Preprocesarea textelor în sistemele de tip „întrebare-răspuns”.

Victoria BOBICEV
Technical University of Moldova
victoria_bobicev@rol.md

***Rezumat* — Sisteme de întrebare-răspuns (Question-Answering systems) au nevoie de procesare intensivă a textului, etapele de preprocesare fiind analiza morfologică şi sintactică. În lucrarea dată este descris procesul de analiză a textelor în sistemul de întrebare-răspuns ce este creat în cadrul proiectului „Cercetarea în domeniul de Regăsire a Informaţiei în scopul creării sistemului electronic de informare publică”. Prima etapă descrisă este marcarea morfologică în baza dicţionarului, a doua este dezambiguizarea morfologică utilizînd metoda statistică. A treia etapă descrisă este analiza sintactică parţială, numită „chunking” ce defineşte grupurile sintactice în propoziţii.**

***Cuvinte cheie* — analiza morfologică automată, analiza sintactică parţială, dezambiguizarea morfologică, procesarea textului, sisteme de întrebare-răspuns.**

I. INTRODUCERE

Sistemele de întrebare - răspuns (în engleză "*question answering systems*", sau sisteme *QA*) sunt caracterizate prin faptul că primesc întrebări formulate în limbaj natural şi, în baza unei colecţii de documente, extrag răspunsul sau un set de răspunsuri găsite în documentele date. Astfel de sisteme sunt considerate ca fiind următorul pas în evoluția motoarelor de căutare a informației în surse textuale [1].

Ca regulă astfe de sistem constă din următoarele module:

1. Modulul de analiză a întrebării – transformă întrebările formulate în limbaj natural uman în interogări pentru motorul de achiziție de documente;

2. Modulul de achiziție de articole – caută în colecția de articole articolele relevante pentru întrebarea formulata de utilizator, în baza datelor primite de la modulul de analiză a întrebării;

3. Modulul de extragere a răspunsului – din colecția de articole returnate de modulul de achiziție de articole, extrage un răspuns succint care constituie răspunsul în limbaj natural uman la întrebarea utilizatorului.

Astfel de sistem are nevoie de pre-procesări elementare asupra textelor din colecția: segmentare lexicală, etichetare morfo-sintactică (POS tagging), recunoaşterea grupurilor frazale (chunking), determinarea dependenţelor între cuvinte sau grupuri frazale (dependency linking) [7].

Sisitemul ce se crează în cadrul proiectului „Cercetarea în domeniul de Regăsire a Informaţiei în scopul creării sistemului electronic de informare publică” la prima etapă va fucționa în baza documentelor prestate de IDSI (Istitutul de Dezvoltare a Societăţii Informaţionale), şi anume:

1. Codul cu privire la ştiinţă şi inovare al Republicii Moldova
2. H O T Ă R Î R E cu privire la aprobarea Acordului de parteneriat între Guvern şi Academia de Ştiinţe a Moldovei pentru anii 2009-2012
3. Cu privire la parcurile ştiinţifico-tehnologice şi incubatoarele de inovare
4. Ştiinţă şi inovare.

Documentele acestea au fost preprocesate cu scopul simplificării căutării şi extragerii răspunsului.

II. MARCAREA MORFOLOGICĂ

Problema marcării automate morfologice a fost studiată intens pe parcursul anilor. Primele încercări de a rezolva problema aceasta sunt bazate pe cunoştinţe [8]. Mai tîrziu au fost efectuate încercările de a aplica metode statistice pentru dezambiguizarea morfologică a cuvintelor textului. Au fost executate încercările de a aplica diferite sisteme statistice în baza lanţului Markov.

În [6] este descrisă o aplicaţie a metodei statistice pentru marcarea automată a textului român cu codurile morfologice. Metoda aceasta este bazată pe modelul Markov şi este practic independentă de limba textului la care este aplicată. Modelul statistic este antrenat în baza unui text marcat cu codurile morfologice, obţinînd statistica secvenţelor din coduri morfologice. Apoi, avînd statistica obţinută metoda aplică algoritmul lui Viterbi pentru dezambiguizarea marcării cuvintelor textului cu codurile respective. Iniţial meoda a fost aplicată pentru textul englez, în [6] aceasta este aplicată pentru marcarea textelor române.

Marcarea morfologică a textelor în rîndul său are nevoie de unele etape preliminare. Una din etapele importante preliminare este împărţirea textului în cuvinte. În cazul textelor procesate etapa dată nu prezintă dificultăţi. În multe cazuri etapa împărţirii textului se uneşte cu marcarea cuvintelor cu codurile morfosintactice cu ajutorul dicţionarului. Dicţionarul utilizat a fost creat la catedra “Informatica Aplicată” şi conţine aproximativ 90 000 de forme de cuvinte. Fiecare formă de cuvînt este însoţită de forma iniţială a lui (lema) şi un cod morfosintactic. Mai jos este prezentat fragmentul dicţionarului dat:

alcoolica	alcoolic	Afsry
abrevierilor	abreviere	Ncfpoy
alegînd	alege	Vmg
ales	alege	Vmp--sm

Fiecare întrare în dicţionarul dat ocupă un rînd şi conţine o formă a cuvîntului, forma lui iniţială (care este utilizată în dicţionare) şi codul morfoogic ce descrie forma dată a cuvîntului. De exemplu, codul Afsry descrie Adjectiv,

feminin, singular, cazul nominativ-acuzativ, articulat, fiecare literă a codului descrie o caracteristică morfologică. Codificarea dată a fost propusă în [4].

Folosind astfel de dicţionar obţinem textul marcat cu codurile morfosintactice. În textul marcat sunt trei tipuri de elemente: cuvintele marcate cu un cod morfologic, cuvintele marcate cu cîteva coduri morfologice – cuvinte ambigue şi cuvintele care nu au fost marcate cu nici un cod morfologic - cuvintele care nu sunt găsite în dicţionar se marchează ca 'unknown'. Un fragment de text marcat este prezentat mai jos:

Condiţiile condiţie/Ncfpry
de =/S
eligibilitate unknown
a avea/Va--3s|al/Tsfs|=/Q
rezidenţilor rezident/Ncmpoy
parcului parc/Ncmsoy
ştiinţifico-tehnologic =/Amsrn|=/Amson

Cuvîntul „a avea/Va--3s|al/Tsfs|=/Q” în exemplul dat este ambiguu, el poate fi verb, articol sau particula ce este indicat de toate codurile posibile delimitate cu linia verticală. În tabelul 1 este prezentată statistica obţinută pentru textele analizate. Din tabel se observă că cuvintele problematice (ambigue şi necunoscute) formează în jur de 20% din text ce înseamnă că din fiecare 10 cuvinte aproximativ unul este ambiguu şi unul necunoscut.

TABELUL I. STATISTICA CUVINTELOR AMBIGUE ŞI NECUNOSCUTE ÎN TEXTELE ANALIZATE DUPĂ PRIMA ETAPĂ DE MARCARE

Document	**N de cuvinte ambigue**	**N de cuvinte necunoscute**	**total cuvinte**
Codul cu privire la ştiinţă şi inovare	5340 (15%)	2856 (8%)	35552
Hotărîre cu privire la aprobarea Acordului de parteneriat	4179 (10,5%)	3928 (9,9%)	39723
parcurile ştiinţifico-tehnologice şi incubatoarele de inovare	651 (16%)	283 (7%)	4034
Ştiinţă şi inovare	5127 (15%)	2670 (7,9%)	33622
Total	15297 (13,5%)	9737 (8,6%)	112931

Pasul următor este dezambiguizarea morfologică.

III. DEZAMBIGUIZAREA MORFOLOGICĂ

Modele statistice pentru rezolvarea problemelor prelucrării automate a limbajului natural sunt create în aşa mod ca să reprezinte regularităţile statistice prezentate în limbă. Cercetătorii au observat că modelele create pentru comprimarea textului sunt foarte asămănătoare cu modelele create pentru prelucrarea lui. Pentru comprimare se crează un model în baza aceluiaşi lanţ lui Markov. Astfel este logic de presupus că modelele create pentru comprimare pot fi folosite pentru rezolvarea problemelor prelucrării limbajului natural. Metoda PPM (prediction by partial matching – prezicere prin corespundere parţială) este o metodă de comprimare în baza contextului limitat (finite-context modeling), care estimează probabilitatea simbolurilor în baza contextului - simbolurilor precedente în text. În [5] a fost descrisă implementarea practică algoritmului optimizat PPMC, care este considerat cel mai bun algoritm de comprimare a textelor. PPM prezintă o variantă de amestecare cînd probabilităţile obţinue în baza contextelor cu lungimea diferită sunt unite într-o probabilitate comună. În cazul general probabiltatea amestecată a simbolului curent **p(s)** poate fi calculată ca:

$$\mathbf{p(s)}= \prod_{i=1}^{o} \mathbf{p'(c_i|\, c_{i-o}\, c_{i-o+1}\, c_{i-o+2} \ldots c_{i-1})} \qquad \mathbf{(1)}$$

unde

$\mathbf{p'(c_i|\, c_{i-o}\, c_{i-o+1}\, c_{i-o+2} \ldots c_{i-1})}$ – probabilităţile simbolului curent c_i, determinate de model pentru toate contextele, începînd cu contextul maximal *î*.

Modelul acesta poate fi creat în baza literelor textului, în baza cuvintelor sau codurilor morfologice. În cazul nostru modelul este creat în baza secvenţelor de coduri în text marcat cu codurile morfologice.

Modelul PPM descris este aplicat pentru rezolvarea problemei ambiguităţii cuvintelor, selectarea codului morfologic potrivit pentru fiecare cuvînt. Cuvintele ambigue ca regulă au două sau trei coduri ataşate din care sistemul trebuie să selecteze unul corect. Cuvintele necunoscute prezintă cea mai mare problemă pentru sistem fiindcă pentru cuvintele acestea el trebuie să selecteze din toate codurile posibile. În experimentele efectuate noi intenţionat am restrîns numărul de coduri posibile pentru cuvintele necunoscute cu scopul micşorării timpului lucrului programului. În marea majoritate cuvintele necunoscute aparţin părţilor de vorbire 'deschise' - substantivelor, adjectivelor şi verbelor. Acestea sunt atît de numeroase, că practic nu pot fi prezentate în dicţionar în întregime.

În metoda descrisă pentru găsirea codului potrivit este folosită o variantă a algoritmului Viterbi [3]. În urmare este prezentat algoritmul folosit în experimentele executate.

Iniţializarea:
Se introduc patru semne speciale care formează contextul primului cuvînt. Pentru primul cuvînt al propoziţiei se calculează probabilităţile tuturor codurilor posibile pentru cuvîntul dat:
$P(t^k_1) = P_{PPM}(t^k_1 \mid B\ B\ B\ B)$ pentru toate $t^k \in t^1 \ldots t^n$;
Iteraţii:
Pentru toate cuvintele propoziţiei w_1 de la al doilea pînă la ultimul $w_2 \ldots w_p$:
Pentru toate codurile posibile cuvîntului dat $t^k_1 \in t^1 \ldots t^n$:
Pentru toate codurile posibile cuvîntului precedent $t^l_{1-1} \in t^1 \ldots t^n$:
Calculăm recursiv probabilitatea codului curent
$P(t^k_1) = P_{PPM}(t^l_{1-1})\, P_{PPM}(t^k_1)$;
Dacă probabiltatea aceasta este maximală, memorizăm valoarea codului precedent în vectorul codurilor finale:
$P(t_{1-1}) = t^l_{1-1}$ pentru care $\max P(t_1)$;
Sfîrşit
Sfîrşit
Sfîrşit
Pentru toate codurile posibile ultimului cuvînt al propoziţiei $t^k_p \in t^1 \ldots t^n$:
Aflăm valoarea maximală a probabilităţii $P(t_n)$

```
     Şi memorizăm codul corespunzător în vectorul
rezultatelor finale.
       Sfîrşit
```

Folosind metoda PPM cu contextul maximal patru coduri precedente am efectuat antrenarea sistemului în baza corpusului de texte marcate morfologic, verificate şi corectată manual. În baza probabilităţilor calculate cu ajutorul algoritmului descris mai sus am executat dezambiguizarea automată a fişierelor deja analizate. Rezultatele obţinute verificate manual sunt prezentate în tabelul 2.

TABELUL 2. STATISTICA CORECTITUDINII DEZAMBIGUIZĂRII MORFOLOGICE ÎN TEXTELE ANALIZATE

Document	**N de cuvinte prelucrate**	**N de cuvinte dezambiguiza-te corect**	**N de cuvinte dezambiguiza te greşit**
Codul cu privire la ştiinţă şi inovare	22,5%	18,9%	3,6%
Hotărîre cu privire la aprobarea Acordului de parteneriat	19,5%	13,3%	6,2%
parcurile ştiinţifico-tehnologice şi incubatoarele de inovare	23,8%	20%	3,8%
Ştiinţă şi inovare	18,8%	13,8%	5%
Total	21,15%	16,5%	4,65%

Din tabel se observă că cuvintele prelucrate sunt cele ambigue şi cele necunoscute (8% + 13% = 21%). Numărul de cuvinte prelucrate este cu o fracţie de procent mai mică decăt suma celor ambigue şi necunoscute din cauză că unele necunoscute nu sunt de fapt cuvinte. În text se întîlnesc unele elemente care nu au fost recunoscute de sistem şi se considerau cuvinte necunoscute (de exemplu, cifre romane: III, V, X). Aceste erori nu au fost luate în consideraţie pe parcursul evaluării calităţii dezambiguizării din cauză că erorile au fost comise la prima etapă de marcare.

A treia coloniţă a tabelului conţine procentul cuvintelor cu coduri atribuite greşit în procesul de dezambiguizare. Se observă că procentul acesta diferă considerabil de la un document la altul. Media cuvintelor cu codul atribuit greşit este 4,65% ce de fapt presupune aproape un cuvînt greşit din fiecare 20 cuvinte din text. Însă analiza mai detaiată a cuvintelor cu coduri greşit detectate arată că în unele cazuri partea de vorbire a fost determinată corect, eroarea a fost comisă în detectarea unelor caracteristici morfologice, de exemplu, cazului sau numărului substantivului. Astfel de erori afectează analiza etapelor următoare, şi anume, analiza sntactică. Dacă calculăm procentajul erorilor grave, cînd este greşit detectată partea de vorbire, obţinem 3,15% din numărul total al cuvintelor textului. Rezultatul acesta este destul de bun şi practic la acelaşi nivel ca şi alte sisteme de dezambiguizare morfologică [6].

IV. CHUNKING

Etapa următoare după analiza morfologică finalizată este analiza sintactică, însă această sarcină este extrem de complicată şi practic nu poate fi realizată în baza textelor documentelor date. Astfel, în locul analizei sintactice totale este propusă analiza sintactică parţială care a obţinut denumirea "chunking" provenind de la cuvîntul "chunk" – bucată, fragment. Textul este împărţit în fragmente care reprezintă grupuri sintactice, şi anume, grup nominal grup verbal, grup prepoziţional şi altele.

În [2] a fost descrisă metoda de creare a gramaticii pentru analiza sintactică parţială în mod semiautomat. Gramatica dată prezintă un set de reguli sintactice ce sunt aplicate la un text marcat morfologic. Regula sintactică a gramaticii date este formată din coduri morfologice şi reprezintă o secvenţă a codurilor ce se unesc într-un grup. De exemplu, pentru grup nominal se formează astfel de reguli:

Nc Ts Ncmsoy Ams
Nc Ts Ncmpoy Amp
Nc Ts Ncmsoy Ams
Nc Ts Ncmpoy Amp

în care sunt descrise codurile cuvintelor consecutive: Nc (Noun, common) – substantiv comun, Ts – articol posesiv, Ncmsoy – substantiv comun, masculin, singular, cazul dativ-genitiv, articulat, Ams – adjectiv, masculin, singular. Regulile date specifică caracteristicile morfologice ale cuvintelor care indică acordul între cuvinte în grupul dat. De exemplu, în grupul nominal este indicat acordul în gen şi număr între substantiv şi adjectiv.

În lucrarea menţionată au fost create doar reguli pentru grupuri nominale. Grupuri nominale sunt de fapt elementele principale în text, mai ales în textele formale care sunt procesate în lucrarea dată. Statistica părţilor de vorbire în textele date demonstrează că numărul de substantive în text este aproximativ de cinci ori mai mare decît numărul de verbe. În propoziţie pe lîngă un verb apar aproximativ cîte cinci substantive care formează grupuri nominale complicate cu articole, adjective şi pronume asociate. Specificul documentelor de aşa tip este aşa numită „nominalizarea" verbelor, cînd acţiunea în propoziţie este redată de substantiv format din verb (verb nominalizat). De exemplu: „a accepta condiţiile" se transformă în „acceptarea condiţiilor", „a asigura integritatea" respectiv în „asigurarea integrităţii", „a aproba acordarea statutului" – în „aprobarea acordării statutului". De menţionat că în ultimul exemplu sunt două verbe nominalizate şi astfel de combinaţii pot fi prelungite din care cauză este uneori greu de înţeles mesajul documentelor formale. La fel de grea este procesarea automată a acestui tip de documente. Una din problemele de bază în traducerea automată este anume procesarea grupurilor nominale compuse din mai multe substantive („compound nouns").

Altă problemă dificilă prezintă grupurile prepoziţionale. Prepoziţiile joacă un rol important în formarea lanţurilor de grupuri nominale, ca de exemplu: „domeniile de activitate prevăzute de proiectele de inovare şi transfer tehnologic". Însă în procesul unirii noi ne conducem de sens şi deosebim cazul „a fost aprobat proiectul de Consilium Suprem" şi cazul „a fost aprobat proiectul de transfer

tehnologic". În primul caz „proiectul" şi „de Consilium Suprem" sunt două grupuri ataşate ambele la grup verbal: „a fost aprobat proiectul" şi „a fost aprobat de Consilium Suprem", într-a doilea caz este un grup „proiectul de transfer tehnologic". Problema ataşării grupurilor prepoziţionale este la fel una din problemele nerezovate în procesarea limbajului natural. În lucrarea de faţă problema aceasta la fel nu este rezolvată; rezolvarea ei este lăsată pe viitor.

Astfel, pentru documentele date au fost formate 53 reguli sintactice, 45 pentru grupuri nominale şi 8 pentru grupuri verbale. Textele documentelor analizate au fost marcate utilizînd regulile date. În urmare este prezentat un fragment de text marcat.

```
<CHUNK type=NP morph=Ncfpry head= Condiţiile len=1>
        Condiţiile   condiţie/Ncfpry
</CHUNK>
        de          =/S
<CHUNK type=NP morph=Ncfsrn head=încetare len=3>
        încetare    ???/Ncfsrn
        al/Tsfs
        statutului  statut/Ncmsoy
</CHUNK>
        de          =/S
<CHUNK type=NP morph=Ncmsrn head=rezident len=4>
        rezident    =/Ncmsrn
        a           al          =/Tsms
        parcului    parc/Ncmsoy
        ştiinţifico-tehnologic  =/Amsrn
</CHUNK>
        şi          =/C
<CHUNK type=NP morph=Ncmsoy head=statutului len=2>
        a           al/Tsfs
        statutului  statut/Ncmsoy
</CHUNK>
        de          =/S
<CHUNK type=NP morph=Ncmsrn head=rezident len=3>
        rezident    =/Ncmsrn
        al          =/Tsms
        incubatorului           incubator/Ncmsoy
</CHUNK>
        de          =/S
<CHUNK type=NP morph=Ncfsrn head=inovare len=1>
        inovare     =/Ncfsrn
</CHUNK>
```

Din fragmentul de text prezentat se observă că fiecare grup sintactic este marcat cu tagul <CHUNK> </CHUNK>, elementul iniţial conţine atribute type ce indică tipul grupului, head şi morph care indică elementul principal în grupul dat şi caracteristicile lui morfologice. Informaţia aceasta este necesară pentru etapele următoare de analiză sintactică care presupune unirea grupurilor date în grupuri mai mari, repetînd acţiunea de unire pînă ce toată propoziţiava fi analizată. În procesul acesta este important de luat în consideraţie informaţia morfologică şi de respectat acordul morfologic între elementele propoziţiei. La fel de importantă este detectarea elementului principal în grupurile deja marcate. Din exemplul prezentat se observă că o mare parte a textului este marcată, în afara tagurilor rămîn preponderent prepoziţiile. Prepoziţiile se unesc cu grupuri nominale care le urmează şi formează grupuri prepoziţionale. De exemplu: „de încetare al statutului" constă din prepoziţie „de" şi grup nominal „încetare al statutului". În cazul dat regula este unică. Mai dificilă este problema unirii grupului prepoziţional la textul ce se află de partea stînga a prepoziţiei, la textul precedent. Un exemplu de ataşare problematică este deja prezentat în articolul dat.

V. CONCLUZIE

Sisteme de întrebare-răspuns (Question-Answering systems) au nevoie de procesare intensivă a textului, etapele de preprocesare fiind analiza morfologică şi sintactică. În lucrarea dată este descris procesul de analiză a textelor în sistemul de întrebare-răspuns ce este creat în cadrul proiectului „Cercetarea în domeniul de Regăsire a Informaţiei în scopul creării sistemului electronic de informare publică". Prima etapă descrisă este marcarea morfologică în baza dicţionarului, a doua este dezambiguizarea morfoogică utilizînd metoda statistică. Metoda statisitcă propusă rezultă în text marcat cu rata aproximativă de erori 3-5%, ce este la nivel cu alte sisteme pentru limba româna. A treia etapă descrisă este analiza sintactică parţială, numită „chunking" ce defineşte grupurile sintactice în propoziţii.

REFERINŢE

[1] R. Baeza-Yates, B. Ribeiro-Nieto, Modern Information Retrieval, Addison Wesley, 2000.

[2] Bobicev V. Metoda semiautomată de creare a gramaticii pentru analiza sintactică a grupurilor nominale în textele române. International Conference Trends in the Development of the Information and Communication Technologies in Education and Management, ASE, Moldova, 2003, pp. 301-304.

[3] Bobicev V., Popescu A. Marcarea morfo-sintactică automată folosind modelul PPM. Conferinţa jubiliară Tehnico-ştiinţifică a colaboratorilor, doctoranzilor şi studenţilor consacrată celei de-a 40-a aniversări a doctoranturii U.T.M., Moldova, 2006, pp. 155-158

[4] Tomaz Erjavec, Nancy Ide, Dan Tufis: Development and Assessment of Common Lexical Specifications for Six Central and Eastern European Languages, ALLC-ACH '98 Conference, 1998.

[5] Moffat, A., 1990. Implementing the PPM data compression scheme. IEEE Transactions on Communications, Vol. 38, No. 11, pp. 1917-1921.

[6] Dan Tufiş, Oliver Mason: Tagging Romanian texts: A Case Study for QTAG, a Language Independent probabilistic tagger, First International Conference on Language Resources and Evaluation, 1998.

[7] Dan Tufiş, Dan Ştefănescu, Radu Ion, and Alexandru Ceauşu. RACAI's Question Answering System at QA@CLEF 2007. In Alessandro Nardi and Carol Peters, editors, Working Notes for the CLEF 2007 Workshop, pages 15–21, 2007.

[8] Dan Tufiş, Octav Popescu, "A Knowledge-Based Approach to Morpho-lexical Processing of Natural Language", in Proceedings of the International Conference for Young Computer Scientists, Beijing, 1991.