

USE OF ISOTOPES IN THE ONLINE COMMENTATORS' CATEGORIZATION

Daniela Gîfu, Scientific Researcher, PhD, "Al. Ioan Cuza" University of Iași

Abstract. The paper presents a study of automatic recognition of discursive isotopy on anonymous readers' comments on online articles, having an important role in the semantic disambiguation. The isotopes ensure also the discursive coherence, which helps to understand the message included in the utterance. The aim of this survey is to identify the isotopes in the online press articles, and the anonymous readers' comments on them, followed by the comparing phase, comment vs. article. Moreover, the paper proposes a methodology concerning the categories of online commentators, through the overlapping of isotopes. Relying upon the theory of discursive isotopes and semantic hierarchies in argumentation, this research establishes a typology able to define an automated evaluation grid. The manually annotated corpus issued from the press monitoring will be the Gold Corpus on which our software applies the lexical-semantic isotopy. This research supports the direct beneficiaries as PR specialists, politicians' counsellors, managers, editors etc., in their effort of public segmentation, but also specialists in the field of natural language processing and linguists.

Keywords: *isotopes, discourse, online press, anonymous readers' comments, commentators' categorization.*

1. Introducere

Lucrarea propune o nouă abordare de clasificare a comentatorilor de pe formurile ziarelor online, pornind de la noțiunea de *izotopie*, esențială pentru analiza discursului, discursul putând fi asociat cu textul (Ricœur, 1995; Plett, 1983).

Opțiunea pentru o asemenea problemă vine din nevoia completării profilului opinitorului online (Gîfu & Cioca, 2013, Gîfu et al., 2013) pornind de la analiza comparativă a două texte, articolul din presa scrisă online cu fiecare comentariu postat pe forumul acelei publicații, calculând frecvența repetării unor unități lingvistice (izotopie) între acestea. Cu alte cuvinte, ne propunem să măsurăm gradul de acoperire a temei unui articol în comentariile pe care le atrage.

Multitudinea de articole, ale căror abordări adesea sunt diametral opuse, chiar dacă pornesc de la un subiect comun, dă naștere la comportamente civice diferite. De pildă, atunci când individul este interesat de subiectul arestării unui anumit ministru (criză de imagine) va începe să caute rapid articole în presa online. Probabilitatea ca un comentator să intervină pe forumul unui ziar care, fie are o argumentare (foarte) asemănătoare cu opiniile jurnalistului, fie se raportează la capătul opus, este extrem de mare. Și, cum a comenta pe forum a devenit o activitate cotidiană, o asemenea analiză considerăm că estede actualitate și, totodată, de interes pentru aria de cercetare interdisciplinară centrată pe prelucrarea limbajului natural (PLN).

Premisa prezentului studiu este aceea că analiza corespondenței dintre textul unui articol și blocul de comentarii de pe forumul publicației online, unde au fost postate, divulgă atitudinea civică, care poate fi ierarhizată.

Lucrarea este structurată în cinci capitole. După o scurtă introducere cu privire la tema propusă, în capitolul doi reținem cele mai importante contribuții pe tema izotopiilor. În capitolul trei amintim câteva aspecte teoretice cu privire la conceptul de izotopie pentru o mai bună contextualizare a cercetării, urmând ca în capitolul patru să descriem metodologia de lucru centrată pe categorizarea comentatorilor de pe forumurile presei online, utilizând o aplicație computațională aflată în faza de implementare. Ultimul capitol reține concluziile prezentului studiu, lanțul de preprocesări textuale bazându-se pe instrumentele implementate de Grupul de Cercetare în Tehnologii ale Limbajului Natural al Facultății de Informatică de la Universitatea “Alexandru Ioan Cuza” din Iași (NLP-Group@UAIC-FII).

2. Context

Dacă înțelegem textul ca manifestare a libertății de exprimare, pornind de la ipoteza ambivalenței constitutive a semnului-text, textul poate fi tratat ca „entitate eminamente verbală, în același timp parte a unui proces semiotic complex și reflex al acestuia” (Vlad 2000, p. 22). Analiza de text, din perspectiva semiotică, propusă de corifeul semiologiei franceze, Greimas (Greimas, 1966), oferă azi cadrul propice de colaborare dintre lingviști și informaticieni, conturându-se o nouă disciplină, lingvistica computațională, al cărei scop este acela de dezvoltare a teoriei computaționale a limbajului făcând uz de elemente ale informaticii.

Metoda lui Greimas devine tehnica de bază a analizei semiotice, influențând precursorii Școlii de la Paris, Barthes (Barthes, 1970), Derrida (Derrida, 1974) etc. Teoria se bazează pe ipoteza existenței unui univers semantic, cu referire la „totalitatea semnificațiilor, postulată pentru un mesaj” (Greimas & Courtés, 1982, p. 361), semnificații ce pot fi desprinse dintr-un text (*izotopia semantică* sau de *profundime*¹). Este și unul dintre obiectivele noastre viitoare, de a interpreta semantic textele (intersecția definițiilor cuvintelor-cheie recunoscute în articol și în comentarii), pornind de la metodele analitice tradiționale, prin utilizarea unei aplicații computaționale.

Subiectul propus se încadrează într-o dezbatere mai amplă, care implică literatura critică în etapa de (re)productibilitate (Riva, 2011), sugerând o regândire a modelelor și metodelor de hermeneutică textuală, folosind o modalitate digitală (Ciotti & Crupi, 2012). Noutatea constă în a face analiza naratologică prin observarea rezultatelor ei macrostructurale și microstructurale (stiluri, lexeme, izotopii), propunând un model hermeneutic care permite indexarea semantică a familiilor și izotopiilor, deductibile prin concepte generale: loc, spațiu, caracter și identitate.

Apariția unui număr din ce în ce mai mare de instrumente și tehnologii - care permit stocarea datelor textuale și a edițiilor electronice în diferite formate (RTF, PDF etc.) și, implicit analizarea lor (cantitativă, mai ales) - scade potențialul hermeneutic computerizat, atunci când textul este împărțit în *atomi* cu aceeași semnificație (Trevisan, 2008). În plus, critica textuală este lipsită adesea de dimensiunea istorică, o soluție fiind TEI (*Text Encoding Initiative*), care codifică unele caracteristici semantice în textele moderne. De pildă, Laboratorul Crilet al Facultății de Arte din cadrul Universității “Sapienza” din Roma, propune extinderea interpretării documentelor (Mordenti, 2007), folosind transcriere digitală și

¹ Este vorba de intersecția definițiilor cuvintelor-cheie obținute într-o primă etapă, cele desprinse din articol și cele care reies din fiecare comentariu de pe forumul unde a apărut publicat.

reformulare prin marcare semantică. O combinație pragmatică între istorie și semiotică, astfel încât documentul digital reprezintă, în interior, modelul interpretativ. Cu alte cuvinte, este posibil ca un corpus narativ să aparțină mai multor familii de sensuri, putând fi analizat: *vertical*, studiind sortarea lexicală de la frecvență maximă hapax-legomena; *semantic*, studiind frecvență și poziția izotopiilor selectate în text (Greimas, 1970); *alfabetic*, generând o ordine alfabetică pentru a identifica familii de sensuri. Având construit un sistem centrat pe text, este util pentru a începe gândirea critică adăugând marcarea XML pentru link-uri către site-urile cu referințe istorice. Modelul TEI explică principiile de codificare ale textului (Cummings, 2007, Romary, 2009, Vanhoutte și Van den Branden, 2010).

Studiul de față dorește să demonstreze potențialitatea analizei textuale, evidențiind natura interdisciplinară a demersului metodologic ce va fi descris.

3. Izotopie – concept și tipologie

Fără a ne propune o descriere prea detaliată a cadrului teoretic, scopul acestui studiu constă în demonstrarea eficienței utilizării unei metode computaționale asupra configurării izotopice rezultate din procesarea automată a unui corpus (articole de presă și comentariile aferente). Devine o metodă eficientă descrierii universului civic identitar în mediul virtual. Practic, izotopiile generează ideile pe care vrem să le extragem și să le comparăm în vederea categorizării profilurilor comentatorilor pe forumurile presei scrise online.

3.1. Originea conceptului de izotopie

Dacă în *semantica structurală*², izotopia³ descrie coerența și omogenitatea textelor, în sens larg, este privită ca un principiu de construcție a textului în semiotică. Coerența ne dă răspunsul la întrebarea „în ce condiții, un discurs, adică o suită de enunțuri, poate fi considerat bine format sau coerent” (Moeschler & Reboul, 1999: 423), ea presupunând cel puțin două aspecte: regulile de coerență și dimensiunile fundamentale ale coerenței discursului (temporalitatea și referențialitatea) (Chiorean, 2007).

Termenul de izotopie a fost transferat pentru prima oară în analiza semantică de Greimas, în 1966, ca fiind „ansamblul redundant de categorii semantice care face posibilă lectura uniformă a povestirii, așa cum rezultă ea din lecturile parțiale ale enunțurilor după reducerea ambiguităților acestora, această reducere fiind ea însăși ghidată de căutarea lecturii unice.” (Greimas, 1970; 1975: 3). Reducerea ambiguității implică, în cele din urmă, căutarea unei interpretări unice a unui text (i.e. coeziune). Cu alte cuvinte, coeziunea trimite spre logica enunțării, fiind “capacitatea elementelor verbale explicite (co-prezente) de a realiza legături intra-textuale de sens. (Vlad, 2000: 71). Spunem că un text este coeziv, “dacă există relații propoziționale între enunțurile care îl compun” (Moeschler & Reboul, 1999: 440).

Conceptul de *izotopie* a fost extrem de influent fiind revizuit și redefinit de pionierii *semanticii interpretative* (Rastier, Arrivé, Groupeu, Kerbart-Orecchinoni, Eco etc.).

² Despre latura metodologică și teoretică a structuralismului a scris Roventă-Frumușani în cartea *Semiotică, societate, cultură*.

³ Împrumutat din fizică și chimie, conceptul de *izotop* numește elementele cu același număr atomic, dar cu numere de masă diferită.

În definirea izotopiei, Rastier nu se îndepărtează de tendința delimitată de Greimas, considerând-o „orice iterație a unei unități lingvistice” (Rastier, 1972: 80), putând fi analizată dintr-un punct de vedere sintagmatic, dar nu și dintr-unul sintactic. Pentru ambii autori, elementul comun este *iterația*, *izotopia* referindu-se la forma conținutului sau la cea a expresiei unui text, din care se pot desprinde și clasificări ale acestei noțiuni. Prin urmare, izotopia desemnează *iterativitatea* unor claseme care asigură discursului-enunț omogenitatea, de-a lungul unui lanț sintagmatic. Eco arată defectele de utilizare a "repetiției", înlocuind-o cu "direcția", redefinind izotopia ca "direcția luată de o interpretare a textului".

Pentru a stabili o izotopie este nevoie de stabilirea unui context minimal (ex. o sintagmă care reunește cel puțin două figuri semice). Categoria semică subsumează doi termeni contrari. Luând în considerare parcursurile cărora le pot da naștere, la nivel tematic, instrumentul care evidențiază valorile vehiculate de un text, pe care vrem să-l analizăm, este careul semiotic – ai căror patru termeni se vor numi izotopii (fig. 1) – considerat o structură binară „care permite vizualizarea organizării profunde a semnificației, cu ajutorul a trei tipuri de relații logico-semantice interdependente.” (Everaert-Desmedt, 2007: 74). Careul semiotic rămâne un instrument esențial în analiza semiotică, constituit pe baza unei axe semantice care se articulează în jurul valorilor S_1 și S_2 , care sunt contrare. (Everaert-Desmedt, 2007: 74)

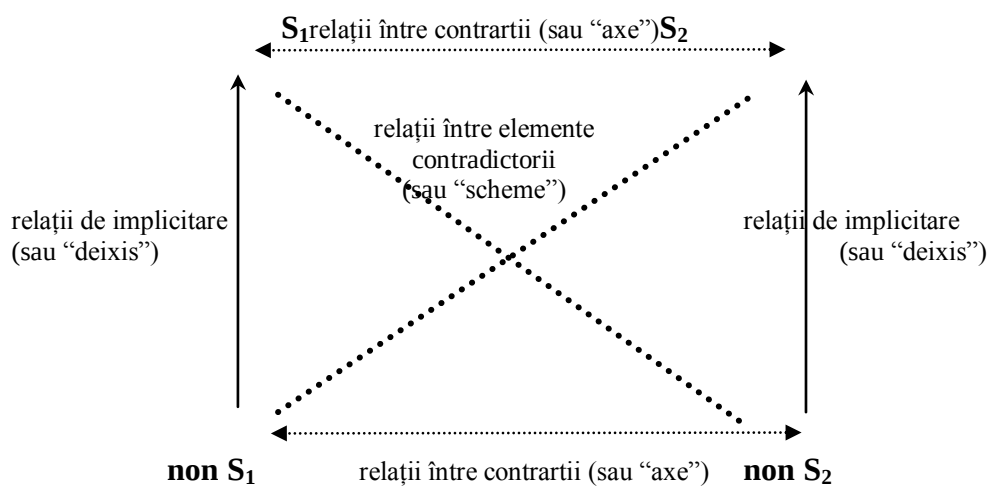


Fig. 1 Careul semic

Dacă ne oprim la forma conținutului, Greimas amintește de izotopie *semantică* (ex. textele poetice) și izotopie *rațională* (ex. textele științifice), diferențierea făcându-se în funcție de valoarea de adevăr. În schimb, dacă ne raportăm la forma expresiei, izotopia iese din cadrul textual, sensul fiind o funcție a intertextului (efectul lecturii). O altă ierarhizare importantă la cercetarea propusă este cea raportată la componenta tematică și cea figurativă a discursului, Greimas și Courtés diferențiind izotopiile *tematice* (situate la nivel profund al parcursului generativ) de izotopiile *figurative* (sprijină configurațiile discursive).

O extindere a conceptului de izotopie o face Kerbart-Orecchinoni pentru a indica repetarea nu numai a unui Sem, ci și a altor unități semiotice (ex. foneme).

Fiind preocupați de parcursul generativ al discursului, reținem și descriem sumar distincția dintre *izotopia sintactică* (recurența categoriilor aferente) și *izotopia*

semantică(favorizează lectura uniformă a discursului, rezultată din lecturile parțiale ale enunțurilor lui), tipologie ce stă la baza metodologiei propuse.

3.2. *Izotopii sintactice*

Cunoscute drept *izotopii de suprafață*, acestea pot fi determinate după frecvența de ocurență a unor termeni, cunoscuți drept *cuvinte-cheie* sau *taguri*, prezente la fiecare articol din presa scrisă online. Pentru cititori - care, de cele mai multe ori, devin și comentatori pe forum – cuvintele-cheie au un rol extrem de important în stabilirea imediată a subiectului articolului fără a parcurge întreg conținutul și, implicit, să-și aleagă articolele care-l interesează. De asemenea, aceste taguri au rolul de a descrie pe scurt motoarele de căutare⁴ conținutul (informația) adăugat. De altfel, în această lucrare, aplicația măsoară frecvențele de ocurență a cuvintelor-cheie (izotopii sintactice) extrase din corpusul analizat pe baza cărora sunt determinate clasele de forumiști la fiecare nivel al ierarhiei stabilite.

3.3. *Izotopii semantice*

Cunoscute drept *izotopii de profunzime*, acestea pot fi identificate la intersecția definițiilor cuvintelor-cheie obținute în etapa precedentă. Altfel spus, prin studierea redundanțelor unităților formale de conținut, când sememele aceluiasi câmp sunt articulate între ele prin relații logice identificabile, câmpul sememic e structurat în cod, iar dacă diverse codaje sunt parțial izomorfe, se constituie izotopii semantice prin redundanța sememelor, ocupând același rang logic în codurile izomorfe. Recunoașterea lor devine unul dintre obiectivele de viitor pentru dezvoltarea a acestei aplicații.

4. Metodologia de lucru

Fie că izotopia este provocată de intenția conștientă a locutorului, fie că nu, structura însăși a manifestării izotopiei rămâne aceeași. Dat fiind numărul mare de izotopii pe care le extrage mașina (fără a lua automat în considerare tagurile date de Internet), am inclus anumite reguli pentru a limita variațiile la nivel de suprafață pe care le vom descrie mai jos.

4.1. *Corpus*

În perioada 5-11 ianuarie 2015 au fost monitorizate și stocate articolele și comentariile în legătură cu acestea din publicația *Adevărul*. Dintre acestea au fost prelucrate semi-automat 100 de articole, împreună cu 1980 de comentarii postate pe forum la trei ore după publicarea articolelor. O pătrime a acestui corpus a fost adnotat manual și trei pătrimi automat.

Verificarea rezultatelor procesării automate (a doua pătrime) a stat la baza îmbunătățirii formulelor ce urmează a fi descrise. Celelalte două pătrimi au fost procesate automat după revizuirea algoritmului.

4.2. *Descrierea instrumentului computațional*

⁴Un exemplu este relevanța informației pentru agregatorii de conținut precum *Yahoo News*, care preiau și postează doar știrile care au un conținut important, ceea ce înseamnă publicitate rapidă, gratuită și la scară impresionantă.

IARC (*Izotopes of Anonymous Readers' Comments*) este o aplicație implementată de curând la Facultatea de Informatică⁵ a Universității „Alexandru Ioan Cuza” din Iași (UAIC) care are o funcționalitate simplă, dar de mare utilitate, având în vedere volumul uriaș de comentarii regăsite pe forumurile ziarelor online.

În momentul încărcării corpusului în aplicație, orice element (grafic sau text), care nu face obiectul corpusului pe care vrem să-l analizăm, este eliminat.



Fig. 2 Titlu articol înainte de încărcare în IARC

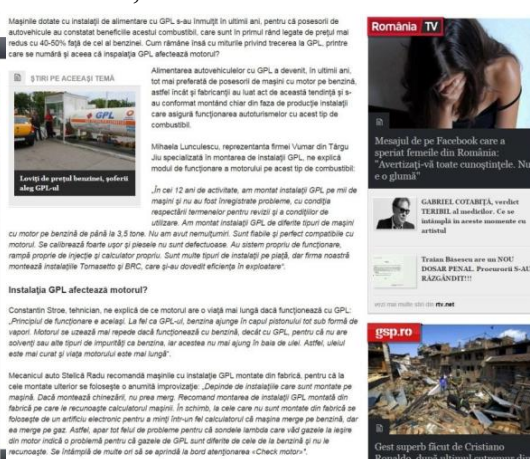


Fig. 3 Conținut articol înainte de încărcare în IARC

Spre ex, dacă pagina unui ziar online arată ca în fig. 1, fig. 2, în aplicație va fi preluat exclusiv corpul de text (v. fig. 2 - articolul comentat și fig 3 - comentariile).

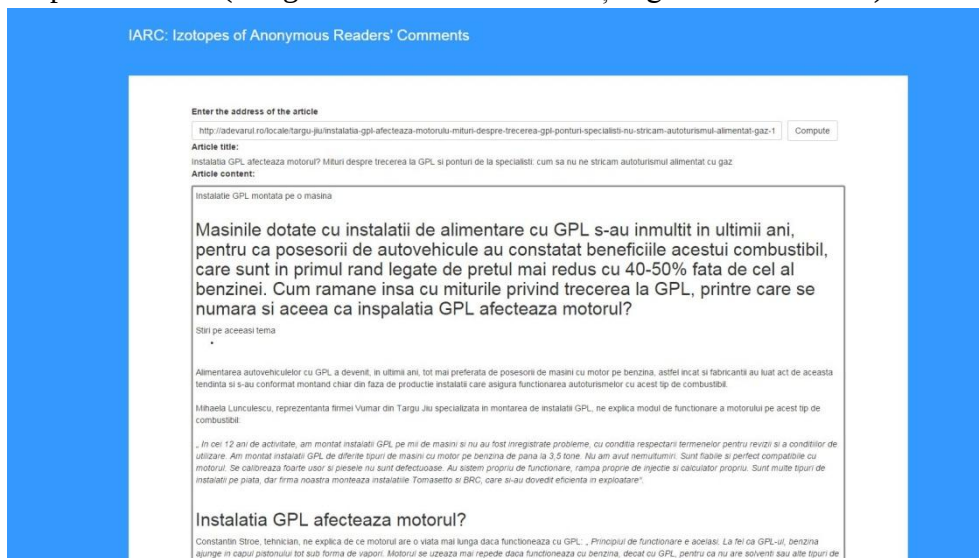


Fig. 4 IARC - sesiune de lucru (articol)

⁵ Proiect realizat de Silviu Pantilimon, student în anul II, master de Lingvistică computațională în anul 2015.



Fig. 2 Titlu articol înainte de încărcare în IARC

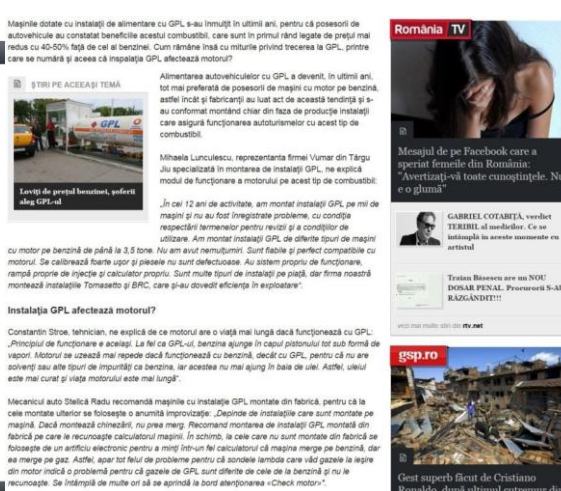


Fig. 3 Conținut articol înainte de încărcare în IARC

Spre ex, dacă pagina unui ziar online arată ca în fig. 1, fig. 2, în aplicație va fi preluat exclusiv corpul de text (v. fig. 2 - articolul comentat și fig 3 - comentariile).



Fig. 4 IARC - sesiune de lucru (articol)

Comments:

Author :
Comment category : 2

Stie cineva daca la noi exista stati de incarcare pentru autovehicule care au instalatie de C.N.G. (nu G.P.L.)? (gaz condensat nu lichefiat)? As vrea sa stiu si eu asta... e mult mai economic - cam la jumatate de pret de GPL. Grestii GNG ul este mai scump decat GPL.

Author :
Comment category : 1

As vrea sa stiu si eu asta... e mult mai economic - cam la jumatate de pret de GPL.

Author :
Comment category : 1

Grestii GNG ul este mai scump decat GPL.

Author :
Comment category : 1

Comentariu neaprobat.

Author :
Comment category : 5

"...Motorul se uzeaza mai repede daca functioneaza cu benzina, decat cu GPL..." "Desi este o ardere mai uscata la motor si asta inseamna o uzura mai rapida din cauza "da" hotaraste-le adata ori... altfel, Insa, GPL-ul este o alternativa valabila. Benzina vine de la rafinarie, integrata cu "solventi", mai precis cu aditivi (onu nu a auzit de aditivi si foloseste alti termeni). Acesti aditivi au diferite roluri: de reglare a cifrei octanice, de stabilizare la anumite conditii meteorologice (foarte frig foarte cald), biocizi (anti-bacterii si fungi), de curatare, de stabilizare a alcalinitatii etc. Unul din cei mai importanti aditivi este de lubrifiere - cine cunoaste standardul EN 228 stie ca deja din septembrie 2014 am trecut pe benzina Euro 6, la care continutul de sulf este < 10 ppm (sulf ce, cu tot raul lui mai avea si un rol de lubrifiere). NU ACELASI LUCRU SE POATE SPUNE SI DESPRE GPL. Exista si cativa aditivi pentru GPL, Insa din cauza ca GPL-ul este un gaz comprimat, efectul aditivilor este minim asupra combustibilului (aditivul nu se afla intr-o cantitate uniforma pe intreg volumul de GPL). De asemenea, nu exista aditivi pentru lubrifiere. In schimb, exista uleiuri speciale care sunt facute special pentru a combate uzura facuta de GPL. Eu am gasit un singur tip: Valvoline GPL, adus din Ungaria, in rest nimic de la Elf, Castrol, Mobil 1 etc. Probabil acestia nu se axeaza pe piata noastra. Deci, domnule mecanic auto, GPL-ul uzeaza motorul mai repede DACA NU SE FOLOSESTE UN ULEI INDEDEBII INUTAT (daca nu unul special facut pentru GPL) macar unul 100% sintetic - care costa o mica de bani. In rest ita: GPL-ul este o solutie mai rapida decat

Fig. 5 IARC - sesiune de lucru (comentarii)

Comments:

Author :
Comment category : 2

Stie cineva daca la noi exista stati de incarcare pentru autovehicule care au instalatie de C.N.G. (nu G.P.L.)? (gaz condensat nu lichefiat)? As vrea sa stiu si eu asta... e mult mai economic - cam la jumatate de pret de GPL. Grestii GNG ul este mai scump decat GPL.

Author :
Comment category : 1

As vrea sa stiu si eu asta... e mult mai economic - cam la jumatate de pret de GPL.

Author :
Comment category : 1

Grestii GNG ul este mai scump decat GPL.

Author :
Comment category : 1

Comentariu neaprobat.

Author :
Comment category : 5

"...Motorul se uzeaza mai repede daca functioneaza cu benzina, decat cu GPL..." "Desi este o ardere mai uscata la motor si asta inseamna o uzura mai rapida din cauza "da" hotaraste-le adata ori... altfel, Insa, GPL-ul este o alternativa valabila. Benzina vine de la rafinarie, integrata cu "solventi", mai precis cu aditivi (onu nu a auzit de aditivi si foloseste alti termeni). Acesti aditivi au diferite roluri: de reglare a cifrei octanice, de stabilizare la anumite conditii meteorologice (foarte frig foarte cald), biocizi (anti-bacterii si fungi), de curatare, de stabilizare a alcalinitatii etc. Unul din cei mai importanti aditivi este de lubrifiere - cine cunoaste standardul EN 228 stie ca deja din septembrie 2014 am trecut pe benzina Euro 6, la care continutul de sulf este < 10 ppm (sulf ce, cu tot raul lui mai avea si un rol de lubrifiere). NU ACELASI LUCRU SE POATE SPUNE SI DESPRE GPL. Exista si cativa aditivi pentru GPL, Insa din cauza ca GPL-ul este un gaz comprimat, efectul aditivilor este minim asupra combustibilului (aditivul nu se afla intr-o cantitate uniforma pe intreg volumul de GPL). De asemenea, nu exista aditivi pentru lubrifiere. In schimb, exista uleiuri speciale care sunt facute special pentru a combate uzura facuta de GPL. Eu am gasit un singur tip: Valvoline GPL, adus din Ungaria, in rest nimic de la Elf, Castrol, Mobil 1 etc. Probabil acestia nu se axeaza pe piata noastra. Deci, domnule mecanic auto, GPL-ul uzeaza motorul mai repede DACA NU SE FOLOSESTE UN ULEI INDEDEBII INUTAT (daca nu unul special facut pentru GPL) macar unul 100% sintetic - care costa o mica de bani. In rest ita: GPL-ul este o solutie mai rapida decat

Fig. 5 IARC - sesiune de lucru (comentarii)

Procedura de utilizare a IARC este simplă, dar foarte eficientă și rapidă. Concret, se accesează site-ul publicației care are forum. Se “citește” site-ul, însemnând articolul și comentariile aferente acestuia. Pe rând aceste texte sunt trecute prin modulul de extragere a cuvintelor-cheie (en. *topic extraction*), care constă în următoarele etape:

1. Conținutul textului se trimite spre pre-procesare cu instrumentele⁶ cunoscute în PLN: segmentarea în propoziții, etichetarea morfo-lexicală, lematizarea și extragerea grupurilor nominale de interes (NP chunking)⁷. Cu alte cuvinte, împărțirea unui text în secvențe de cuvinte corelate sintactic.

⁶<http://nlptools.infoiasi.ro/>

⁷ Noun Phrase chunking (NP-chunk) se referă la extragerea grupurilor/expresiilor substantive dintr-o propoziție.

2. Se verifică doar cuvintele din grupurile nominale, din care se rețin adjectivele și substantivele care sunt contorizate.

3. Cuvintele găsite se trec printr-un filtru de curățare (en. *cleaning*), fiind eliminate cuvintele de legătură (conjunție și prepoziție), dar și adverbele și pronumele care nu fac obiectul etapei de preprocesare NP chunking.

4. Fiecare cuvânt-cheie din lista găsită de aplicație (articolul) se verifică cu celelalte cuvinte-cheie (unu – la – unu) de la fiecare comentariu în parte. În cazul în care sunt sinonime, ponderile acestora se cumulează, fiind reținut doar cuvântul-cheie cu cea mai mare frecvență.

5. Se analizează toate propozițiile unde apare cel mai important cuvânt-cheie (cu cea mai mare frecvență) sau unde regăsim sinonime cu acesta.

6. Pentru fiecare cuvânt care nu este sinonim cu cel mai important cuvânt, dar care se găsește în aceeași propoziție cu el, sau cu sinonime ale acestuia, se adaugă +0.5 pentru fiecare potrivire (en. *math*)⁸.

7. Din lista totală, pentru fiecare cuvânt, se elimină sinonimele acestuia⁹.

8. Pentru articole care sunt comentate, eliminăm cuvintele cu ponderea ≤ 2 , fiind considerate ca de importanță minoră, numărul crescut al acestora putând afecta rezultatul final (compararea izotopiilor din articol cu cele din comentarii).

Din cuvintele rămase (*totalSize*) se păstrează doar cuvintele $\text{validWords} = 5 + (\text{totalSize} / (3 + \text{totalSize} / 10))$.

9. În vederea *categorizării comentatorilor*, calculele s-au făcut după unele teste, care pot fi îmbunătățite pe măsură ce corpul de antrenare crește.

Exemplu: dacă într-un articol se vorbește despre armamentul Rusiei (tipuri de tancuri, avioane etc.), iar în unele comentarii se vorbește despre planurile Rusiei și pe cine vrea să atace, aceste comentarii vor avea o categorie mică, deoarece nu se vorbește despre armamentul Rusiei, care este recunoscut ca grup nominal.

Acesta a fost și motivul folosirii celor trei tipuri de calcule folosite la clasificarea comentatorilor, pe care le descriem mai jos:

a) Pentru fiecare cuvânt din comentariu (considerat ca fiind izotopic) se verifică în lista de cuvinte izotopice ale articolului. În cazul în care unul dintre cuvintele-cheie reținute în lista analizei articolului sau este sinonim cu acesta -> se adaugă la rezultatul final.

$\text{pondere cuvânt articol} * \text{pondere cuvânt comentariu}$

ponderea 100% este considerată ca fiind suma pătratelor ponderilor articolului.

Categoriile comentatorilor (*none*, *low*, *medium*, *high* și *expert*) se vor încadra în următoarele intervale:

1 ->[0-4] -> none (neglijabil)

2 ->(4-10] -> low (slab)

3 ->(10-14] -> medium (mediu)

4 ->(14-20] -> high (ridicat)

5 ->(20-100] -> expert (expert)

b) Se verifică numărul de cuvinte care apar în ambele liste, atunci ponderea 100% va fi

⁸Notă: un cuvânt poate avea ponderea maximă egală cu ponderea celui mai important cuvânt.

⁹Notă: din lista de sinonime se păstrează cuvântul cu ponderea cea mai mare.

numărul total de cuvinte din articol.

Exemplu: într-un comentariu identificăm cuvintele-cheie x,y,z , iar la articol x,y,z,t .

Rezultatul va fi: $3 \cdot 100 / 4 = 75\%$

Cele cinci categoriile amintite se vor încadra în următoarele intervale:

- 1 ->[0-20] -> none (neglijabil)
- 2 ->(20-30] -> low (slab)
- 3 ->(30-50] -> medium (mediu)
- 4 ->(50-75] -> high (ridicat)
- 5 ->(75-100] -> expert (expert)

c) Pentru fiecare cuvânt-cheie care apare în ambele liste (articol și comentariu), se adaugă ponderea din articol, pondere 100% fiind dată de suma tuturor ponderilor din articol.

Exemplu:

Dacă în articol avem ponderile: $x = 5\%$ $y = 6\%$ $z = 10\%$ $t = 3\%$

iar în comentariu găsim doar cuvintele-cheie y, z, t, ponderea acestuia va fi dată de formula:

$$(6+10+3) \cdot 100 / (5+6+10+3) = 86\%$$

În acest caz, categoriile stabilite se vor încadra în următoarele intervale:

- 1 ->[0-20] -> none (neglijabil)
- 2 ->(20-30] -> low (slab)
- 3 ->(30-50] -> medium (mediu)
- 4 ->(50-75] -> high (ridicat)
- 5 ->(75-100] -> expert (expert)

Rezultatul final va fi media aritmetică dintre cele 3 calcule. Menționăm, că aceste calcule au rezultat în urma unor teste (adnotare manuală vs. adnotare automată) care dezavantajau comentariile scurte și foarte scurte, așa cum sunt multe pe forumurile publicațiilor online.

Deși statisticile în acest moment sunt mulțumitoare, este prematur să avansăm niște concluzii ferme cu privire la acuratețea datelor obținute.

5. Concluzii și direcții viitoare de cercetare

Lucrarea prezintă o metodologie capabilă să depisteze și să compare izotopiile (cuvintele-cheie) extrase din două tipuri de texte (articol vs. comentariu). Considerăm că instrumentul computațional prezentat aici (încă în fază experimentală) răspunde nevoii de a interpreta cantitativ aprecierile forumiștilor și de a le ierarhiza, operații care validează ipoteza cercetării.

O asemenea aplicație poate fi extrem de utilă jurnaliștilor, managerilor, specialiștilor în relații publice, politicienilor și staff-urilor de campanie, dar și specialiștilor în domeniul PLN, dornici să simplifice volumul uriaș de muncă pentru diferitele tipuri de analiză textuală.

Rămân încă probleme de rezolvat. Ne gândim la comentariile care apar după un timp de la demersul analitic propus (nu pot fi interceptate automat și trebuie revenit și verificat ce informații s-au mai adăugat). De asemenea, se poate întâmpla ca jurnalistul să mai facă adăugiri la articolul postat. Cel puțin două soluții sunt luate în calcul: adăugarea de câmpuri, unde să se completeze conținutul articolului/comentariilor sau apăsarea pe butonul dreapta al mouse-ului și salvarea cu Save as (cum este, spre exemplu, pentru Google Chrome). La a

doua variantă, se generează un fișier HTML care conține toate datele necesare, după care acesta poate fi încărcat în aplicație.

Mai mult, ne gândim și la o analiză comparativă a comentariilor și, implicit, la o categorizare a lor din perspectiva importanței subiectului opinat în comunitatea forumiștilor. Poate fi o modalitate de a descoperi acele comentarii centrate pe un subiect important, dar diferit de cel al articolului.

Deloc de neglijat este și clasificarea conținutului articolului raportat la titlul acestuia. Cu alte cuvinte, să analizăm dacă titlul reflectă textul articolului, dat fiind că presa românească abundă de titluri care seduc, dar care în corpul articolului se regăsesc prea puțin.

Mulțumiri: Această lucrare a fost publicată cu sprijinul Proiectului ERASMUS MUNDUS (Erasmus Mundus European Mobility with Neighbouring Region in the East), Action 2 – Strand 1 (2009-2013), Acord de finanțare nr. 2011-2576/001-001-EMA2, (Lot 8: Moldova, Ukraine, Belarus), finanțat de Uniunea Europeană.

REFERINȚE BIBLIOGRAFICE:

Barthes, R.: *S/Z*, transl. by Richard Miller, New York: Hill & Wang, 1970.

Ciotti, F. & Crupi, G.: *Dall'Informatica umanistica alle culture digitali. Atti del Convegno di studi in memoria di Giuseppe Gigliozzi* (Roma, 27-28 ottobre 2011), Roma, Università La Sapienza, 2012

Chiorean, L.: *Discurs eseistic. Coerență, referențialitate, coeziune*, în *Lucrările Conferinței Internaționale „Integrarea europeană – între tradiție și modernitate”*, ediția a II-a, Târgu-Mureș, 20-21 septembrie 2007.

Cummings, J.: *The Text Encoding Initiative and the Study of Literature*, in Ray Siemens and Susan Schreibman (eds.), *Blackwell Companion to Digital Literary Studies* (Blackwell: Malden), 451-476, 2007.

Derrida, J.: *Of grammatology*, transl. by G.C. Spivak, Baltimore: John Hopkins University Press, 1974.

Everaert-Desmedt, N.: *Sémiotique du récit*, De Boeck Université, Bruxelles, colecția „Culture&Communication”, seria „Licence, Master, Doctorat”, 2007.

Gîfu, D. and Cioca, M.: *Online Civic Identity. Extraction of Features* in *Procedia – Social and Behavioral Sciences*, vol. 76/15, edited By Emanuel Soare, ELSEVIER, 2013, pp. 366-371.

Gîfu, D., Stoica, D. and Cristea, D.: *Virtual Civic Identity* in *Proceedings of The 9th International Conference Linguistic Resources and Tools for Processing The Romanian Language, ConsILR-2013*, 16-17 May 2013, Miclăușeni, Elena Mitocariu, Mihai Alex Moruz, Dan Cristea, Dan Tufiş, Marius Clim (eds.), "Alexandru Ioan Cuza" University Publishing House, Iași, 2013, pp. 139-148

Greimas, A. J.: *Structural Semantics: An Attempt at a Method*, 1966, transl. by Daniele McDowell, Ronald Schleifer, Alan Velie, Lincoln, Nebraska: University of Nebraska Press, 1983.

Greimas, A. J.: *Del senso*, Milano, Bompiani, 1970.

Greimas, A. J.: *Despre sens. Eseuri semiotice*, București, Ed. Univers, 1975.

Greimas, A. J. and Courtés, J.: *Semiotics and Language: An Analytical Dictionary*,

Bloomington: Indiana University Press, 1982.

Moeschler, J., Reboul, A.: *Dicționar enciclopedic de pragmatică*, coord. și trad. de Carmen Vlad și Liana Pop, Cluj-Napoca, Ed. Echinoc, 1999.

Mordenti, R.: *L'altra critica. La nuova critica della letteratura fra studi culturali, didattica e informatica*, Roma, Meltemi, 2007.

Plett, H. F.: *Știința textului și analiza de text*, Ed. Univers, București, 1983.

Rastier, F.: *Systématique des isotopies*, în A.J. Greimas (ed.), *Essais de sémiotique poétique*. Paris, Larousse, 1972.

Ricœur, P.: *De la text la acțiune. Eseuri de hermeneutică*, vol. 1, Ed. Humanitas, București, 1995.

Riva, M.: *Il futuro della letteratura. L'opera letteraria nell'epoca della sua (ri)producibilità digitale*, Scriptaweb, 2011.

Romary, L.: *Questions & Answers for TEI Newcomers*, in: *Jahrbuch für Computer philologie*. (see, <http://computerphilologie.de/jg08/romary.pdf>, accesat pe 7 mai 2015, 2009.

Roventă-Frumușani, D.: *Semiotică, societate, cultură*, Institutul European, Iași, 1999.

Vanhoutte, E. și Van den Branden, R.: *The Text Encoding Initiative*, in: Marcia J. Bates and Mary Niles Maack (eds.), *Encyclopedia of Library and Information Sciences* (1), 2010.

Vlad, C.: *Textul aisberg*, Casa Cărții de Știință, Cluj-Napoca, 2000.