



HAL
open science

Analiza automată a auto-explicațiilor

Bogdan Oprescu, Mihai Dascălu, Ștefan Trăușan-Matu, Traian Rebedea,
Philippe Dessus, Maryse Bianco

► **To cite this version:**

Bogdan Oprescu, Mihai Dascălu, Ștefan Trăușan-Matu, Traian Rebedea, Philippe Dessus, et al..
Analiza automată a auto-explicațiilor. *Revista Română de Interacțiune Om-Calculator*, 2012, 5 (2),
pp.71-76. hal-01075551

HAL Id: hal-01075551

<https://hal.science/hal-01075551v1>

Submitted on 18 Oct 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Analiza automată a auto-explicațiilor

Bogdan Oprescu, Mihai Dascălu, Ștefan Trăușan-Matu,
Traian Rebedea

Universitatea Politehnica din București

313 Splaiul Independenței, 060042 București, România

bogdan.oprescu@cti.pub.ro, mihai.dascalu@cs.pub.ro,
stefan.trausan@cs.pub.ro, traian.rebedea@cs.pub.ro

Philippe Dessus, Maryse Bianco

Université Pierre-Mendès France

F-38040 Grenoble CEDEX 9,
Franța

philippe.dessus@upmf-
grenoble.fr,
maryse.bianco@upmf-grenoble.fr

REZUMAT

Auto-explicațiile reprezintă verbalizări pe care un cititor și le formulează în timpul lecturii unui text în vederea înțelegerii mai eficiente a conținutului. Sistemul implementat este proiectat să analizeze în mod automat aceste explicații, permițându-i astfel profesorului să evalueze mai în detaliu nivelul de înțelegere al materialelor citite; implementări similare există doar pentru limba engleză. Lucrarea de față prezintă pe scurt arhitectura aplicației și tehnologiile folosite în implementare și descrie modul în care s-a realizat evaluarea verbalizărilor. Metoda propusă se bazează pe tehnici specifice de prelucrare a limbajului natural adaptate pentru limba franceză și se adresează utilizării în clasele din școala primară. În plus, în cadrul procesului de analiză am integrat o euristică proprie la nivel de concepte pentru a putea evalua similaritatea dintre textele inițiale și verbalizările elevilor.

Cuvinte cheie

Verbalizare, exercițiu de lectură prin auto-explicare (SERT), analiză semantică latentă (LSA), analiză automată a auto-explicațiilor.

Clasificare ACM

I.2.7 Natural Language Processing.

INTRODUCERE

Studii psihologice și pedagogice au arătat că oamenii tind să înțeleagă mai bine un text atunci când încearcă să își explice ceea ce au citit și asimilat pe parcursul lecturii [1], [2]. Pornind de la aceste observații, au fost dezvoltate tehnici, precum SERT [3], care să ajute elevii să înțeleagă mai bine textele științifice în vederea eficientizării procesului de învățare prin focalizarea acestuia pe înțelegere, mai degrabă decât pe memorare.

Experimentele educaționale desfășurate de noi decurg în următorul fel: la momente predefinite, elevii sunt opriți în timpul lecturii și li se cere să explice ceea ce au citit până la momentul respectiv. Explicațiile lor sunt înregistrate și apoi transcrise, evaluate de doi experți umani și clasificate conform unei scheme folosite de McNamara în aplicații similare pentru limba engleză [4]. În sistemele mai avansate de asistență la învățare, elevilor li se prezintă mai multe metode de verbalizare și sunt încurajați să le folosească în mod alternativ. Așadar, evaluarea explicațiilor date de elevi este un pas cheie în a-i ajuta să își îmbunătățească înțelegerea în timpul citirii unui text. Criteriul nostru de evaluare este reprezentat de

cunoștințele folosite de cititor pentru a formula propriile explicații. În consecință, o verbalizare poate fi:

Parafrazare – o reformulare a ultimului paragraf citit folosind alte cuvinte. Parafrazarea textelor obligă elevii să reformuleze textele într-o manieră cât mai familiară. De asemenea, îi forțează să facă o reprezentare mintală a cunoștințelor din text și să înțeleagă structura generală a contextului; aceștia pot fi considerați primii pași în procesul de înțelegere a textului dat.

Predicție – o explicație care anticipează parțial o parte din informația care urmează să apară în text.

Cauzal-relevantă – o frază apropiată oarecum de o alta cauzal relevantă din ultimul paragraf.

Verbalizare bazată pe cunoștințe anterioare – o explicație în care cititorul folosește informații anterioare împreună cu informațiile găsite în text.

Corelație – tip de explicație în care cititorul leagă fragmente de informație din ultimul paragraf sau din paragrafe mai vechi din text care îl ajută să înțeleagă cum diferite alte părți din text sunt legate și să își producă o imagine globală asupra întregului material citit.

Dacă dorim ca elevii să fie mereu asistați vom avea nevoie de un pedagog specializat care să supravegheze un număr mic de elevi, ceea ce face aceste tehnici greu de aplicat la o scară largă. În plus, evaluarea conținutului unei verbalizări este o activitate subiectivă care poate fi asistată de tehnici computaționale. Din acest motiv a apărut ideea folosirii unei aplicații pe calculator, care să asiste activitatea unui cadru didactic.

Primele experimente au fost publicate de McNamara et al. [4], iar iSTART poate fi considerat prima implementare care tratează auto-explicațiile [10]. Acesta funcționează exclusiv pentru limba engleză și conține mai multe module: unul care explică sistemul SERT studenților, altul care face o demonstrație de utilizare a auto-explicațiilor folosind un student și un tutore virtual și un al treilea care se ocupă de antrenarea utilizatorului propunând texte spre lectură, cerând explicații și oferind ajutor în formularea lor. Principala provocare ridicată de un astfel de sistem este evaluarea verbalizărilor oferite de studenți în concordanță cu materialele citite, aceasta fiind componenta pe care încercăm să o implementeze aplicația noastră, de aceasta dată pentru limba franceză.

Astfel, scopul proiectului nostru a fost de a face posibilă integrarea de noi texte fără intervenția vreunui specialist, iar prelucrarea lor manuală să fie redusă sau chiar absentă.

iSTART împarte verbalizările în patru categorii principale: *irelevante, parafraze, verbalizări* conținând *cunoștințe prezente în alte zone ale textului* decât în pasajul tocmai citit, și *verbalizări* care folosesc *cunoștințe anterioare* ale elevului, neîntâlnite în text. După cum rezultă și din datele prezentate în [5], este mai ușor să se identifice parafraze sau explicații irelevante, însă este mult mai dificil de identificat verbalizările care fac parte din celelalte două categorii.

Scopul nostru a fost să creăm un sistem similar cu iSTART pentru limba franceză care să poată fi folosit pentru texte cu o dificultate corespunzând cunoștințelor elevilor din clasele primare. În acest context, am conceput un modul de evaluare care utilizează Analiza Semantică Latentă (LSA) și o euristică bazată pe vocabularul folosit, ca tehnici de prelucrare a limbajului natural.

Metoda folosită pentru a include automat verbalizările într-o categorie a fost să comparăm explicația dată de utilizator cu paragraful citit ultima dată, cu precedentul și cu cel care îl succede, așa cum se observă în Figura 1:

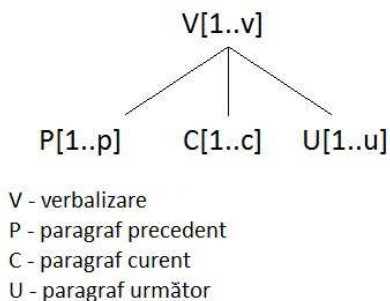


Figura 1. Tehnica de comparare

Modul în care o verbalizare este apoi inclusă în una dintre categorii este detaliat în Tabelul 1. În continuare această lucrare se va focaliza pe a stabili ce înseamnă „apropiat” și „oarecum apropiat” în termeni de prelucrare de limbaj natural și pe detecția verbalizărilor în care elementele de parafrază predomină.

Următoarele secțiuni descriu arhitectura aplicației și rolul fiecărui modul în încercarea noastră de a determina natura verbalizărilor furnizate de utilizatori. Ulterior prezentăm experimentele realizate și deciziile pe care le-am luat pe baza rezultatelor obținute prin măsurarea similarităților și prin detecția erorilor.

ARHITECTURA

Aplicația e compusă din mai multe module (Figura 2), unele fiind folosite în interacțiunea directă cu utilizatorul. În cadrul acestei lucrări ne vom concentra în special asupra prezentării modulelor care se ocupă de evaluarea auto-explicațiilor.

Fluxul de date

La pornirea aplicației, un fișier de configurare este parsat, iar drept rezultat un graf de stări este construit pentru a dicta comportamentul aplicației. Atunci când este cerută o verbalizare, textul este corectat pe măsură ce utilizatorul scrie, folosindu-se modulul Jazzy, pentru a elimina greșelile apărute la dactilografie, ca de exemplu caractere lipsă sau caractere interschimbate. Apoi, textul

corectat este trimis către modulul „Stare”, care cere modulului „Test” să evalueze verbalizarea. În funcție de răspunsul primit de la modulul „Test”, modulul „Stare” poate decide să ceară o nouă explicație sau să treacă la următorul paragraf.

Tabelul 1. Logica de decizie

Verbalizare	Criteriu de decizie
Parafrază	V apropiat de C
Predicție	V oarecum apropiat de U
Cauzal-relevantă	V apropiat de o propoziție causal-relevantă (selectată manual) din C
Cunoștințe externe	V aproioat de un rezumat al textului
Corelare	P, U, C, V apropiate între ele

Modulul „Test” primește verbalizarea și o compară cu paragraful curent, cu cel precedent și cu următorul paragraf. Funcția de determinare a similarității e bazată pe Analiza Semantică Latentă (LSA) și pe o listă de cuvinte relevante. Funcția obține un rating de la cele două metode de comparare și calculează un rating unic. Modul de combinare și de folosire în determinarea tipului verbalizării au fost determinate experimental, iar pragurile identificate au fost introduse în logica aplicației. Modul efectiv în care au fost stabilite pragurile va fi detaliat în secțiunile următoare.

Pentru a realiza comparația folosind LSA, modulul ”Test” trece mai întâi informația prin modulul ”Converter”, care elimină punctuația, elimină terminațiile (dacă LSA a fost antrenat pe un corpus similar) și înlocuiește diacriticele specifice limbii franceze. La sfârșit, modulul LSA calculează un vector al paragrafului și returnează cosinusul dintre vectorii celor două paragrafe comparate. Corpusul de antrenare folosit conține mai multe texte pentru copii. Dimensiunea totală a acestuia a fost de 6Mb de text folosit pentru crearea spațiului semantic. Corpusul de antrenare a fost segmentat și punctuația a fost eliminată, iar diacriticele limbii franceze au fost înlocuite. Doar segmentele între cincizeci și o sută de cuvinte au fost reținute pentru antrenare.

În ceea ce privește euristica bazată pe cuvinte importante, modulul „Test” construiește o listă de cuvintele din text separate pe părți de vorbire: substantiv, verb, adjectiv, adverb și din sinonimele lor. Pentru determinarea părții de vorbire și a formei de bază a cuvântului se folosește Tree Tagger [8]. Fiecare cuvânt este căutat apoi în WOLF [6], o variantă de WordNet open-source pentru limba franceză, și sinonimele fiecărui cuvânt sunt identificate în lista de cuvinte importante. Ulterior sunt numărate cuvintele din verbalizare care se regăsesc în listele de cuvinte relevante. Se calculează astfel un raport între cuvintele regăsite și numărul total de cuvinte din listă și se obține un rating unic, făcându-se o medie ponderată între rapoartele corespunzătoare celor patru părți de vorbire.

Odată scorul final calculat, acesta este transmis modulului „Stare” care returnează feedback personalizat utilizatorului.

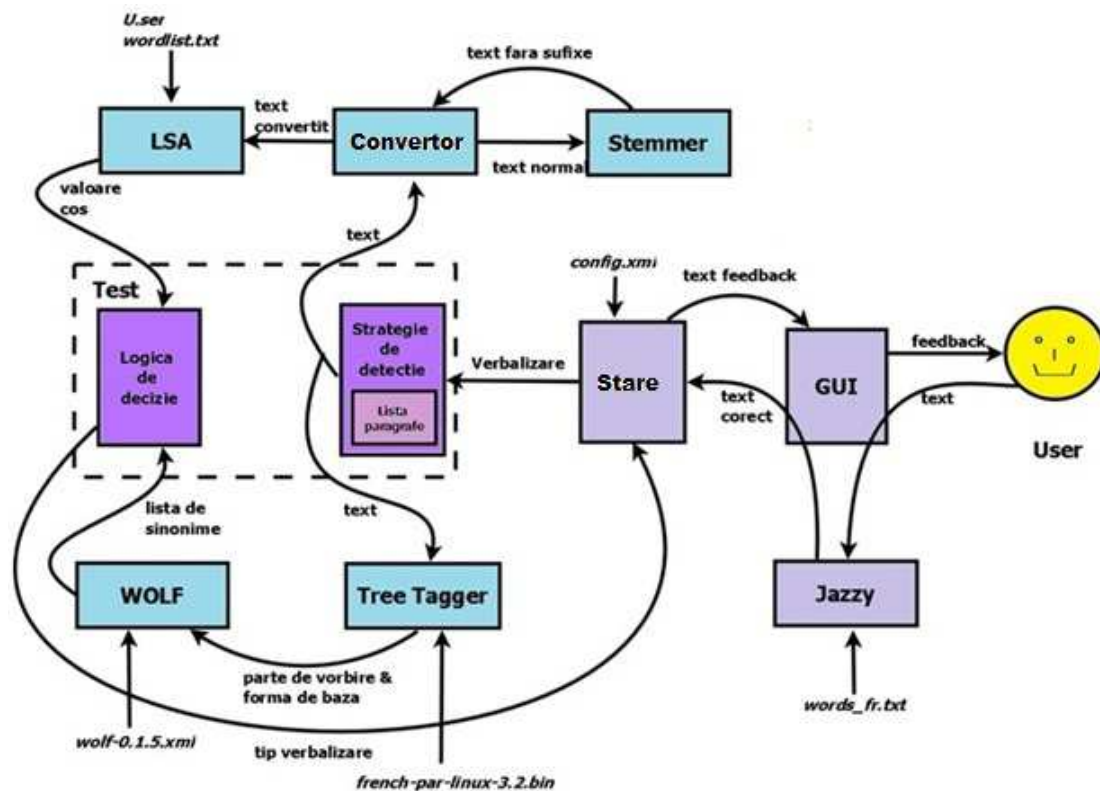


Figura 2. Fluxul de date

TEHNOLOGII FOLOSITE ȘI ABORDĂRI

Spellchecking cu Jazzy

Modulul Jazzy este responsabil de corectarea greșelilor de ortografie din fereastra de input [9]. Modulul folosește un dicționar și încearcă să aproximeze forma corectă a cuvântului folosind distanța Levenshtein și o listă cu toate cuvintele din limba franceză. Lista de cuvinte este obținută prin parsarea unui dicționar de forme flexionale al limbii franceze, Morphalou¹, disponibil online.

Converter

Acest modul este folosit pentru a pregăti textul pentru analiza cu LSA. Primul pas în conversie are în vedere eliminarea punctuației și a altor elemente în afară de formele cuvintelor. Dacă folosim un LSA antrenat pe un corpus pe care s-a efectuat *stemming*, atunci se va efectua *stemming* și pe textul introdus de utilizator.

Deoarece diacriticele din limba franceză au fost înlocuite în corpusul de antrenare al LSA, acestea sunt înlocuite și în textele analizate în prezent.

LSA

Analiza semantică latentă (LSA) este o teorie și metodă de extragere și reprezentare a sensului cuvintelor. Sensul este estimat folosind calcule statistice aplicate pe corpusuri mari de text. Un corpus lingvistic reprezintă un set de constrângeri pe care LSA le extrage pentru a determina sensul cuvintelor prin intermediul conceptelor [5]. Ca aparat matematic, LSA folosește metode de algebră

liniară, principalul procedeu fiind descompunerea în valori proprii. Întrucât măsoară similaritatea dintre cuvinte, LSA se comportă mai bine pe corpusuri de text din zone specializate ale limbajului, cum ar fi vocabularul științific din diferite domenii. Pentru a funcționa eficient, e important ca textele pentru care metoda e apelată să fie din aceeași zonă de vocabular cu corpusul de antrenare.

Principala calitate a LSA este capacitatea de a exploata constrângerile mutuale. Astfel, înțelesul unui paragraf poate fi calculat pe baza sensului cuvintelor din care acesta este compus. LSA tratează corpusul ca pe un număr de paragrafe individuale care au un înțeles coerent, le convertește pe fiecare într-o ecuație în care fiecare cuvânt este o variabilă iar numărul său de apariții coeficientul corespunzător, iar prin rezolvarea sistemului este calculată valoarea fiecărui cuvânt. De aceea, prin LSA este posibil să se calculeze valoarea unui paragraf prin însumarea valorilor fiecăruia din cuvinte. În această abordare, înțelesul fiecărui cuvânt depinde de cuvintele care fac parte din acel paragraf. Din această cauză, pentru a aplica LSA eficient, corpusul de antrenare trebuie să fie suficient de mare, comparabil ca dimensiune cu volumul de text de care are nevoie un om pentru a învăța să vorbească. Din cauza dimensiunii mari a sistemului de ecuații, calculele necesare pentru antrenare consumă mult timp și resurse, chiar și pentru sistemele puternice de calcul folosite astăzi. După ce se realizează descompunerea în valori proprii, spațiul vectorial rezultat este proiectat pe aproximativ 300 de dimensiuni, obținându-se un spațiu semantic de vectori care urmează să fie folosit în calcularea valorii fiecărui paragraf.

Vectorul unui paragraf este estimat ca suma componentelor sale, iar similaritatea între paragrafe este

¹ <http://cnrtl.fr/lexiques/morphalou/>

măsurată ca valoarea cosinusului dintre vectorii celor două paragrafe. S-a mai introdus și un parametru menit să amortizeze efectul cuvintelor care apar prea des și într-un număr mare de paragrafe, Tf-IDf (Term frequency, Inverse Document frequency) [5], pentru a mări acuratețea măsurătorilor. Formula de calcul a componentei vectorului unui paragraf, pe dimensiunea i , este următoarea:

$$p_i = \sum_{i=0}^k x_i \frac{1}{f_i} (1 + \log n)$$

unde x_i este valoarea componentei vectorului cuvântului x pe dimensiunea i , n este numărul de apariții al cuvântului în paragraf, f_i este frecvența cuvântului în corpusul de antrenare și k este dimensiunea spațiului de vectori, care în cazul nostru are valoarea 300.

Cosinusul unghiului dintre doi vectori este calculat cu formula:

$$\cos(\vec{x}, \vec{y}) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

Stemmer

În prelucrarea limbajului natural, operația de *stemming* presupune extragerea rădăcinii comune din forma flexională a unui cuvânt. În acest scop trebuie eliminate sufixele și prefixele cuvintelor astfel încât cuvinte din aceeași familie să se reducă la aceeași formă, nefiind obligatoriu ca această să fie o formă de dicționar. Implementarea noastră integrează Snowball [7], un stemmer open-source bazat pe reguli.

Tree Tagger

Tree Tagger [8] este un instrument care etichetează cuvintele în funcție de categoria morfologică din care fac parte și ajută la identificarea cuvintelor din principalele patru categorii recunoscute de WordNet (substantiv, verb, adjectiv și adverb). O altă funcție foarte importantă îndeplinită de Tree Tagger este aceea de identificare a rădăcinii cuvintelor, astfel încât ele pot fi mai ușor căutate în dicționar. Principalul avantaj al acestei implementări este că aceasta poate funcționa independent de limbă, necesitând doar un fișier specific de configurare, particularizat pentru fiecare limbă în parte.

WOLF

WordNet este o ontologie lexicalizată pentru limba engleză [6]. Acesta grupează cuvintele în seturi de sinonime, numite synset-uri, oferă definiții scurte și generale pentru ele și ține cont de relațiile semantice între synset-uri. Scopul final este de a produce o combinație de dicționar și de tezaur care să fie intuitivă și care să suporte analize automate. Este important de specificat diferența dintre WordNet și un tezaur. Un tezaur este o mulțime de cuvinte grupate împreună în funcție de sens. WordNet realizează suplimentar și legătura între concepte, devenind astfel un instrument important pentru prelucrarea limbajului natural.

Întrucât scopul programului nostru a fost să prelucrăm limba franceză, a trebuit să găsim o alternativă la WordNet care să funcționeze pentru limbă franceză.

Singura bază de date similară, open-source pentru franceză este WOLF (*WordNet Libre du Français*), un proiect dezvoltat de Benoît Sagot la universitatea Paris 7 în Paris. Această versiune conține aproximativ treizeci de mii de synset-uri, cam o treime din câte conține versiunea pentru engleză, iar câmpurile care desemnează sensul sunt completate cu informații cu privire la sursele din care fost preluat cuvântul, și nu cu numărul aferent sensului. Baza de date este păstrată într-un format XML și respectă sintaxa folosită de BalkaNet [11], un proiect similar dezvoltat pentru mai multe limbi est-europene. În mod evident WOLF nu este la fel de performant ca varianta pentru engleză a WordNet, dar este instrumentul cel mai potrivit pentru scopurile noastre.

Măsurarea similarității

Modulul de măsurare a similarității ocupă rolul central al aplicației, conectând toate celelalte module între ele și realizând operațiile cele mai importante. Acesta primește intrarea de la modulul „Stare” și de la fișierele de configurare, apelează celelalte module pentru a evalua verbalizările și trimite răspunsul final către modulul „Stare”.

Pentru abordarea bazată pe cuvinte importante au fost folosite Tree Tagger și WOLF cu scopul de a crea liste de cuvinte relevante pentru fiecare paragraf. Când un paragraf este creat, cuvintele din componența lui sunt etichetate și apoi este creată o listă conținând toate sinonimele pentru substantivele, verbele, adverbele și adjectivele din text. Toate aceste cuvinte sunt considerate ca fiind cuvintele relevante din text.

Mai târziu cuvintele din verbalizare sunt și ele etichetate și apoi sunt numărate cuvintele din fiecare categorie. Se calculează raportul dintre cuvintele aflate în verbalizare care se regăsesc în lista de cuvinte importante ale paragrafului și numărul total de cuvinte din verbalizare, pentru fiecare din cele patru părți de vorbire luate în considerare. Pentru calcularea unui grad unic de asemănare se folosește formula:

$$R_W = \frac{W_s \frac{n_s}{N_s} + W_v \frac{n_v}{N_v} + W_{aj} \frac{n_{aj}}{N_{aj}} + W_{av} \frac{n_{av}}{N_{av}}}{W_s + W_v + W_{aj} + W_{av}}$$

unde R_W este valoarea returnată de funcție, n_s , n_v , n_{aj} și n_{av} sunt numerele de substantive, verbe, adjective și respectiv adverbe din verbalizare prezente în lista de cuvinte relevante a paragrafului, N_s , N_v , N_{aj} și N_{av} reprezintă numerele totale de cuvinte din cele patru părți de vorbire din cele patru paragrafe, iar W_s , W_v , W_{aj} și W_{av} sunt ponderile lor în calcularea mediei. Ponderile au fost determinate experimental, încercându-se multiple combinații pentru diferite valori ale ponderilor.

O funcție separată este folosită pentru a calcula similaritatea cu ajutorul LSA. Aceasta compară fiecare propoziție din verbalizare cu întreg paragraful și valorile obținute sunt păstrate într-o listă. Apoi sunt eliminate cele mai mici două valori și se returnează o medie a valorilor rămase, ponderată cu lungimea propozițiilor. În final, întreaga verbalizare este comparată cu paragraful în cauză și valoarea obținută este introdusă în calcul, cu o pondere egală cu jumătate din lungimea ei.

După calculul acestor doi parametri, modulul "Test" poate lua o decizie cu privire la paragraful analizat. Au fost calculate experimental un prag minim și un prag maxim, iar în urma observațiilor făcute pe multiple categorii de texte, s-a ajuns la concluzia că cea mai bună variantă ar fi setarea manuală a pragurilor la valori convenabile determinate experimental. Dacă scorul paragrafului este scăzut pentru ambele criterii, atunci explicația nu poate fi luată în considerare. Dacă scorurile depășesc pragurile superioare, atunci aceasta este considerată o parafrază. Altfel, explicația este considerată drept având legătură cu paragraful, dar nu suficient de apropiată pentru a fi considerată o parafrază.

REZULTATE

Am efectuat mai multe teste folosind cele două metrici (similaritate LSA și co-apariții ale cuvintelor) și am reușit să formulăm mai multe concluzii pe baza acestor rezultate. Corpusul nostru de test a constat într-un text împărțit în șase paragrafe, de aproximativ cinci propoziții fiecare, și din verbalizările date de cinci elevi de clasele primare pentru paragrafele respective. Participanții la experiment au fost rugați să se oprească din lectură după fiecare secțiune și să explice ce au citit până în momentul respectiv. Verbalizările au fost apoi evaluate manual și au fost identificate elementele de parafrază, corelare, elaborare sau predicție din cadrul fiecăreia. Am folosit ambele metrici pentru a compara verbalizările date de elevi cu paragraful citit înainte, cu paragraful care îl precede și cu următorul paragraf din text și am încercat să determinăm natura verbalizărilor bazându-ne pe rezultatele acestor comparații.

Primul aspect care ne-a interesat presupune măsurarea distribuției valorilor returnate de funcțiile de evaluare pe codomeniul acestora, respectiv intervalul $[0, 1]$, pentru a verifica dacă distribuția acestor valori este uniformă. În consecință am ales în mod arbitrar rezultatele returnate de cele două metrici pentru o serie de verbalizări din corpusul nostru de test, am sortat crescător rezultatele și le-am reprezentat în graficele de mai jos (Figurile 4 și 5).

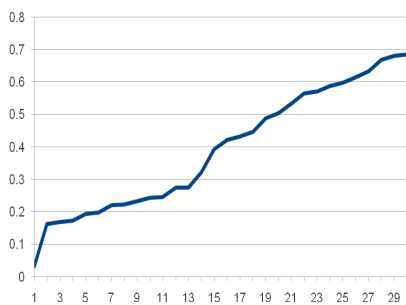


Figura 3. Distribuția valorilor pentru metrica bazată pe vocabular

Se observă că LSA variază într-un interval mai mic, între 0 și 0.5, în timp ce euristica bazată pe co-apariții ale cuvintelor returnează valori între 0 și 0.7; în orice caz, ambele evoluează aproximativ liniar. Această analiză ne ajută să stabilim un prag peste care putem considera o verbalizare ca fiind o parafrază.

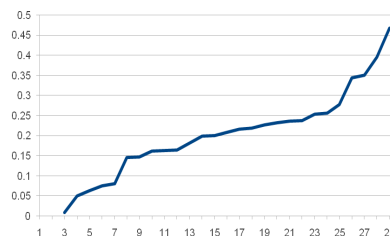


Figura 4. Distribuția valorilor pentru metrica LSA

În acest moment avem două metrici, ambele indicând gradul de asemănare între două paragrafe, dar trebuie să decidem dacă rezultatele acestor două metrici sunt sau nu coerente. În consecință, am încercat să evaluăm gradul în care rezultatele returnate de acestea sunt coerente. Figura 5 prezintă rezultatele comparative ale celor două metrici atunci când sunt furnizate aceleași intrări.

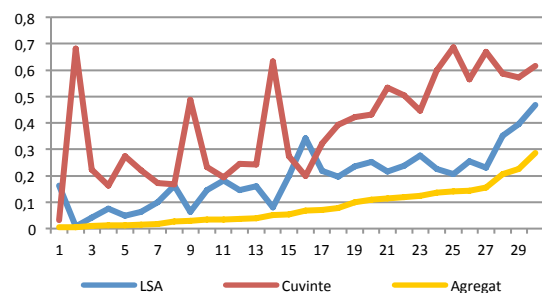


Figura 5. Comparație între cele două metrici folosite

Pe baza acestor observații, am decis că cea mai bună metodă de a combina aceste două metrici este să le înmulțim. Rezultatul metricii combinate este, de asemenea, prezentată pe grafic, cu culoarea galben.

Corelația Pearson pentru cele două metrici, calculată pentru datele pe care a fost testată este de 34.02%, iar față de valoarea agregată, LSA are o corelație de 87.73%, în timp ce euristica bazată pe vocabular are o corelație de 68.16%, ceea ce înseamnă că prima euristica are o pondere mai mare în rating-ul agregat.

Pe baza acestor rezultate am decis să stabilim un prag superior de 0.07 pentru metrica combinată pentru a decide dacă o verbalizare este parafrază sau nu. Acest prag ne-a permis să identificăm 19 din 27 de parafraze în testele efectuate, ceea ce înseamnă o precizie de aproximativ 70%.

Astfel, în acest moment știm că putem identifica parafrazele cu o precizie destul de bună, dar trebuie să evaluăm dacă datele obținute ne pot ajuta și în privința altor tipuri de verbalizări. În consecință am comparat valoarea returnată de metrica pentru paragraful curent, pentru cel precedent și pentru cel care îl succede, pentru a găsi similități între verbalizările de același tip. Am reprezentat separat variațiile date de cele două metrici pentru verbalizările în care predomină elementele de parafrază.

Figura 6 prezintă valorile returnate de LSA pentru 11 parafraze comparate cu paragraful tocmai citit, cu cel precedent și cu următorul. Este evident că asemănarea cu paragraful curent este mult mai mare decât cu cele învecinate, unde valoarea returnată de funcția de comparație se apropie de zero.

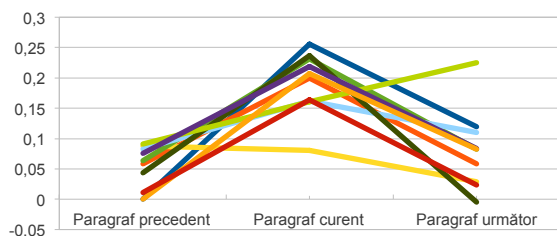


Figura 6. Verbalizări conținând parafraze comparate cu LSA

Figura 7 arată același test pentru metrica bazată pe vocabular. Se observă că graficul are aceeași caracteristică, cu unele variații, ceea ce ne face să constatăm că metrica LSA este mai precisă decât cealaltă, chiar dacă valorile medii returnate sunt oarecum mici.

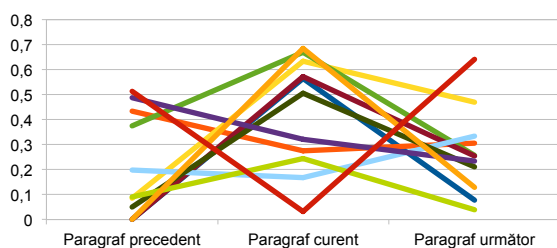


Figura 7. Verbalizări conținând parafraze comparate folosind metrica bazată pe vocabular

CONCLUZII

Pornind de la cercetările realizate de McNamara am început dezvoltarea unei aplicații care integrează multiple tehnici de prelucrare a limbajului natural și care vizează să evalueze în mod automat auto-explicațiile date de elevi în timpul lecturii, să le împartă în categorii și să ofere indicații corespunzătoare cititorilor.

Astfel, pentru a determina natura verbalizărilor am folosit LSA și o euristică bazată pe vocabularul folosit pentru a compara verbalizarea furnizată cu paragrafele alăturate; pe această abordare a oferit rezultate încurajatoare dintr-o perspectivă, însă limitate vizavi de mulțimea tuturor structurilor care ar trebui identificate automat.

Până în prezent am reușit să identificăm parafrazele cu o precizie destul de bună și să realizăm că nu se pot obține alte informații utile doar prin tehnicile folosite până în prezent. În consecință este nevoie să implementăm și alte strategii și să folosim cu atenție instrumentele pe care le avem la dispoziție pentru a îmbunătăți rezultatele și a identifica un număr mai mare de tipuri de verbalizări.

Cercetările viitoare se vor focaliza pe găsirea de similarități între verbalizări și diferite părți din text pentru a înțelege mai bine cât din informația prezentată a fost folosită și înțeleasă de utilizator, pentru a oferi explicații suplimentare și pentru a evalua rezultatele folosind un model formal de analiză a discursului.

REFERINȚE

- Chi, M.T.H., de Leeuw, N., Chiu, M.H., & LaVancher, C.. Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 439-477, 1994
- McNamara, S. D., Scott, F. J.. *Training Reading Strategies*, Old Dominion University, Department of Psychology, p387-392, Norfolk USA, Erlbaum, (1999)
- McNamara, D.S.. *Reading comprehension strategies: theories, interventions, and technologies*, 398-403, New York, Erlbaum, 2007
- O'Reilly, T., McNamara, S.D., Sinclair, G.P.. *iSTART: a web-based reading strategy a Intervention that improves students' science comprehension*, University of Memphis, 2004
- Dennis, S., Kintsch, W., Landauer, T.K., McNamara, S.D., *Handbook of latent semantic analysis*, (2007).
- Fellbaum, C., *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998
- Porter, M.F.. - *Snowball: A language for stemming algorithms*, available online at <<http://snowball.tartarus.org/texts/introduction.html>> 2001
- Schmid, H.. *Probabilistic Part-of-Speech Tagging Using Decision Trees*. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, available online at <ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf>, 1994
- Idzelis, M.. *The Java Open Source Spell Checker*, available online at <<http://jazzy.sourceforge.net/>>
- Jackson, G., Guess, R., McNamara, D.. *Assessing Cognitively Complex Strategy Use in an Untrained Domain*, University of Memphis, 2009
- Stamou S., Oflazer K., Pala K., Christodoulakis D., Cristea D., Tufis D., Koeva S., Tot-kov G., Dutoit D., Grigoriadou M.. "Balkanet: A Multilingual Semantic Network for Balkan Languages". In Proceedings of the 1st Global Wordnet Conference, Mysore, India, 2002