TEORIA DEPENDENȚEI CONCEPTUALE – INSTRUMENT PENTRU SUMARIZARE

Autor: Victoria LAZU

Universitatea Tehnică a Moldovei
Email: lazu_vica@mail.utm.md

Abstract: *Sumarizarea reprezintă un aspect foarte important în ceea ce priveşte coordonarea numărului vast de texte citite de oameni. Sumarizarea are rolul de a reduce cantitatea de informaţie care trebuie citită şi oferă posibilitatea de a decide gradul de relevanţă al informaţiei necesare. Odată cu utilizarea calculatoarelor în cadrul procesării textelor scrise, una dintre primele sarcini apărute este cea de a realiza sumarul unui text care va prezenta cît mai succint conţinutul, prin reducerea lungimii documentului, menţinînd totodată sensul şi păstrînd ideile de bază.*

Cuvinte cheie: *sumarizare, abstracţie, reprezentarea cunoştinţelor, dependenţa conceptuală, entităţi, acţiuni primitive, atribute, reguli semantice.*

1. Introducere

Sumarizarea automată a textelor este un domeniu destul de vast şi variat şi include sarcini ce nu au fost încă exploatate sau nu au adus rezultate destul de relevante, deoarece limbajul natural este destul de ambiguu şi întîmpină dificultăţi în procesul de procesare a lui. În era calculatoarelor este firesc să apară dorinţa de a face computerele să înţeleagă limba vorbită şi scrisă, să caute în texte întrebările şi răspunsurile, să deducă concluzii logice asupra conţinutului lor, să le poată sumariza şi generaliza, adică să realizeze ceea ce face, o persoană, în mod normal cu textul. Sumarizarea are rolul de a reduce cantitatea de informaţie care trebuie citită şi oferă posibilitatea de a decide gradul de relevanţă al informaţiei necesare. Dependenţa conceptuală este o teorie despre modul de reprezentare a diferitelor tipuri de cunoştinţe, despre evenimentele conţinute în mod uzual în propoziţii exprimate în limbaj natural. Ea exprimă relaţiile dintre concepte într-o manieră independentă de limba în care au fost exprimate la origine propoziţiile.

2. Sumarizarea

Sumarizarea automată este crearea unei versiuni mai scurte a unui text de către un program. Rezultatul acestei operaţii conţine totuşi majoritatea punctelor importante din textul original. Tehnologiile care construiesc un sumar coerent, dintr-un text de orice natură, trebuie să ia în considerare diverse variabile precum lungimea, stilul de scriere şi sintaxa pentru a realiza un sumar util. Un sumarizator este un sistem de prelucrare automat a unuia sau mai multor texte, cu scopul obţinerii unui rezumat (sumar) util unui utilizator uman. După forma de rezumare se disting două posibilităţi de sumarizare: *extracţie* şi *abstracţie*.
Tehnicile de extracţie copiază informaţiile considerate cele mai importante din text în sumar adică sumarul este complet compus din secvenţe de cuvînte (propoziţii, fraze sau cuvinte cheie), copiate din documentul original în timp ce *tehnicile de abstracţie* implică parafrazarea unor secţiuni din documentul sursă, ele conţin cuvinte, secvenţe ce nu sunt prezente în originale. În general, abstracţia poate condensa un text mai bine decît extracţia, dar astfel de programe sunt mai greu de dezvoltat şi necesită utilizarea tehnologiei de generare de limbaj natural, domeniu care este în dezvoltare.
Modelul de abstracţie conţine atît analiza morfologică a propoziţiilor cît şi cea sintactică şi semantică.
Aşa cum arată figura 1, abstractizarea are două abordări de bază. Prima (partea de sus a figurii) foloseşte o metodă tradiţională lingvisticii care analizează propoziţii sintactic. Această metodă utilizează informaţii semantice pentru a adnota arborii de analiză. Procedurile de compactare funcţionează direct pe baza arborilor pentru a elimina şi regrupa o parte din propoziţii de exemplu, prin tăiere arborilor în conformitate cu criteriile structurale.
A două abordare îşi are rădăcinile în inteligenţa artificială şi se concentrează asupra înţelegerii limbajului natural. Inteligenţa artificială în special necesită cunoaştere.

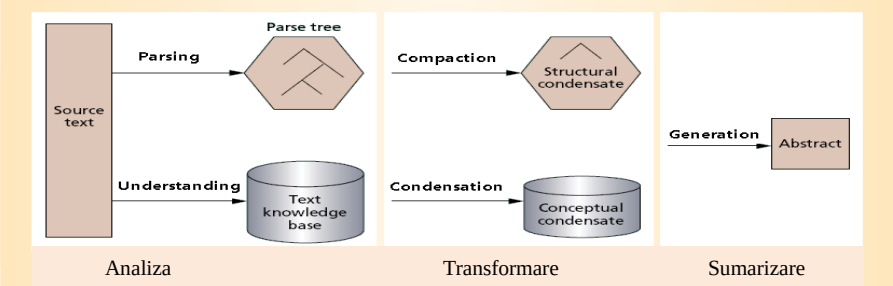

Fig. 1. Arhitectura de abstracție

Reprezentarea cunoştinţelor poate fi dificilă: cunoaşterea e voluminoasă, e greu de caracterizat cu precizie şi e în permanentă schimbare.
Modalităţile tradiţionale de reprezentare a cunoştinţelor sunt: - Reţelele semantice (apropiate de limbajul natural, permiţînd ierarhizarea conceptelor); - Dependenţa conceptuală (independentă de limbajul natural); - Cadrele (pentru situaţii stereotipe); - Scenariile (pentru secvenţe de evenimente stereotipe).

3. Dependenţa conceptuală

Construirea bazelor de cunoştinţe mari şi a sistemelor care exploatează aceste cunoştinţe trebuie să ţină cont de eficienţa reprezentărilor interne asociate unui model, de capacitatea de acces şi modificare a cunoştinţelor. În plus, s-a constatat ca şi la nivel simbolic, achiziţia cunoştinţelor, înţelegerea, utilizarea şi întreţinerea bazelor de cunoştinţe este dificilă în lipsa unei organizări conceptuale adecvate. Aceste motive au dus la apariţia unei noi clase de modele de reprezentare a cunoştinţelor: cunoştinţele structurate.
Cele mai importante reprezentări structurate ale cunoştinţelor sînt: reţelele semantice şi grafurile conceptuale, unităţile, numite în diverse abordări şi cadre sau prototipuri, dependenţele conceptuale şi scenariile. Toate aceste modele prezintă o caracteristică comună: oferă posibilitatea de a organiza şi structura cunoştinţele în funcţie de relaţiile existente între obiectele universului problemei.
Paradigmele de reprezentare se numesc modele de cunoştinţe structurate tocmai datorită accentului fundamental pe care îl pun pe structura cunoştinţelor în reprezentare, prin indexarea cunoştinţelor în funcţie de obiectele importante în sistem. Dacă sistemul bazat pe cunoştinţe are nevoie de una din informaţiile asociate unui astfel de obiect, structura de cunoştinţe asociată obiectului este regasită şi toate faptele relevante despre acel obiect sunt identificate deodată. Modelele cunoştinţelor structurate au asociate metode de inferenţă specifice, incluse direct în reprezentare.
Dependenţele conceptuale şi scenariile pot fi caracterizate drept "structuri tari" de reprezentare a cunoştinţelor, deoarece ele includ noţiuni specifice despre tipul obiectelor şi relaţiilor existente între acestea. Structurile tari de reprezentare a cunoştinţelor pot fi puse în corespondenţă, la nivelul controlului, cu strategiile de căutare informată, în care fiecare algoritm de căutare include informaţie euristică specifică domeniului problemei.
Dependenţele conceptuale sunt un model structurat care permite reprezentarea cunoştinţelor conţinute în propoziţiile limbajului natural. Scopul dependenţelor conceptuale este acela de a reprezenta cunoştinţele astfel încît:
- reprezentarea să fie independentă de limbajul în care au fost formulate propoziţiile;
- propoziţii diferite, dar avînd aceeaşi semnificaţie, să aibă aceeaşi reprezentare sub forma dependenţelor conceptuale;
- reprezentarea semnificaţiei propoziţiilor să fie neambiguă;
- să faciliteze execuţia inferenţelor determinate de informaţiile conţinute în propoziţii.

Din aceste motive, dependenţele conceptuale folosesc o reprezentare a propoziţiilor limbajului natural care nu se bazează pe cuvintele propoziţiei ci pe o mulţime de entităţi primitive conceptuale care pot fi combinate pentru formarea semnificaţiei cuvintelor şi a propoziţiei. Acest model a fost propus de Schank în 1972 şi a fost folosit în diverse programe de înţelegere a limbajului natural.
În teoria dependenţelor conceptuale se disting cinci tipuri de elemente componente (primitive ontologice) care constituie componentele constructive ale reprezentării. Aceste elemente de bază sunt: entităţi, acţiuni, cazuri conceptuale, timpuri conceptuale şi dependenţe conceptuale, fiecare element avînd o serie de subtipuri.

Entităţi

Obiectele sau producătorii de scenarii, numite pe scurt PP, reprezintă persoanele (actorii) sau obiectele fizice (inclusiv memoria umană) care actionează în universul discursului.

Atributele sau ucenicii scenariilor, numite pe scurt PA, reprezintă proprietăţile producătorilor de scenarii.

Acţiuni

Acţiunile primitive, numite pe scurt ACT. În reprezentarea dependenţelor conceptuale acţiunile sunt formate dintr-o mulţime de acte primitive. Un set tipic de astfel de acţiuni primitive este următorul:

- ATRANS - Transferul făcut de o relaţie abstractă, de exemplu "a da"
- PTRANS - Transferul locaţiei fizice a unui obiect, de exemplu "a merge"
- PROPEL - Aplicarea unei forţe fizice unui obiect, de exemplu "a împinge"
- MOVE - Mişcarea părţii unui corp de către acel corp, de exemplu "a lovi"
- GRASP - Apucarea unui obiect de către un actor, de exemplu "a smulge"
- INGEST - Ingerarea unui obiect de către o fiinţă, de exemplu "a mînca"
- EXPEL - Expulzarea unui obiect din corpul unei fiinţe, de exemplu "a plînge"
- MTRANS - Transferul unei informaţii mentale, de exemplu "a destăinui"
- MBUILD - Construirea informaţiilor noi din cele vechi, de exemplu "a decide"
- SPEAK - Producerea sunetelor, de exemplu "a vorbi"
- ATTEND - Concentrarea atenţiei unui organ de simt asupra unui stimul, de exemplu a asculta"

Ucenicii acţiunilor, numiţi pe scurt AA, reprezintă atribute sau proprietăţi ale acţiunilor primitive.

Cazuri conceptuale

Caz obiect (o)
Caz direct (D)
Caz instrument (I)
Caz recipient (R)

Timpuri conceptuale

Condiţional (c)	Interogativ (?)
Continuu (k)	Tranziţie (t)
Prezent (nil)	Începutul tranziţiei (ts)
Trecut (p)	Sfîrşitul tranziţiei (tf)
Viitor (f) Negativ (/)	Atemporar (delta)

Timpurile conceptuale oferă o modalitate de modificare a descrierii evenimentelor prin indicarea timpului, modului sau aspectului unui verb, deci a informaţiilor care apar de obicei în limbajul natural.

Dependenţe conceptuale

Aceste dependenţe reprezintă reguli semantice de formare a structurilor de dependenţe pe baza entităţilor şi a acţiunilor, cum ar fi relaţia între un actor şi un eveniment sau relaţia între o acţiune primitivă şi un instrument. Există o serie de astfel de dependenţe conceptuale stabilite, o parte dintre cele mai importante fiind prezentate în Figura 2. În această figură prima coloană conţine forma dependenţei conceptuale, a doua coloană conţine un exemplu de utilizare a dependenţei conceptuale şi a treia coloană conţine exprimarea în limbaj natural a acestei dependenţe.

Elementele componente prezentate sunt folosite pentru crearea *structurilor conceptuale,* numite şi structuri de dependenţe sau conceptualizări, care pot descrie adecvat înţelesul unei propoziţii în limbaj natural. O astfel de structură conceptuală se reprezintă printr-o formă particulară de graf. În reprezentarea grafică a dependenţelor conceptuale nodurile reprezintă entităţi sau acţiuni, săgeţile indică direcţia dependenţelor, săgeţile duble indică legături bidirecţionale între actori (entităţi) şi acţiuni. În plus, toate legăturile (arcele) din graf au asociate etichete corespunzătoare relaţiilor de caz, deci cazurilor conceptuale, sau relaţii temporare, deci timpurilor conceptuale.

Regula 1 descrie relaţia între un actor şi evenimentul cauzat de acesta. Relaţia de dependenţă este bidirecţională deoarece nici actorul nici evenimentul nu pot fi considerate mai importante în relaţie. Litera p cu care este etichetata legatura indică timpul trecut.

Regula 2 descrie relaţia între un producător de scenarii şi un ucenic de scenarii. Multe descrieri de stări, cum ar fi înalţime, sunt reprezentate în dependenţele conceptuale prin valori cuantificate.

Regula 3 descrie relaţia între o acţiune primitivă şi instrumentul cu care se execută acea acţiune. Instrumentul trebuie să fie întotdeauna o conceptualizare, i.e. trebuie să conţină o acţiune şi nu un simplu obiect fizic.
Regula 4 descrie relaţia între o conceptualizare şi o altă conceptualizare care a cauzat-o pe prima. Se observă că sagetile indică dependenţa unei conceptualizări de o alta şi de aceea au o direcţie opusă direcţiei implicaţiei existente între concepte. Cele două forme ale regulii descriu cauza unei acţiuni (a) şi cauza unei schimbări de stare (b).
Regula 5 descrie relaţia între o conceptualizare şi o altă conceptualizare ce reprezintă timpul în care s-a produs prima.

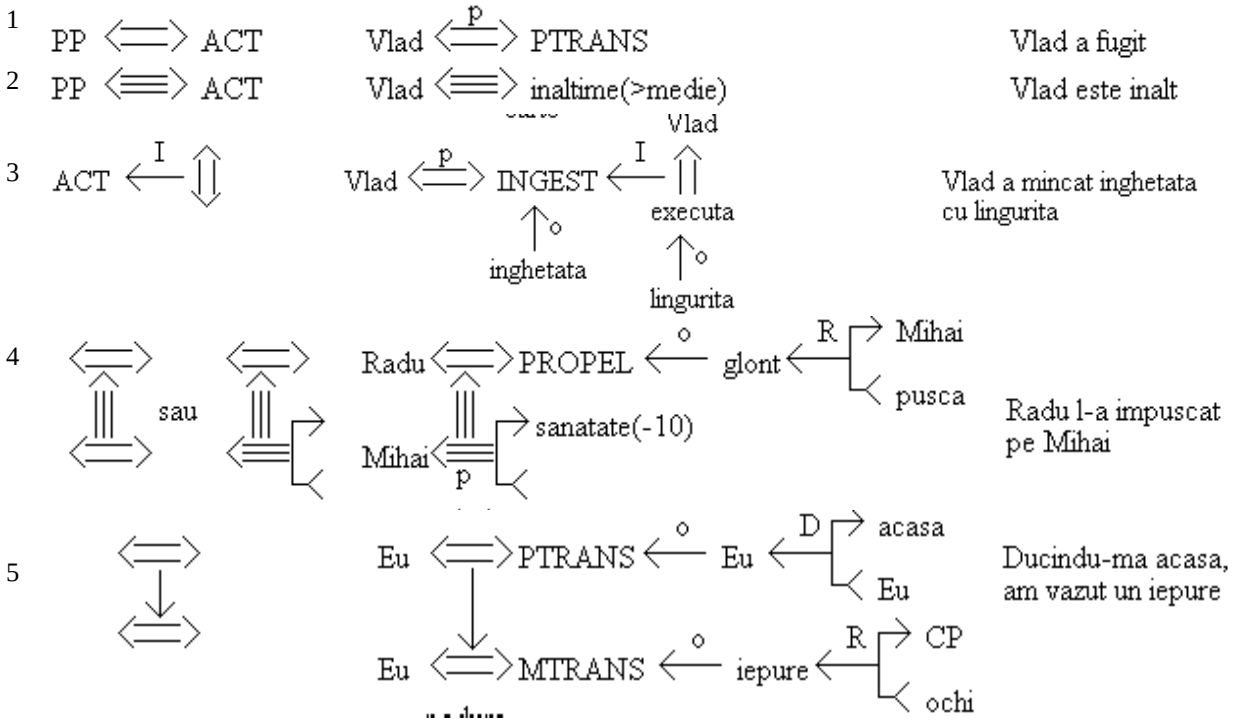

Fig. 2. Dependenţe conceptuale

Exemplul acestei reguli pune în evidenţă modul în care dependenţele conceptuale imită o prelucrare informaţională specifică modelului mental uman: acţiunea "a vedea" este reprezentată ca un transfer de informaţie între ochi şi un element de prelucrare conştient.

Utilizînd aceste structuri conceptuale de bază, se pot construi structuri complexe corespunzătoare semnificaţiei propoziţiilor în limbaj natural.

Din punct de vedere al reprezentării cunoştinţelor şi a facilităţilor de raţionament, dependenţele conceptuale oferă următoarele avantaje:

(1) Numărul de reguli de inferenţă necesare rezolvării problemei scade.
(2) Multe reguli de inferenţă sunt deja incluse în reprezentare.
(3) Structura iniţial construită pentru reprezentarea informaţiei dintr-o propoziţie conţine şabloane care trebuie completate. Aceste şabloane pot fi folosite ulterior ca un mecanism de focalizare a atenţiei programului pentru înţelegerea propoziţiilor următoare.

4. Concluzie

Sumarizarea automată a textului ia unul sau mai multe texte şi extrage "cea mai importantă" informaţie sau informaţia legată de aspectul ales de către utilizator. Există sisteme care ar putea fi foarte utile pentru cercetători care au nevoie să stabilească rapid conţinutul unui articol. Din păcate, cu tehnologia actuală, este dificil să se producă automat un rezumat, care să înlocuiască întregul document.
Reprezentarea cunoştinţelor prin dependenţă conceptuală presupune o abordare decompoziţională, care analizează propoziţia în limbaj natural pentru a reţine structura sa de bază, elementele semantice

fundamentale şi facilitează execuția inferențelor determinate de informațiile conținute în propoziții. Ceea ce va duce la uşurarea sumarizării informației.

Bibliografie

1. Mani, Inderjeet (2001). *Automatic Summarization.* ISBN 1-58811-060-5
2. Marcu, Daniel (2000). *The Theory and Practice of Discourse Parsing and Summarization.* ISBN 0-262-13372-5
3. Schank, R., "*A Conceptual Dependency Representation for a Computer-Oriented Semantics*", Ph.D. Thesis University of Texas, Austin 1969
4. Schank, R.C. (1975). *Conceptual Information Processing.* New York: Elsevier.
5. Roger C. Schank, *The Cognitive Computer*, On Language, Learning and Artificial Intelligence. Addison, Wesley Publishing Company, Inc., Reading. 1984