_i = \sqrt{\frac{1}{s}\sum_{j=1}^{s}(\hat{P}(t_j) - \hat{P})^2} \quad , \quad \hat{P} = \frac{1}{s}\sum_{j=1}^{s}\hat{P}(t_j) \tag{9}$$

4. Evaluare

4.1 Date folosite pentru experimente

Aşa cum se arată în Hinrichs et al. (2010), „familia de corpusuri Brown" cuprinde corpusul Brown, corpusul LOB (Lancaster-Oslo/Bergen) pentru engleza britanică, corpusul Frown (Freiburg-Brown, 1992) pentru engleza americană şi corpusul FLOB (Freiburg Lancaster-Oslo/Bergen, 1991) pentru engleza britanică. Aceste corpusuri sunt asemănătoare în ceea ce priveşte dimensiunea şi structura pe categorii, arătând, pentru o perioadă determinată, diferenţele dintre engleza britanică şi cea americană şi, totodată, schimbările gramaticale apărute în istoria limbii engleze.

Corpusul Brown descris în (Francis & Kucera, 1979) a fost primul corpus consistent pentru limba engleză şi este format din texte publicate în Statele Unite ale Americii în anul 1961. Corpusul Brown are şase versiuni, care conţin acelaşi text de bază, dar diferă prin format.

Din punct de vedere structural, corpusul Brown iniţial este împărţit în 15 categorii de texte:

A. Reportaje pe diverse teme: politică, sport, societate, cultură
B. Editoriale
C. Recenzii
D. Texte religioase
E. Publicaţii având ca subiect aptitudini şi hobby-uri
F. Folclor literar
G. Biografii, eseistică
H.Scrieri care combină diverse genuri (documente ale guvernului, rapoarte ale unor fundaţii, rapoarte industriale, cataloage de prezentare ale

unor universităţi, publicaţii sau reviste ale întreprinderilor pentru proprii clienţi sau angajaţi)

J. Lucrări ştiinţifice
K. Romane şi povestiri de ficţiune în sens generic
L. Romane şi povestiri poliţiste sau de groază
M. Literatură SF
N. Romane şi povestiri de aventură
P. Romane şi drame romantice
R. Texte umoristice

În secţiunea următoare, vom face referire la aceste categorii de texte prin litera de identificare care este trecută la începutul categoriei. Din punctul de vedere al dimensiunii, corpusul Brown conţine un număr de peste un milion de cuvinte. Din totalul de 50.000 de cuvinte distincte din corpusul Brown, aproximativ jumătate din cuvinte apar o singură dată în corpus. Este de remarcat faptul că articolul hotărât „the" constituie 7% din corpus, în timp ce prepoziţiile „of" şi „to" formează 6% din corpus.

Setul de etichete asociat părţilor de vorbire, folosit la etichetarea corpusului Brown (http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml), pe care l-am utilizat în testele din lucrare, conţine 472 de etichete. O descriere parţială a etichetelor din acest set poate fi găsită la adresa: http://www.scs.leeds.ac.uk/ccalas/tagsets/brown.html. Cea mai mare parte din aceste etichete este reprezentată de combinaţii de două sau mai multe etichete simple. Verbele la formă negativă au asociată o etichetă care se termină cu asterisc (de exemplu, cuvântul „wasn't" are eticheta „BEDZ*", unde „BEDZ" este eticheta pentru verbul „to be" la timpul trecut, persoana întâi şi a treia singular). Pentru forma contractată a cuvintelor, se foloseşte simbolul „+" ce separă etichetele pentru fiecare cuvânt component (de exemplu, cuvântul „nobody'd" are eticheta „PN + HVD", unde „PN" este eticheta pentru pronumele personal, iar „HVD" este eticheta pentru verbul „to have", la timpul trecut). Pentru cuvintele împrumutate din alte limbi, etichetele încep cu prefixul „FW-".

4.1 Rezultate

Algoritmul de etichetare cu părţi de vorbire, ce foloseşte modelul HMM bazat pe trigrame, a fost testat în diferite moduri utilizând corpusul Brown. Pentru început, a fost necesară determinarea parametrilor care maximizează

precizia POS tagger-ului pentru cuvintele necunoscute şi anume, valoarea de prag a frecvenței pentru cuvintele din setul de antrenare, respectiv numărul optim de litere din terminația cuvintelor din setul de antrenare. Sufixele cuvintelor din corpusul de antrenare, care au o frecvență mai mică sau egală cu frecvența de prag sunt folosite pentru construcția arborilor de sufixe.Valorile pentru frecvența de prag avute în vedere sunt: 1, 3, 4, 5, 6, 8, 11 şi 15, respectiv pentru sufixe au fost luate în considerare: 1, 2, 3, 4 şi 5 litere.

Pentru acest prim test s-a folosit validarea încrucişată („cross validation"). Corpusul Brown a fost împărțit în 6 diviziuni, fiecare conținând în proporții aproximativ egale propoziții din cele 15 categorii ale corpusului. De fiecare dată, una din cele 6 secțiuni este folosită ca set de testare şi restul de 5 seturi formează corpusul de antrenare. Precizia medie a POS tagger-ului pentru cuvintele necunoscute este media aritmetică a preciziilor rezultate repetând experimentul pentru toate combinațiile din secțiunile construite. În tabelul 1 sunt precizate valorile obținute pentru preciziile medii.

Tabelul 1. Precizia medie a POS tagger-ului obținută pentru cuvintele necunoscute prin variația frecvenței şi lungimii sufixelor cuvintelor din setul de antrenare

Frecvența de prag	Lungime sufix = 1	Lungime sufix = 2	Lungime sufix = 3	Lungime sufix = 4	Lungime sufix = 5
1	74.14	78.40	77.47	75.20	73.99
3	74.39	78.85	78.58	76.97	75.71
4	74.57	79.02	78.81	77.25	75.95
5	74.54	79.01	78.84	77.31	76.05
6	74.52	78.95	78.93	77.36	76.16
8	74.48	78.91	78.85	77.34	76.13
11	74.41	78.92	78.90	77.41	76.21
15	74.31	78.77	78.76	77.34	76.13

Rezultatele arată că cele mai bune valori pentru precizia medie au fost obținute pentru lungimea sufixelor de 2, apoi de 3, indiferent de valoarea frecvenței de prag. Sufixele de lungime 1 conțin cea mai puțină informație pentru etichetarea corectă a cuvintelor necunoscute, rezultând precizia POS tagger-ului cea mai scăzută. De asemenea, un sufix de lungime mai mare sau egală cu 4 scade precizia POS tagger-ului pentru că se includ şi litere care nu fac parte din terminațiile propriu-zise ale unor părți de vorbire. Pe măsură ce am crescut frecvența de prag până la o anumită valoare, precizia POS tagger-ului pentru cuvintele necunoscute a crescut indiferent de

lungimea sufixelor, după care, crescând în continuare frecvența, precizia a scăzut, dar mult mai uşor față de cum a crescut. Explicația pentru această situație constă în faptul că acele cuvintele care au frecvența de apariție în corpusul de antrenare mai mare pot oferi mai multă informație despre posibilele contexte în care pot fi decât cuvintele care sunt mai rar întâlnite în acelaşi corpus. Pe măsură ce frecvența lor de apariție creşte, modelul probabilistic are tendința să țină cont din ce în ce mai puțin de cuvintele cu o frecvență mai mică. Cea mai bună precizie de 79.02% a fost obținută pentru frecvența de prag 4 şi lungimea sufixelor de 2, aceste valori fiind folosite în testele care urmează.

Următorul test arată cum se modifică precizia POS tagger-ului atunci când dimensiunea setului de antrenare creşte. Folosind diviziunile corpusului Brown, formate pentru testul anterior, setul de antrenare va include pe rând fiecare diviziune. Apoi setul de antrenare va creşte incluzând celelalte părți, păstrându-i-se, totodată, omogenitatea categoriilor de texte componente. Părțile neincluse în corpusul de antrenare vor forma setul de testare. Cele mai semnificative rezultate ale studiului sunt arătate în tabelele 2 şi 3.

Tabelul 2. Precizia POS tagger-ului variind dimensiunea setului de antrenare şi testare

Nr. de propoziții set de antrenare	Nr. de cuvinte set de antrenare	Nr. de cuvinte set de testare	Nr. de cuvinte necunoscute	Precizie cuvinte necunoscute	Precizie cuvinte cunoscute
9583	191803	969389	79698	78.30	95.27
19268	385294	775898	45119	78.44	95.61
29065	578362	582830	27323	78.14	95.85
38817	773237	387955	16704	77.95	95.95
48270	973047	188145	7801	77.79	95.78

Tabelul 3. Precizia POS tagger-ului variind dimensiunea setului de antrenare şi testare

Nr. de propoziții set de antrenare	Nr. de cuvinte set de antrenare	Nr. de cuvinte set de testare	Nr. de cuvinte necunoscute	Precizie cuvinte necunoscute	Precizie cuvinte cunoscute
9070	188145	973047	78250	78.74	95.13
18523	387955	773237	43185	79.69	95.73
28275	582830	578362	27656	79.86	95.86
38072	775898	385294	16055	80.11	96.13
47757	969389	191803	6543	79.5	96.30

Pe măsură ce dimensiunea setului de antrenare a crescut, în toate testele precizia a fost mai mare pentru cuvintele cunoscute. Precizia pentru

cuvintele necunoscute nu are o creştere liniară prin mărirea corpusului de antrenare, ci creşterile de precizie alternează cu scăderea sa. Creşterea de aproximativ cinci ori a setului de antrenare a dus la o îmbunătăţire a preciziei POS tagger-ului pentru cuvintele cunoscute de +0.51% în setul de teste, prezentat în tabelul 2 şi +1.17% în celălalt set de teste, din tabelul 3. Pe aceleaşi seturi de date, în cazul cuvintelor necunoscute precizia a scăzut cu -0.51% într-un set de teste şi s-a îmbunătăţit cu +0.76% în celălalt set de teste. Pentru cuvintele cunoscute, precizia cea mai mare a fost de 96.30%, iar pentru cuvintele necunoscute de 80.11%.

Dacă în testele anterioare am păstrat omogenitatea conţinutului pentru corpusurile de antrenare şi testare, în continuare dorim să arătăm cum variază precizia POS tagger-ului atunci când este antrenat şi testat cu texte din domenii sursă diferite. În tabelele 4 şi 5 am reprezentat matricea cu valorile preciziilor obţinute de POS tagger pentru cuvintele necunoscute, iar în tabelele 6 şi 7, pentru cuvintele cunoscute.

Tabelul 4. Precizia POS tagger-ului pentru cuvintele necunoscute, setul de antrenare/testare fiind o categorie de texte din corpusul Brown

	Cat. A	Cat. B	Cat.C	Cat. D	Cat. E	Cat. F	Cat. G	Cat. H
Categorie A	76.79	77.42	76.94	77.98	74.72	80.45	80.59	75.31
Categorie B	75.19	73.68	74.45	74.45	73.95	78.44	78.80	72.92
Categorie C	73.82	74.33	72.12	76.85	71.19	77.00	78.06	72.29
Categorie D	68.93	70.89	70.50	75.37	68.76	74.28	75.97	72.18
Categorie E	72.12	72.93	73.72	76.14	76.86	76.61	77.29	72.96
Categorie F	75.73	75.88	73.86	76.94	72.67	79.24	79.14	73.21
Categorie G	75.75	74.33	76.25	78.54	71.87	79.99	78.21	71.23
Categorie H	65.46	69.22	66.77	72.90	69.85	70.52	71.80	69.74
Categorie J	72.60	72.53	74.63	74.86	72.12	75.60	77.17	70.21
Categorie K	69.10	71.74	71.26	75.44	69.85	76.43	77.24	69.04
Categorie L	69.68	71.70	70.23	73.11	70.01	75.24	75.84	69.85
Categorie M	60.50	64.65	62.12	67.70	65.19	68.90	69.07	64.26
Categorie N	67.83	69.19	69.73	72.24	69.80	74.72	75.18	67.16
Categorie P	67.27	70.58	70.93	74.48	70.27	75.51	76.06	69.06
Categorie R	68.60	70.30	70.10	73.18	69.34	75.48	75.37	68.17
Precizia medie per categorie de testare	**70.19**	**71.84**	**71.54**	**74.63**	**70.69**	**75.66**	**76.26**	**70.56**

Seturile de antrenare şi testare vor fi setate cu fiecare categorie de texte a corpusului Brown, rezultând o matrice de dimensiune pătratică cu valorile de precizie ale POS tagger-ului, unde pe prima coloană în matrice am reprezentat categoria de text din corpusul Brown folosită pentru setul de

antrenare, iar pe prima linie, categoria de text din corpusul Brown folosită pentru setul de testare.

Tabelul 5. Precizia POS tagger-ului pentru cuvintele necunoscute, setul de antrenare/testare fiind o categorie de texte din corpusul Brown

	Cat. J	Cat. K	Cat. L	Cat. M	Cat. N	Cat. P	Cat. R	**Precizia medie per categorie de antrenare**
Categorie A	75.30	77.18	78.64	79.84	78.84	77.36	77.23	**77.70**
Categorie B	74.94	76.04	74.11	77.13	74.42	73.84	75.69	**75.31**
Categorie C	75.46	74.48	72.97	75.55	74.57	73.13	75.98	**74.69**
Categorie D	75.04	71.43	69.06	72.86	69.96	69.61	72.18	**71.55**
Categorie E	78.49	73.43	71.82	74.77	72.93	73.10	75.19	**74.39**
Categorie F	77.53	77.07	77.41	75.91	77.37	76.96	76.19	**76.13**
Categorie G	75.73	78.24	79.10	79.35	77.16	76.98	78.79	**76.67**
Categorie H	75.78	65.59	62.61	64.72	64.62	60.63	65.98	**67.60**
Categorie J	75.73	73.77	71.02	70.90	72.54	70.40	73.82	**73.01**
Categorie K	72.11	75.84	78.08	77.74	78.69	77.44	75.18	**74.24**
Categorie L	74.08	76.32	73.97	75.36	78.40	76.53	73.85	**73.59**
Categorie M	68.63	71.18	69.60	68.24	72.25	70.35	70.16	**67.47**
Categorie N	72.17	77.31	77.62	76.93	74.75	77.12	74.84	**72.99**
Categorie P	72.69	77.81	77.89	78.45	78.56	75.16	75.00	**73.90**
Categorie R	73.48	75.45	74.52	76.35	76.20	74.44	65.74	**72.93**
Precizia medie per categorie de testare	**74.39**	**74.66**	**73.89**	**75.42**	**74.75**	**73.42**	**74.29**	

Pentru obținerea valorilor preciziei de pe diagonala principală a matricei, textele din respectiva categorie au fost împărțite în proporţii egale pentru seturile de antrenare şi testare. Se observă că pentru unele categorii de texte, precizia obținută este mai mică decât dacă POS tagger-ul ar fi fost antrenat pe întregul corpus al categoriei respective şi testat pe texte dintr-un alt domeniu al corpusului Brown. Acest lucru este datorat faptului că setul de antrenare a avut o dimensiune redusă, iar modelul probabilistic nu a „învățat" suficiente cazuri pentru a face predicția pentru situații noi, chiar şi atunci când textele folosite pentru antrenare şi testare sunt din acelaşi domeniu.

Tabelul 6. Precizia POS tagger-ului pentru cuvintele cunoscute, setul de antrenare/testare fiind o categorie de texte din corpusul Brown

	Cat. A	Cat. B	Cat. C	Cat. D	Cat. E	Cat. F	Cat. G	Cat. H
Categorie A	94.11	94.27	93.89	94.16	93.68	95.19	95.21	92.58
Categorie B	93.70	92.97	93.72	93.72	93.46	94.86	95.20	92.68

Categorie C	92.80	93.41	93.99	93.82	93.33	94.54	94.87	91.20
Categorie D	92.24	93.00	93.54	93.50	93.36	94.63	94.93	91.47
Categorie E	92.90	93.32	93.97	93.94	93.48	95.05	95.16	92.26
Categorie F	93.69	93.99	94.06	94.78	94.22	95.55	95.88	92.73
Categorie G	93.51	94.30	94.49	95.18	93.87	95.78	78.21	92.74
Categorie H	92.73	93.22	93.38	93.85	93.76	94.36	94.71	94.08
Categorie J	93.32	93.71	93.91	94.47	94.09	95.23	95.57	93.47
Categorie K	92.06	92.62	92.92	93.63	92.65	94.27	94.62	90.61
Categorie L	92.41	92.72	92.61	93.23	92.64	94.20	94.23	90.59
Categorie M	92.04	92.40	92.53	92.68	92.38	93.55	93.71	89.97
Categorie N	92.01	92.64	92.41	93.41	92.44	94.12	94.18	89.72
Categorie P	92.29	92.63	92.63	93.37	92.70	94.09	94.45	90.26
Categorie R	92.10	92.21	93.10	93.26	92.71	93.88	94.13	89.89
Precizia medie per categorie de testare	**92.7**	**93.17**	**93.37**	**93.82**	**93.24**	**94.55**	**94.74**	**91.44**

Tabelul 7. Precizia POS tagger-ului pentru cuvintele cunoscute, setul de antrenare/testare fiind o categorie de texte din corpusul Brown

	Cat. J	Cat. K	Cat. L	Cat. M	Cat. N	Cat. P	Cat. R	**Precizia medie per categorie de antrenare**
Categorie A	94.58	95.42	95.21	95.32	95.37	95.29	95.13	**94.66**
Categorie B	94.51	95.23	94.77	94.63	94.62	94.82	94.85	**94.34**
Categorie C	94.07	94.72	94.17	94.24	94.26	94.54	94.59	**93.90**
Categorie D	94.25	94.93	94.06	94.37	94.18	94.55	94.75	**93.88**
Categorie E	94.78	94.92	94.67	94.48	94.70	94.75	94.73	**94.26**
Categorie F	95.08	96.02	95.52	95.48	95.52	95.73	95.29	**94.86**
Categorie G	95.30	96.05	95.45	95.47	95.44	95.82	95.54	**94.92**
Categorie H	94.89	93.76	93.12	93.02	93.28	93.05	93.92	**93.65**
Categorie J	94.62	95.00	94.24	94.45	94.30	94.63	94.54	**94.35**
Categorie K	93.56	95.60	95.81	95.36	95.86	95.89	95.12	**93.93**
Categorie L	93.37	95.93	95.31	95.72	96.01	96.01	95.28	**93.93**
Categorie M	93.08	95.10	94.90	95.17	95.20	95.23	95.09	**93.42**
Categorie N	93.17	95.98	95.99	95.57	95.84	96.01	95.25	**93.78**
Categorie P	93.51	96.07	95.95	95.77	95.98	95.79	95.34	**93.93**
Categorie R	93.43	95.07	94.93	95.14	94.86	95.14	94.53	**93.56**
Precizia medie per categorie de testare	**94.11**	**95.30**	**94.91**	**94.93**	**94.97**	**95.10**	**94.96**	

Ultima coloană din matrice conține precizia medie per categorie de antrenare, iar ultima linie cuprinde preciza medie per categorie de testare. Precizia medie per linie/coloană a fost obținută prin calcularea mediei aritmetice a preciziilor rezultate setând corpusul de antrenare la o categorie de texte şi testând POS tagger-ul pentru celelalte categorii.

Atunci când POS tagger-ul a fost antrenat pe o categorie de texte din corpusul Brown şi testat pe celelalte categorii, cea mai mare precizie pentru cuvintele cunoscute, de 96.07% a fost obținută pentru setul de antrenare cuprinzând romane şi drame romantice şi pentru setul de testare conținând texte de ficțiune. În cazul cuvintelor necunoscute, cea mai mare precizie (80.59%) s-a obținut pentru POS tagger-ul antrenat cu texte de reportaje şi testat pe texte biografice şi eseistice.

În interpretarea rezultatelor din ultimele patru tabele vom ține cont de faptul că anumite categorii de texte din corpusul Brown prezintă multe asemănări din punct de vedere al domeniului la care se referă conținutul lor. Însă, cele mai multe categorii ale corpusului nu sunt similare.

Domeniul de texte cuprinzând romane şi povestiri de ficțiune în sens generic (categoria K a corpusului Brown) include categoriile L, M, N, P, aspect confirmat şi de rezultatele obținute de noi. Astfel, pentru corpusul de antrenare format din categoria K, folosind pentru testarea POS tagger-ului categoriile L, M, N, P, au fost obținute cele mai mari valori ale preciziei de 95.81%, 95.36%, 95.86%, 95.89%, pentru cuvintele cunoscute, iar pentru cuvintele necunoscute valorile au fost de 78.0%, 77.74%, 78.69%, 77.44%. Dintre celelalte categorii de texte folosite pentru testare (pentru aceeaşi categorie de antrenare K), cea mai scăzută precizie a fost obținută pentru categoria H, de 90.61% în cazul cuvintelor cunoscute (69.04% pentru cuvintele necunoscute). Alte asemănări între categoriile corpusului Brown, confirmate de rezultatele POS tagger-ului, sunt între categoriile de antrenare/testare: (L, N), (G, K), (F, K), (A, G).

Valorile preciziilor POS tagger-ului pentru corpusul de testare setat la categoria H şi pentru setul de antrenare format, pe rând, din categoriile L, M, N, P, R sunt mai mici față de celelalte categorii de texte de antrenare, concluzia fiind că aceste categorii de texte nu sunt similare. Rezultate scăzute ale preciziei POS tagger-ului sunt şi pentru perechile de seturi de antrenare/testare: (E, A), (D, A), (H, A), (K, A).

Din punct de vedere al preciziei medii per categorie de antrenare, cele mai bune rezultate, în cazul cuvintelor cunoscute, au fost obținute pentru setul de antrenare format din categoriile de texte: reportaje (94.66%), folclor literar (94.86%), biografii şi eseuri (94.92%), iar cele mai slabe rezultate au fost obținute pentru categoria de texte umoristice (93.56%), urmată de literatura SF (93.42%). Pentru cuvintele necunoscute, POS tagger-ul fiind antrenat cu texte din categoriile folclor literar, biografii şi eseuri, precum şi

reportaje, cele mai bune rezultate au fost de 76.13%, 76.67%, respectiv 77.7%, iar cele mai nesatisfăcătoare rezultate s-au obținut pentru literatura SF (67.47%), scrierile care combină diverse genuri (67.6%) şi textele religioase (71.55%).

Cele mai scăzute rezultate obținute de POS tagger în acelaşi context pot fi analizate prin construcția matricei de confuzie. În matricea de confuzie, denumirile rândurilor şi coloanelor sunt etichete de părți de vorbire. Pentru POS tagger-ul antrenat cu texte din categoria scrieri care combină diverse genuri şi testat pentru categoria romane şi drame romantice din acelaşi corpus Brown, a fost obținută cea mai scăzută precizie, de 60.63%, pentru cuvintele necunoscute. O porțiune din matricea de confuzie corespunzătoare este arătată în tabelul 8.

Tabelul 8. Matricea de confuzie pentru cel mai slab rezultat pentru cuvintele necunoscute

	JJ	NN	NP	VBD	VBN	IN	RB
JJ	76.50	15.02	0.33	0.29	4.34	-	3.48
NN	5.38	92.80	0.13	0.18	0.61	0.02	0.84
NP	3.10	7.18	80.40	0.28	0.14	1.55	7.32
VBD	0.41	3.03	-	65.80	30.50	0.03	0.20
VBN	5.29	2.49	-	7.86	84.20	-	0.15
IN	0.13	0.20	-	-	-	98.81	0.85
RB	3.55	2.91	0.75	0.07	0.03	98.81	89.74

De asemenea, în tabelul 9 este arătată o porțiune din matricea de confuzie pentru cazul în care cea mai scăzută valoare a preciziei pentru cuvintele cunoscute a fost de 89.72%, setul de antrenare fiind format din texte din categoria romane şi povestiri de aventură şi testat pe categoria scrieri care combină diverse genuri.

Tabelul 9. Matricea de confuzie pentru cel mai slab rezultat pentru cuvintele cunoscute

	JJ	NN	NP	VBD	VBN	IN	RB
JJ	86.34	10.57	0.63	0.19	1.24	0.04	0.95
NN	6.95	90.82	1.41	0.24	0.37	0.03	0.14
NP	4.44	3.53	89.63	-	-	-	2.39
VBD	0.49	0.24	-	86.63	12.62	-	-
VBN	5.01	0.32	0.18	27.69	66.69	-	0.09
IN	0.14	0.60	0.11	0.01	-	98.96	0.15
RB	3.48	6.40	0.42	0.07	0.07	5.33	84.19

5. Concluzii

Modelul HMM bazat pe trigrame se poate folosi cu succes pentru etichetarea cu părți de vorbire a cuvintelor unui text, obținându-se o precizie ridicată. Algoritmul foloseşte atât contextul în care apare cuvântul, cât şi informații despre cuvântul în sine, ținând cont de sufixul său şi de frecvența sa de apariție în corpusul de antrenare. Aceste informații sunt necesare pentru construcția arborilor de sufixe. Un factor important în obținerea unei precizii ridicate în cazul unui model probabilistic pentru POS tagging este reprezentat de utilizarea unui set de antrenare de dimensiune cât mai mare, astfel încât modelul să „învețe" cât mai multe situații în care un cuvânt poate apărea în contexte diferite.

Pe măsură ce am mărit dimensiunea setului de antrenare, precizia POS tagger-ului pentru cuvintele cunoscute a crescut în toate testele. Precizia pentru cuvintele necunoscute are o evoluție neuniformă, deşi numărul acestor cuvinte scade. De asemenea, numărul cuvintelor necunoscute a crescut atunci când am antrenat POS tagger-ul pe texte dintr-o anumită categorie şi l-am testat pe alte categorii de texte ale corpusului Brown. De remarcat este faptul că atunci când POS tagger-ul a fost antrenat şi testat pe corpusuri de texte din toate categoriile corpusului Brown, precizia pentru cuvintele cunoscute a fost între 95-96%, iar pentru cuvintele necunoscute, între 78-80%, pe când pentru POS tagger-ul antrenat separat pentru fiecare categorie de texte şi testat pe celelalte categorii de texte, precizia obținută pentru cuvintele cunoscute a scăzut la 90-95%, iar pentru cuvintele necunoscute, la 65-75%.

Pentru viitor, un aspect pe care îl avem în vedere este analiza rezultatelor POS tagger-ului antrenat cu un corpus de texte vechi şi testat pe un set de date recente, identificând diferențele de gramatică ale limbii. O direcție de cercetare interesantă este şi analiza diferențelor de etichetare gramaticală a cuvintelor între un corpus de texte în limba standard şi dialectele acelei limbi.

Referințe

Baayen, H., R. Sproat. *Estimating lexical priors for low-frequency morphologically ambiguous forms.* Computational Linguistics 22(2), pp. 155–166, 1996.

Brants, T. *TnT – A statistical part-of-sppech tagger.* In Proc. of the 6th Applied NLP Conference, 2000.

Brill, E. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics, 1995.

Denis, P., Sagot, B. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In PACLIC, Hong Kong, 2009.

DeRose, S.J. Grammatical category disambiguation by statistical optimization. Computational Linguistics, 1988.

Francis, W.N., Kucera., H. *Brown Corpus Manual.* Providence, Rhode Island Department of Linguistics Brown University, 1979.

Garside, R., Leech, G., McEnery, A. *Corpus Annotation.* Longman, London and New York, 1997.

Giménez, J., Márquez, L. *SVMTool: A general POS tagger generator based on Support Vector Machines.* Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal, 2004.

Hinrichs, L., Smith, N. Waibel, B. Manual of information for the part-of-speech tagged, post-edited „Brown" corpora. ICAME Journal, 34, pp.189-231, 2010.

Jurafsky, D., Martin, JH. Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition. Prentice Hall, 2000.

Leech, G., Garside, R., Bryant, M. *CLAWS4:The tagging of the British National Corpus.* Proceedings of COLING-94, 1994.

Manning, C.D., Schutze, H. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA, 1999.

Marcus, M.P., Santorini, B., Marcinkiewicz, M.A. *Building a large annotated corpus of English: The Penn treebank.* Computational Linguistics, 19(2), 1993.

Samuelsson, C. and Reichl, W. A class-based language model for large vocabulary speech recognition extracted from part-of-speech statistics. In IEEE ICASSP-99, pp. 537–540, 1999.

Shen, L., Satta, G., Joshi, A. *Guided learning for bidirectional sequence classification.* Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL), 2007.

Simionescu, R. *Graphical grammar studio as a constraint grammar solution for part of speech tagging*. In Proceedings of "ConsILR" conference, 2011.

Simionescu, R. *Hybrid POS Tagger.* In Proceedings of "Language Resources and Tools with Industrial Applications" Workshop (Eurolan 2011 summerschool), 2011.

Spoustová, D., Hajič, J., Raab, J., Spousta, M. *Semi-supervised training for the averaged perceptron POS Tagger*. Proceedings of the 12 EACL, 2009.

Søgaard, A. *Semi-supervised condensed nearest neighbor for part-of-speech tagging.* The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), Portland, Oregon, 2011 .

Tufiş, D., Irimia, E., Ion, R., Ceauşu, A. *Unsupervised Lexical Acquisition for Part of*

Speech Tagging In Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008, Marrakech, Morocco, 2008.

Viterbi, A. J. Error bounds for convolutional codes and asymptotically optimal decoding algorithm. IEEE Transactions on Information Theory 13, 260–269, 1967.

Voutilainen, A. *A syntax-based part of speech analyser.* In EACL 7, pp.157-164, 1995.

Voutilainen, A., Heikkilä, J., Anttila. A. *Constraint Grammar of English. A Performance Performance-Oriented*. Publications 21, Department of General Linguistics, University of Helsinki, 1992.

Weischedel, R., Meteer, M., Schwartz, R., Ramshaw, L., Palmucci, J. *Coping with ambiguity and unknown words through probabilisticmodels.* Computational Linguistics 19(2), pp. 359–382, 1993.