Semantică distribuțională pentru analiza poeziilor

Adrian Manea, 510 SLA

3 aprilie 2020

Cuprins

Introducere

Proiectul de față prezintă aspecte introductive privitoare la *semantica distribuțională*, împreună cu două aplicații relevante care se concentrează pe studiul poeziilor scrise în limba engleză.

Astfel, materialul urmărește în primul rînd prezentarea elementelor fundamentale de semantică distribuțională, într-o manieră potrivită pentru cei nefamiliarziați cu domeniul (cum a fost și cazul autorului la începutul documentării pentru această lucrare). În prima parte a lucrării, așadar, se vor introduce principalele noțiuni teoretice privitoare la această metodă de studiu empiric al semanticii, folosind obiecte matematice surprinzătoare, anume spații vectoriale. Ne vom limita, însă, la noțiunile ce vor fi de folos efectiv în restul lucrării, iar pentru mai multe detalii, vom indica referințe bibliografice relevante.

Metodele distribuționale de atribuire a semanticii se vor folosi apoi pentru expunerea a două exemple concrete, ambele legate de texte poetice. Primul exemplu studiază cîteva poezii moderne și contemporane din literatura engleză, pentru a justifica faptul că, deși semantica poetică nu este mereu la îndemîna tuturor celor familiarizați doar cu limbajul comun, se pot identifica suficiente indicii că ea este bazată pe acest limbaj. Se urmărește, în particular, *coerența subiectelor* din textele poetice, adică, odată identificate tipare privitoare la subiectul poeziilor, sîntem interesați dacă vocabularul din contextul subiectului respectiv este relevant pentru subiectul în sine.

A doua aplicație pe care o prezentăm încearcă să răspundă la o controversă privitoare la opera poetică a lui Lord Byron și cea a lui Thomas Moore. Cum ambii au fost poeți reprezentativi pentru un anume curent literar — orientalismul romantic — se pot remarca suficiente conexiuni între poeziile celor doi. Folosind metode de semantică distribuțională, se încearcă răspunsul la întrebarea dacă aceste conexiuni sînt inerente, cumva impuse și induse de încadrarea în (sub)genul literar respectiv sau într-adevăr se poate vorbi despre imitație.

Înainte de a începe efectiv prezentarea, mai menționăm încă o dată că lucrarea se dorește a fi una introductivă și de ansamblu, astfel că majoritatea subiectelor incluse sînt tratate la un nivel elementar. Detalii suplimentare pot fi găsite atît în referințele indicate explicit, cît și în bibliografiile acestora.

1 Aspecte teoretice

1.1 Premisele semanticii distribuționale

Așa cum se explică în [Boleda, 2019] și [Boleda şi Herbelot, 2016], semantica distribuțională (SD) s-a dovedit a fi foarte utilă în științele cognitive, în cercetări privitoare la asimilarea și atribuirea sensurilor unor cuvinte din limbajul comun, însă aplicațiile la lingvistica teoretică și computațională au fost, comparativ, mai rare.

Această ramură a lingvisticii se bazează pe *ipoteza distribuțională*, care afirmă că similaritatea semantică rezultă în similarități ale unor distribuții statistice. Spus altfel, cuvintele relaționate semantic apar în contexte similare. Însă SD procedează invers: pornește de la corpusuri de text și induce reprezentări semantice pe baza distribuțiilor observate.

O remarcă istorică relevantă este că această metodă a apărut în anii 1950 în Statele Unite, încadrată într-un curent mai general al lingvisticii, *distribuționalismul*, care urmărea să folosească metode de teoria distribuțiilor pentru diverse studii cantitative și calitative ale limbajului. Părinții curentului au fost Leonard Bloomfield și Zellig S. Harris, iar teoria gramaticilor generative formulată de Noam Chomsky este considerat a fi puternic influențată de aceasta. Totodată, metodele distribuționaliste specifice și-au adus contribuția inclusiv în didactica limbilor străine.

Intuitiv, ideile SD sînt foarte simple: în multe situații, putem ghici sensul unui cuvînt din contextele în care acesta este folosit. Astfel, este ca și cum reprezentăm si interpretăm sensul ca pe o distribuție în fundalul contextelor lingvistice în care se observă utilizarea cuvintelor-cheie. Ideea poate fi trasată încă de la filosoful Ludwig Wittgenstein, care afirma în *Tractatus Logico-Philosophicus* (1922) că „sensul unui cuvînt este utilizarea lui în limbaj“.

Partea dificilă în acest studiu este, însă, să se explice precis ce se înțelege prin „context“. În varianta naivă (și chiar utilizată în multe studii), contextul înseamnă pur și simplu vecinătatea cuvîntului-țintă, implementată cu ajutorul unei ferestre de conținut, i.e. un interval de lungime fixată, centrat în cuvîntul căruia dorim să îi analizăm semantica.

De obicei, metodele matematice cu ajutorul cărora se implementează modelele distribuționale sînt, oarecum surprinzător, *spațiile vectoriale.* Dimensiunea spațiilor este legată de cuvintele din context, într-un mod pe care îl detaliem mai jos, iar astfel, coordonatele depind de cuvintele co-ocurente. Rezultă că, prin această abordare, vecinătatea în context devine vecinătate în spațiu, într-un mod cît se poate de concret.

Din punctul de vedere al implementării informatice, modelele distribuționale pot fi învățate dintr-un corpus într-o manieră nesupervizată.

1.2 Construcția spațiului vectorial semantic

Vom prezenta foarte pe scurt modul în care se construiește spațiul vectorial cu ajutorul căruia se poate face studiul semantic. Mai multe detalii se pot găsi într-un articol mult mai tehnic pe care îl recomandăm, [Turney şi Pantel, 2010].

Concret, se dă un corpus de cuvinte, de exemplu cîteva cărți sau pagini de Internet din care se poate extrage informație text suficient de multă. Se dă, de asemenea, un cuvînt-țintă căruia se dorește să i se stabilească sensul, în cazul acestei abordări. Cuvîntul-țintă se poate schimba pe parcursul analizei, bineînțeles, însă acum presupunem că avem un cuvînt-țintă fixat la care ne referim. Se înregistrează co-ocurențele relativ la acest cuvînt-țintă, adică ce alte cuvinte apar în vecinătatea lui, de exemplu, într-o fereastră de conținut de 10 cuvinte. Mai precis, pentru fiecare apariție a cuvîntului-țintă în corpusul folosit, se iau în considerare 5 cuvinte care urmează după el și 5 cuvinte care-l preced. În majoritatea situațiilor, în funcție de partea de vorbire care este cuvîntul-țintă, se pot lua în considerare doar părți de vorbire relevante. De exemplu, dacă vrem să studiem un cuvînt-țintă care este un substantiv, am putea ignora prepozițiile, punctuația, pronumele și numele proprii și vom păstra verbele, adjectivele, adverbele.

Odată stabilite și filtrate aceste co-ocurențe, se construiește un vector asociat cuvîntului-țintă, sortat descrescător după numărul de apariții. De exemplu, dacă corpusul este, să zicem, Enciclopedia Britannica, iar cuvîntul-țintă este „coleopter“, vectorul său asociat poate fi de forma:

$$v = \begin{pmatrix} 244 & 111 & 35 & 21 & 5 & 2 & 2 & 1 & 1 & 1 & 1 \end{pmatrix}^t,$$

corespunzător co-ocurențelor *aripă, larvă, copac, floare, cîmpie, cald, roșu, frumos, reprezentant, trifoi.* Deducem că cel mai adesea, cuvîntul „coleopter“ a apărut în vecinătatea (contextul) cuvintelor *aripă, larvă, copac, floare*, celelalte co-ocurențe fiind semnificativ mai rare.

Atunci, spațiul vectorial (numit și *spațiu semantic*) asociat acestui cuvînt are dimensiune 10 (dată de fereastra de conținut, de fapt). Evident, se pot face unele optimizări, precum ignorarea frecvențelor sub un anumit prag (în cazul nostru, frecvențele ≤ 5 am putea spune că sînt irelevante, dată fiind mărimea corpusului). O altă optimizare este să se ignore cuvintele prea generale. De exemplu, dacă întîlnim în preajma cuvîntului-cheie și cuvinte precum *foarte, mare, mic*, acestea pot fi ignorate, pentru că se potrivesc în prea multe contexte pentru a fi specifice cuvîntului-cheie.

Odată construit, spațiul semantic poate fi folosit în mai multe moduri. În cele mai multe cazuri, se asociază spațiului un cvadruplu $\langle A, S, B, M \rangle$, unde:

- A este o funcție care asociază ponderi relevante co-ocurențelor (depinde de la caz la caz);

- S este un coeficient de similaritate, de obicei calculat drept cosinusul unghiului dintre vectorii asociați cuvintelor, folosind produsul scalar euclidian;
- B sînt elementele din bază (sau dimensiunea);
- M este o transformare a spațiului, de obicei una care poate reduce dimensiunea, pentru a face calculele mai relevante.

Un exemplu des întîlnit este *analiza semantică latentă* (LSA) , introdusă în 1997, care folosește:

- A — un indicator numit tf-idf, acronim provenit de la denumirea în engleză, *term frequency-inverse distribution frequency*, care produce rezultate mai relevante decît numărul de co-ocurențe;
- S este cosinusul dintre vectori;
- B este o bază spațiului semantic;
- M este o descompunere folosind valori singulare, notată folosind acronimul englezesc SVD (*singular value decomposition*).

Nu insistăm, însă, pe această metodă, deoarece vom întîlni o îmbunătățire a acesteia într-o secțiune ulterioară, sub forma *analizei semantice explicite*, ESA. Indicăm doar ca referință pentru studiu suplimentar articolul relevant din Scholarpedia.

Alte abordări mai sofisticate înlocuiesc vectorii cu matrice multidimensionale (numite și *tensori*), precum și metode complexe de analiza bayesiană, care calculează funcții de distribuție și pot extrage subiectul despre care se vorbește în text (eng. *topic*).

1.3 Problema semanticii

Înainte de a merge mai departe cu metodele distribuționale specifice, împreună cu studiile de caz, ne oprim pentru o vreme asupra problemei semanticii în sine. Prezentăm cîteva aspecte istorice și filosofice, dar și literare care arată de ce semantica limbajului natural este o problemă, care au fost abordările cele mai cunoscute pentru rezolvarea acestei probleme și de ce, în particular, cazul poeziilor necesită atenție deosebită.[1]

Începînd cu lucrările lui Gottlob Frege (*Sens și referință*, 1892) și Alfred Tarski (*Conceptul semantic de adevăr și fundamentele semanticii*, 1944), semantica formală s-a plasat în contextul teoriei mulțimilor. În acest sens, cuvintele sînt gîndite precum *concepte*, care au *extensiune*. De pildă, cuvîntul „pisică“ este un concept, în a cărui extensiune intră toate pisicile din lume, adică identificăm extensiunea conceptului cu mulțimea tuturor obiectelor care îl instanțiază.

[1] Aplicațiile pe care le vom prezenta au în vedere poezii scrise în limba engleză. De aceea, vom da titlurile și citatele în original, fără a încerca o traducere.

Prin această abordare, se poate vorbi ușor despre adevăr și falsitate, în sensul că o propoziție care enunță o proprietate a unui obiect (concept) este luată ca adevărată numai atunci cînd extensiunea conceptului respectiv conține extensiunea proprietății *la nivel de incluziune între mulțimi*. De exemplu, propoziția „Toți caii sînt albi“ este adevărată dacă întreaga mulțime a cailor (i.e. extensiunea cuvîntului [conceptului] „cal“) este inclusă în mulțimea obiectelor albe (i.e. extensiunea cuvîntului [conceptului] „alb“).

Mai mult, această abordare ne permite să facem distincția semantică între construcții precum „soarele amiezii“ și „soarele apusului“, chiar dacă, de fapt, este vorba despre același obiect fizic.

Semantica poeziilor și, de fapt, întreaga întrebare *dacă* poeziile au semantică sau sens, în general, a fost una îndelung dezbătută. Poeziile sînt greu de studiat folosind tehnici de lingvistică, deoarece o bună parte a poeziilor, în special cele moderne și contemporane sau cele foarte vechi, nu respectă regulile stricte de sintaxă și semantică.

Totuși, semantica distribuțională se poate dovedi o unealtă foarte puternică în acest sens, deoarece tiparele de distribuție care se remarcă pot fi relevante și în cazul textelor poetice. Ideea din spatele acestei încrederi este că, deși în multe situații, limbajul poeziilor nu este reprezentativ pentru limbajul comun, anumite tipare pe care le distingem în poezii sînt, într-un fel, latente în limbajul nostru fundamental. De exemplu, sînt multe cuvinte al căror sens este neschimbat în poezii: cuvîntul „copac“, cel mai probabil (deși nu cert) apare în poezii cu același sens cu care îl găsim și în limbajul comun. Aceasta ne arată că, deși limbajul poetic poate fi considerat unul special, el cu siguranță *nu este complet disociat* de acesta, în sensul că orice text poetic, oricît de dificil, poate fi recunoscut ca fiind alcătuit din elemente de limbaj natural și mai mult, în majoritatea cazurilor, se poate identifica faptul că este alcătuit din construcții cu sens.

Un exemplu (preluat din [Herbelot, 2015]) de schimb de idei privitor la semantica poeziilor, prin comparație cu cea a limbajului natural este corespondența între filosoful Philip Wheelwright (1901-1970) și poeta și criticul literar Josephine Miles (1911-1985), din anii 1940. Conform lui Wheelwright, în lucrarea *On the Semantics of Poetry*, limbajul poeziilor este complet diferit de cel al științei. El consideră că sensul cuvintelor folosite în teorii științifice se bazează pe concepte, în vreme ce cuvintele din poezii au ceea ce el a numit *înțeles metalogic*, adică dat de o semantică nebazată pe logică. Miles a replicat afirmînd că ambiguitate există și în limbajul comun și că nu este specific poeziilor ca apariția unui cuvînt să depindă de context (spre deosebire de știință, unde același cuvînt apare de fiecare dată cu același înțeles, independent de context).

În favoarea lui Wheelwright este, într-adevăr, greu de susținut că expresii poetice precum:

- "*Music is the exquisite knocking of the blood*" (Rupert Brooke);
- "*Your huge mortgage of hope*" (Ted Hughes);
- "*Skeleton bells of trees*" (Avery Slater)

au o interpretare folosind teoria mulțimilor.

Dar în același timp, este greu și să combatem teza lui Miles, conform căreia semantica poeziilor își are rădăcinile în semantica limbajului comun. Într-adevăr, fără cunoașterea acesteia din urmă, construcțiile poetice sînt lipsite de orice fel de semantică, iar orice încercare de descifrare este sortită eșecului.

O poziție care iese din această dezbatere este aceea a lui Gerald Bruns, de la Universitatea Notre Dame din Indiana, SUA. În cartea [Bruns, 2005], el afirmă că *poezia este alcătuită din limbaj, dar nu este o utilizare a acestuia*, în sensul că acele cuvinte care apar în poezii nu ar trebui să fie privite ca fiind definite de contextul poetic. Bruns continuă prin a afirma că, asemenea lui Wittgenstein (*înțelesul este utilizarea*, eng. *meaning is use*), el consideră că extensiunea unui concept nu poate fi închisă de nicio frontieră, lăsînd, astfel, loc pentru utilizări atipice ale limbajului, precum poezia.

Aceasta sugerează că semantica, cel puțin în cazul poeziilor, ar trebui ancorată în context. Acest lucru relaxează frontierele, prin comparație cu teoria mulțimilor și afirmă că, de exemplu, sensul cuvîntului „pisică“ nu mai este legat de pisicile din lume, ci de modul în care oamenii vorbesc despre pisici. Regăsim aici primele indicații în direcția semanticii distribuționale: „responsabilitatea“ sensurilor pică pe context și, indirect, pe vorbitorii care plasează termenul în context.

Teoria distribuțională, însă, nu a prins roade pînă în a doua parte a anilor 1950, cînd dezvoltarea puterii de calcul a contat foarte mult pentru analiza corpusurilor de text.

Legătura istorică este una foarte puternică: o studentă a lui Wittgenstein, Margaret Masterman, a fondat Cambridge Language Research Unit (CLRU), una dintre instituțiile care s-au implicat timpuriu în lingvistica computațională în Marea Britanie. Tot ea a fost interesată si de aspectele de generare a limbajului, fiind una dintre cei care au lucrat la o primă variantă software de generare a poeziilor. În fapt, programul dezvoltat de ei nu producea poezie propriu-zisă, ci oferea sugestii contextuale pentru a umple spațiile dintr-un haiku (lucrare publicată, de exemplu, sub forma [Masterman, 1971]).

Cum puterea de calcul era încă limitată, comparativ cu zilele noastre, echipa lui Masterman nu a putut analiza corpusuri mari de text în sine. Ei au trebuit să dezvolte rețele semantice, care să le permită „comunicarea“ dintre diverse seturi, pe baza cărora au extras apoi informații statistice. Aceasta a fost una dintre mișcările de bază în direcția teoriei distribuționale.

2 Aplicație: Coerența semantică în poezia modernă

Prezentăm acum aplicația din articolul [Herbelot, 2015], asupra coerenței semantice în poezia modernă și contemporană, aplicație care vine în continuarea considerațiilor teoretice generale din capitolul anterior.

2.1 Teoria lucrării

Pentru a susține teza că poeziile folosesc o structură similară a limbii cu limbajul comun atunci cînd produc construcții cu sens, ar trebui să arătăm că limbajul poetic este caracterizat de o *coerență a subiectelor* (eng. *topic coherence*). Ideea de bază a acestei cercetări este să folosească metode distribuționale pentru a arăta că, de exemplu, mulțimea de subiecte { scaun, masă, birou, echipă } are o coerența mai mare decît { scaun, rece, elefant, nor }.

Poeziile studiate sînt scrise în perioada 1881-2008 și variază, ca stil, de la unele considerate dificile, la altele, „transparente".

Se definește *coerența unei mulțimi de cuvinte* (cf. [Newman et al., 2010]) $\{w_1, w_2, \dots, w_n\}$ ca fiind media similarităților lor în perechi (două cîte două):

$$\overline{S}(w) = \overline{\{\text{Sim}(w_i, w_j) \mid 1 \leq i < j \leq n\}},$$

unde bara superioară notează media. Similaritatea se definește cu ajutorul cosinusului:

$$\text{Sim}(A, B) = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}},$$

unde $n = 2000$ (vezi mai jos).

Corpusul folosit face parte din British National Corpus, care a fost lematizat și s-au asociat etichete folosind sistemul CLAWS, pe care nu îl detaliem, pentru că modul său de funcționare și implementare este irelevant pentru restul prezentării.

În plus, se ignoră punctuația, iar cuvintele-cheie fac parte din categoriile: substantive, verbe, adjective și adverbe. Fiecare poezie studiată se convertește în secvențe de cîte 11 cuvinte, iar contextul asociat unui cuvînt-cheie înseamnă cele 5 cuvinte care îl precedă și cele 5 care îl urmează.

Co-ocurențele se calculează folosind formulele:

$$\text{freq}(c_i) = \sum_t \text{freq}(c_i, t)$$
$$\text{freq}(t) = \sum_{c_i} \text{freq}(c_i, t)$$
$$\text{freq}(\text{total}) = \sum_{c_i, t} \text{freq}(c_i, t),$$

unde:

- $\text{freq}(c_i, t)$ este frecvența cuvîntului c_i, care este context, făcînd parte din contextul asociat cuvîntului-țintă t;
- $\text{freq}(\text{total})$ este numărul total de cuvinte;
- $\text{freq}(t)$ este frecvența cuvîntului-țintă t;
- $\text{freq}(c_i)$ este frecvența cuvîntului de context c_i.

Apoi, se calculează ponderea fiecărui termen de context cu formula:

$$v_i(t) = \frac{p(c_i \mid t)}{p(c_i)} = \frac{\text{freq}(c_i, t) \cdot \text{freq}(\text{total})}{\text{freq}(t) \cdot \text{freq}(c_i)}.$$

Se aleg primele 2000 cele mai frecvente cuvinte din corpus ca bază a spațiului semantic. Numărul ales s-a dovedit a fi relevant și în alte experimente.

2.2 Experimentul

Au fost alexe 8 poezii scrise în limba engleză modernă, de diverse grade de dificultate, adică cerînd diferite nivele de profunzime a analizei pentru a identifica sensul. Se mai folosesc și două texte de control: un articol din Wikipedia și un text generat aleatoriu, pentru a servi drept margine superioară, respectiv inferioară în ce privește structura frazeologică, semantică și coerență. În fine, s-au atribuit scoruri de dificultate de către autor și doi referenți independenți, unde scorul de 1 înseamnă foarte ușor de înțeles, iar scorul de 5 înseamnă foarte greu de înțeles.

Textele alese sînt prezentate în tabelul din figura 2.1. Mai multe detalii privitoare la autori și texte considerăm că sînt irelevante pentru scopul lucrării prezente.

Scorurile de dificultate au fost:

Poeziile au fost etichetate POS[1] automat, dar etichetarea a fost apoi verificată manual.

S-a calculat coerența pentru poeziile care au o structură frazeologică destul de clară (Brooke, Duffy, Slater, Stein, Wilde, random, Wikipedia), iar pentru celelalte, s-a efectuat o împărțire

[1] parts-of-speech

Autor	Titlu	An
Brooke	Day That I Loved	1911
Coolidge	Argument Over, Amounting	1990
Duffy	Valentine	1993
Ginsberg	Five A. M.	1996
MacCormack	At Issue III	2001
Slater	Ithaca, Winter	2008
Stein	If I Told Him,A Completed Portrait of Picasso	1924
Wilde	In the Gold Room	1881
Wikipedia	The Language Poets	?
Random	Psychologist. String	N/A

Figura 2.1: Poeziile alese în experimentul pentru coerență din [Herbelot, 2015]

Textul	Autorul	Referent 1	Referent 2	Media
Random	5	5	5	5
MacCormack	5	5	5	5
Coolidge	4	5	5	4.67
Ginsberg	5	4	3	4
Stein	5	3	3	3.67
Slater	2	3	4	3
Brooke	2	4	3	3
Wilde	1	1	2	1.33
Duffy	1	1	2	1.33
Wikipedia	1	1	1	

Figura 2.2: Scorurile de dificultate atribuite textelor alese în [Herbelot, 2015]

în diviziuni cu același numă de cuvinte. De asemenea, în conformitate cu ipotezele de lucru, s-au luat în considerare doar părțile relevante de propoziție care au apărut de cel puțin 50 de ori. În total, s-a folosit aproximativ 72% din conținutul textual, iar lungimea medie a frazelor a fost de 4 cuvinte relevante.

2.3 Rezultate și concluzii

Cel mai clar se pot vedea rezultatele pe graficul din figura 2.3.

Se poate vedea aici că textele poetice analizate se situează aproape la mijloc din punctul de vedere al coerenței între textul aleatoriu și cel de pe Wikipedia. Pentru a elimina anumite îndoieli, autoarea au mai adăugat cîteva texte aleatorii și cîteva de pe Wikipedia, graficul accentuîndu-se și mai bine, ca în figura 2.4.

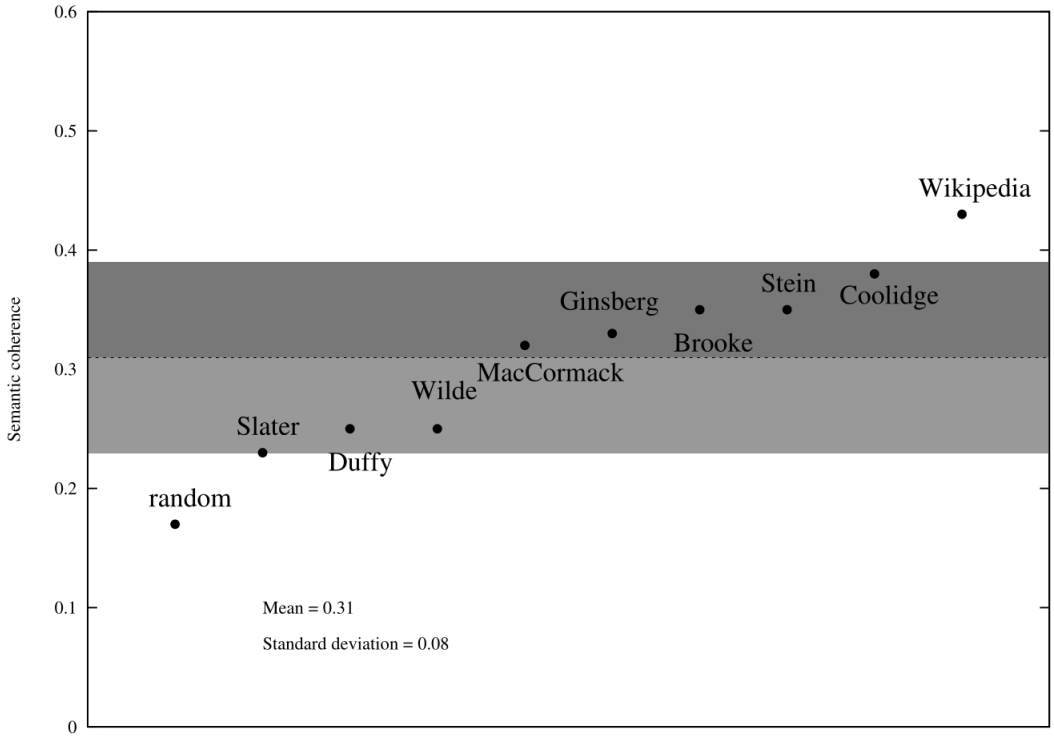

Figura 2.3: Coerența subiectelor din experimentul prezentat în [Herbelot, 2015]

Ideile de bază care se desprind din aceste rezultate sînt că, pe de o parte, poezia produsă de oameni se poate distinge destul de ușor de texte aleatorii, dar și că există o diferență semnificativă între textele științifice ori factuale (precum articolele de pe Wikipedia) și cele poetice. Totodată, coerența poeziilor este mai scăzută decît cea a textelor factuale, rezultat care se explică prin creativitatea autorului.

Mai trebuie remarcat și faptul că nu există nicio corelație între dificultatea textului, așa cum a fost ea percepută de oameni, și coerența care rezultă din calcule. De exemplu, poeziile lui Duffy și Wilde, considerate ”ușoare”, nu au un factor de coerență mare, lucru care este pus pe seama creativității autorilor. Rezultă, deci, că semantica nu lipsește cu precădere din textele complicate, așa cum afirmă unii critici.

Concluziile pe care le putem desprinde, împreună cu autoarea [Herbelot, 2015], sînt de felul următor. Folosind punctul de vedere distribuțional asupra sensului, este, într-adevăr, posibil să se pună în evidența relația între limbajul comun și cel „neobișnuit“, al poeziei. În același timp, tot modelul distribuțional arată clar distincția între texte umane și cele produse aleatoriu, indiferent de transparența sau dificultatea textului. Acest lucru se poate vedea foarte bine pe graficele care arată coerența, indicate mai sus.

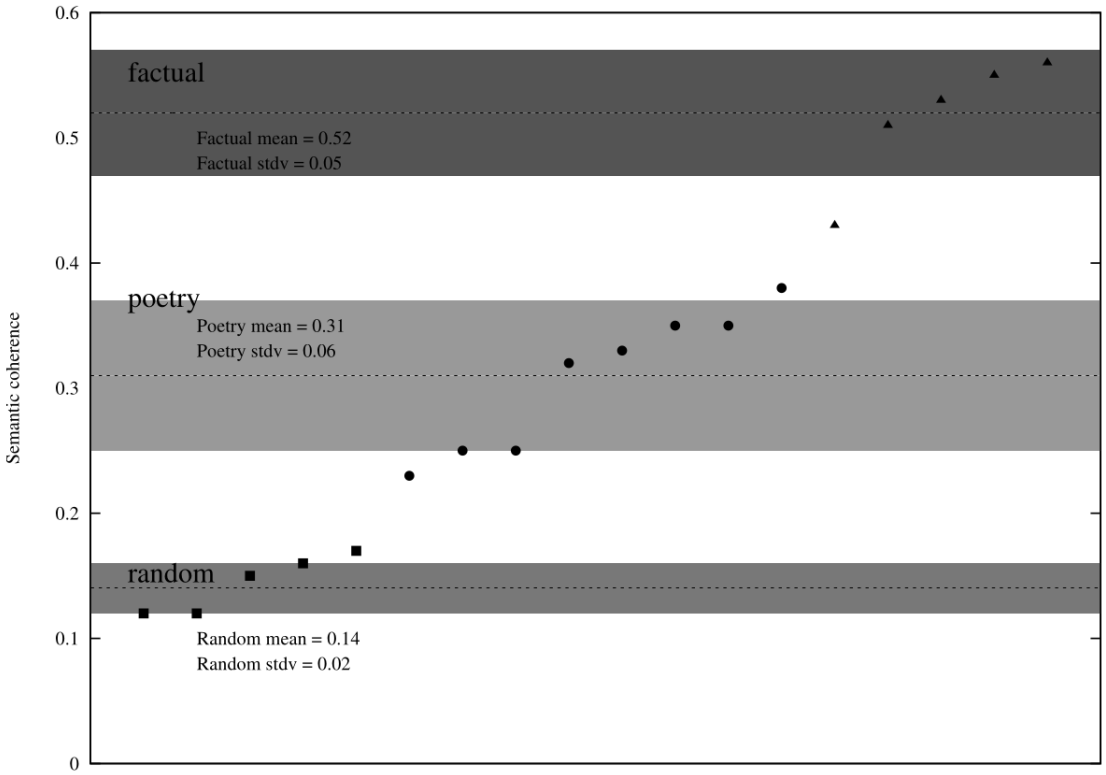

Figura 2.4: Coerența subiectelor din experimentul prezentat în [Herbelot, 2015], cu texte de control adăugate suplimentar

3 Aplicație: Legăturile între Lord Byron și Thomas Moore

3.1 Introducere și motivație

Ideea articolului [Aggarwal şi Tonra, 2014], din care preluăm aplicația de față, este să utilizeze un model de semantică distribuțională pentru a verifica dacă poetul Lord Byron a fost imitat de Thomas Moore, așa cum consideră unii critici sau, de fapt, este vorba despre o caracteristică generală a vocabularului poeziilor din secolul al XIX-lea (sau, mai precis, vocabularul specific sub-genului literar în care s-au încadrat cei doi), caracteristică pe care nu puteau să nu o aibă și poeziile lui Byron și Moore.

În perioada 1813-1817, poeții prieteni Lord Byron și Thomas Moore au scris o serie de poezii care sînt privite astăzi drept reprezentative pentru *orientalismul romantic*, o subcategorie a literaturii romantice, caracterizată prin temele și plasarea în contexte orientale și din Orientul Mijlociu. Printre poeziile publicate de cei doi, s-au descoperit foarte multe coincidențe, precum desfășurări similare de acțiuni, decoruri și nume de personaje similare.

Articolul [Aggarwal şi Tonra, 2014] își propune să folosească o metodă empirică nouă pentru a încerca să clarifice aceste coincidențe. De exemplu, ar putea răspunde la întrebarea dacă cumva orientalismul romantic este, în sine, un gen de poezie care folosește un vocabular limitat și atunci, similaritățile sînt inevitabile. Totodată, întrebarea mai generală poate fi și dacă se poate caracteriza un gen literar prin tocmai vocabularul său.

3.2 Metodologia și experimentul

Autorii au utilizat o metodă de semantică distribuțională prin *analiză semantică explicită* (ESA), ca alternativă la analiza latentă, prezentată în prima parte a lucrării. ESA a fost introdusă în [Gabrilovich şi Markovitch, 2007]. În varianta originală, metoda folosește texte din Wikipedia pentru a stabili unele sensuri, iar apoi, textul dat este interpretat prin intermediul conexiunilor deja stabilite cu ajutorul Wikipedia. Informal, este ca și cum clasificatorul semantic se „antrenează“ cu Wikipedia, pentru ca apoi să rezolve problema textului dat.

ESA permite reprezentarea unor vectori de sens într-un spațiu vectorial cu un număr foarte

mare de dimensiuni, dată fiind complexitatea și diversitatea articolelor din Wikipedia, iar coeficientul de relaționare semantică se calculează din nou folosind cosinusul unghiului celor doi vectori care reprezintă cuvintele analizate.

Experimentul a constat prin alegerea a 4 poezii lungi (poeme narative, în fapt, de cîteva mii de versuri) publicate de Byron între 1813 și 1814 și 4 poezii lungi publicate de Moore în 1817. Aceste poezii au fost împărțite în grupuri de cîte 227 versuri în cazul lui Byron și 246 de versuri în cazul lui Moore. După aceea, s-a calculat scorul ESA pentru grupurile din poeziile lui Byron și același lucru pentru operele lui Moore.

Însă abordarea a fost ceva mai complexă decît în modelul original ESA. S-au folosit, de fapt, 2 modele: unul care folosește Wikipedia, ca în articolul original, iar altul a folosit un corpus dat de 892 de poezii lungi (poeme narative) din secolele XVIII și XIX. Pentru ambele corpusuri s-a folosit ponderea tf-idf, pe care am introdus-o în primul capitol.

3.3 Rezultate si concluzii

Rezultatele, sortate după coeficienții de relaționare, au fost clasificate în „foarte legate“, „posibil legate“ și „nelegate“. În prima categorie au intrat aproximativ 1000 de perechi de versuri. S-au analizat (manual, uman) 15 perechi obținute cu metoda Wikipedia și 15 perechi obținute cu cealaltă metodă.

Drept concluzii, autorii remarcă faptul că, pe de o parte, este surprinzător că această metodă de analiză automată a găsit similarități exact în ce privește conceptele așteptate. Este vorba exact despre elementele remarcate de critici, specifice curentului literar: personaje și decoruri. Astfel, s-a constatat că perechile de versuri care au intrat în categoria „foarte legate“, atît în clasificarea folosind Wikipedia, cît și cu cealaltă metodă, sînt legate de sentimente și de cadre naturale. Ambele concepte caracterizează într-o măsură destul de mare curentul orientalismului romantic și se poate deduce de aici că, într-adevăr, inspirația dintre cei doi, dacă a existat, s-a manifestat exact unde se bănuiește.

Pe de altă parte, atît corpusul de texte analizate, cît și rezultatele sînt, din punct de vedere cantitativ, nu foarte reprezentative. Pentru o concluzie mai precisă, idei de îmbunătățire pot privi atît rafinarea modelelor ESA, cît și îmbogățirea corpusului de texte studiate.

Bibliografie

[Aggarwal şi Tonra, 2014] Aggarwal, N. şi Tonra, J. (2014). Using distributional semantics to trace influence and imitation in romantic orientalist poetry. In Akbik, A. şi Visengeriyeva, L., editors, *Proceedings of the AHA Information Discovery Workshop*. ACL.

[Boleda, 2019] Boleda, G. (2019). Distributional semantics and linguistic theory. *CoRR*, abs/1905.01896.

[Boleda şi Herbelot, 2016] Boleda, G. şi Herbelot, A. (2016). Formal distributional semantics: Introduction to the special issue. *Computational Linguistics*, 42(4):619–635.

[Bruns, 2005] Bruns, G. (2005). *The Material of Poetry: Sketches for a Philosophical Poetics*. Number v. 1 in Georgia Southern University. Jack N. and Addie D. Averitt lecture series. University of Georgia Press.

[Erk, 2012] Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

[Gabrilovich şi Markovitch, 2007] Gabrilovich, E. şi Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In Veloso, M. M., editor, *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pages 1606–1611.

[Herbelot, 2015] Herbelot, A. (2015). The semantics of poetry: A distributional reading. *Digit. Scholarsh. Humanit.*, 30(4):516–531.

[Lenci, 2018] Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, 4(1):151–171.

[Masterman, 1971] Masterman, M. (1971). Computerized haiku. In Reichardt, J., editor, *Cybernetics, art and ideas*, pages 175–184. Studio Vista.

[Newman et al., 2010] Newman, D., Lau, J. H., Grieser, K., şi Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108, Los Angeles, California. Association for Computational Linguistics.

[Turney şi Pantel, 2010] Turney, P. D. şi Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res.*, 37:141–188.