RECUNOAȘTEREA UMORULUI ÎN TEXTE

Marius MANTALUȚA[1]

Universitatea Tehnică a Moldovei, Facultatea Calculatoare, Informatică și Microelectronică, Departamentul Informatică și Ingineria Sistemelor, gr. IA-191, Chișinău, Republica Moldova

*Autorul corespondent: Marius Mantaluța, mantaluta.marius@iis.utm.md

Rezumat. *Comunicarea om calculator nu mai constituie demult un deziderat iluzoriu al inteligenței artificiale. Pentru ca această comunicare să fie una cât mai apropiată de comunicarea interumană, calculatorul (va trebui nu numai să recunoască, ci și să folosească și umorul. Mai mult, umorul oferă profunzimi ale limbajului uman - referindu-ne la cel real, complex, un limbaj creativ nu doar o mulțime de propoziții standard. Reușind să modelăm înțelegerea și generarea umorului de către calculatoare, câștigăm o mai bună imagine asupra modului în care creierul uman funcționează nu doar în privința umorului ci și a limbajului și cunoașterii în general. Sunt multe situații în interacțiunea interumană unde umorul joacă un rol important permițând continuitatea conversației, întărind relațiile interumane.*

Cuvinte cheie*: umor, prelucrarea textuală, tehnologiile limbajului natural, umor.*

Introducere

Umorul are un caracter specific (literar, subtil sau fin) prin soluții neașteptate, caraghioase care pot produce ilaritate. Persoanele cu umor sunt acele persoane care prin comportare sau prin vorbe, în anumite contexte, declanșează râsul. Simțul umorului este influențat de tradițiile, cultura, istoria unui popor, sau diferă după poziția pe scara ierarhiei sociale sau după etate. Nu doar variază de la o persoană la alta, dar aceași persoană poate să găsească o glumă ca fiind amuzantă într-o zi și în altă zi nu, depinzând de starea de spirit a persoanei, de evenimentele recent petrecute în viața persoanei respective. Umorul poate fi usturător prin - satiră, ironie, batjocură, cinic, sau blând, binevoitor, plin de înțelegere, autocritic. Umorul computațional este un domeniu în care există unele abordări de găsire a unui șablon universal pentru generarea și recunoașterea umorului în texte. Studierea prin mijloace computaționale a umorului este un domeniu ale cărei baze au început să se pună abia în ultimii ani, neexistând o teorie general acceptată.

Generatoare de umor

- **LBJOJG -** (Light Bulb Joke Generator), dezvoltat de Attardo si Raskin în 1993 și este un generator de glume de tipul "*De câți (substantiv) este nevoie pentru a (verb)*", însă era foarte limitat deoarece nu ansamblează sau analizează atribute ale glumelor.
- **HAHAcronym -** dezvoltat de Stock si Straparava pentru un sistem care generează versiuni amuzante ale acronimelor deja existente. Efectul comic s-a obținut mai ales prin exploatarea teoriei de nepotrivire. Se va păstra o parte din cuvintele care definesc acronimul iar înlocuirea celorlalte cuvinte se face prin: utilizarea unui câmp semantic opus și păstrarea literei inițiale, a ritmului și rimei. Exemplu**:** **ACM** (Association for Computing Machinery) devine Association for Confusing Machinery.

Recunoșterea umorului

Pentru a recunoaște umorul fără a face toate conexiunile dintre cuvintele unei propoziții, se calculează valori ale unor atribute ce pot caracteriza un text. În funcție de valorile acestor caracteristici se încearcă o clasificare a textelor. Se folosesc combinații de atribute pentru a vedea dacă acestea sunt suficiente pentru recunoașterea umorului sau nu:

- **Cea mai apropiată glumă**: Din datele de antrenament se caută gluma cea mai apropiată de textul pe care îl testăm. Gradul de apropiere dintre 2 texte se calculează după numărul de cuvinte comune celor 2 texte *[Sjobergh and Araki, 2007].*

- **Cea mai apropiată non-glumă**: Pe același principiu se determină cea mai apropiată non-glumă [*Sjobergh and Araki, 2007]*.
- **Cuvinte amuzante**: S-a observat că unele cuvinte sunt comune dar unele sunt specifice doar glumelor. Pentru a suprprinde acest aspect, cuvintele care apar măcar de 5 ori în datele de antrenament și dacă apar de 5 ori mai des în glume decât în non-glume sunt păstrate într-o listă *[Sjobergh and Araki, 2007]*.
- **Ambiguitatea cuvintelor**: se calculează uitându-se într-un dicționar online și numărând sensurile cuvintelor.
  - Ambiguitatea medie: numărul mediu de ambiguități într-o propoziție (media numărului de sensuri pentru fiecare cuvânt);
  - Ambiguitatea maximă: cea mai mare valoare a numărului de sensuri pentru un cuvânt dintr-o propoziție.
- **Cuvinte murdare**: numărul de cuvinte murdare prezente în propoziții. O listă cu 2500 de cuvinte murdare downloadată de pe Internet a fost folosită pentru a se decide dacă un cuvânt este murdar sau nu *[Sjobergh and Araki, 2007]*.

Se poate încerca recunoașterea umorului folosind învățarea automată, utilizând clasificatorul *Naive Bayes si Support Vector Machine*. Rada Mihalcea și Carlo Strapparava au ales să își restricționeze studiul la *one-linere*. [Mihalcea, May 2006]. Un *one-liner* este o propoziție cu efect comic și cu o structură lingvistică interesantă: sintaxă simplă, folosirea deliberată a unor instrumente retorice. În timp ce glumele mai lungi tind să producă umor printr-o structură narativă mai complexă, one-linerele produc efectul comic dintr-o lovitură, cu foarte puține cuvinte. Acest lucru face ca, acest tip de text să fie folosit pentru recunoașterea automată a umorului, deoarece efectul comic se produce în prima și singura propoziție. *Exemplu*: ”Everyone has a photographic memory. Not everyone has film.” Datele umoristice au fost alcătuite din one-linere colectate de pe Internet folosind procesul de bootstrapping. Evaluarea *Bootstrap* este o procedură care utilizează un computer care oferă o alternativă flexibilă și automată. Calculatorul preia mii de probe de bootstrap din datele observate și din aceste probe de bootstrap, estimează precizia statisticii. În special, bootstrap-ul poate fi folosit pentru a găsi erori standard aproximative. Distribuția probelor bootstrapping este concretă și permite compararea valorilor parametrice față de probele repetate care au fost trase (cu înlocuire) din eșantionul inițial.

Concluzie

Umorul este una din cele mai dificile caracteristici de recunoscut nu doar pentru un calculator ci și pentru oameni în general. Odată cu dezvoltarea Inteligenței artificiale va fi necesară și dezvoltarea diferitor limbaje de comunicare, umorul fiind unul din ele. Integrarea umorului computațional în tehnologii va provoca o comunicare mai efectivă, dintre om și mașină. În viitor, recunoașterea umorului de către roboți, și chiar a unor sentimente (bucurie, agresivitate, tristețe, etc.), în texte, poate permite roboților cu o inteligență artificială dezvoltată să emite niște emoții, dar poate crea și unele neînțelegeri și conflicte demonstrate în unele filme și jocuri (iRobot; Detroid Become Human).

Mulțumiri

Ținem să mulțumim pentru ajutorul acordat în realizarea acestui articol dnei Daniela Istrati, lector universitar la Departamentul Informatcă și Ingineria Sistemelor.

Bibliografie

1. Recunoașterea umorului în texte – articol ralizat de Țifrea Oana. https://profs.info.uaic.ro/~corinfor/Humor-Oana.pdf
2. Characterizing Humour: An Exploration of Features in Humorous Texts https://link.springer.com/chapter/10.1007/978-3-540-70939-8_30
3. Technologies That Make You Smile: Adding Humor to Text-Based Applications https://ieeexplore.ieee.org/document/1705426