Semantică distribuțională pentru analiza poeziilor

Adrian Manea, 510 SLA

Ideea

Lucrarea prezintă aspecte introductive de *semantică distribuțională* (SD), aplicate în studiul poeziilor scrise în limba engleză.

După o introducere teoretică privitoare la subiectul și unele dintre metodele utilizate în SD, prezentăm ca aplicație studiul coerenței subiectelor în poezia engleză modernă și contemporană.

Subiectul și metodele SD

Această ramură a lingvisticii se bazează e *ipoteza distribuțională*, formulată în anii '50, care afirmă că similaritatea semantică rezultă în similarități ale unor distribuții statistice. Abordarea este, însă, *invers*: se pornește de la corpusuri de text și se induc reprezentări semantice pe baza distribuțiilor observate.

Intuitiv: „Sensul unui cuvînt este utilizarea lui în limbaj" (L. Wittgenstein, 1922). În multe situații, putem ghici sensul unui cuvînt din contextele în care acesta este folosit.

Modelul matematic: spațiile vectoriale, numite în acest caz special, *spații semantice*.

Metoda, într-o prezentare simplificată, se bazează pe ambiguitatea termenului de „context". Astfel, se stabilește un cuvînt-cheie căruia se caută semantica și o fereastră de conținut, care în varianta cea mai simplă înseamnă un interval centrat în cuvîntul-cheie, de o lungime prestabilită, preluat din textul studiat. De obicei, se aplică rafinări suplimentare contextului, e.g. ignorarea prepozițiilor, a punctuației sau chiar concentrarea pe părțile de vorbire asociate, de obicei, cuvîntului-cheie (e.g. dacă se studiază un substantiv, se poate acorda atenție suplimentară adjectivelor și verbelor).

Se parcurg toate ferestrele de conținut de această formă și se înregistrează cele mai numeroase co-ocurențe de cuvinte, asociate cuvîntului-cheie. Rezultatele se stochează într-un vector asociat cuvîntului respectiv. Exemplu: dacă corpusul studiat este *Encyclopedia Britannica*, iar cuvîntul-țintă este *coleopter*, putem obține un vector de co-ocurențe de forma:

$$v = \begin{pmatrix} 5221 & 4431 & 3111 & 2104 & 987 & 511 & 251 & 41 \end{pmatrix},$$

asociate cuvintelor, respectiv, *mare, aripă, larvă, copac, floare, cîmpie, cald, roșu.* Putem decide să ignorăm primul cuvînt, deoarece este unul comun, care apare în preajma prea multor cuvinte pentru a fi decisiv în ce privește sensul. La fel, ultimul cuvînt, care nu doar că ar putea fi irelevant pentru semantica cuvîntului-țintă, dar are și un număr de apariții comparativ mult mai mic față de celelalte. Însă celelalte sînt relevante și ne dau de înțeles că ar putea fi vorba despre o insectă.

În exemplul de mai sus, spunem că folosim un spațiu vectorial cu un număr de dimensiuni egal cu numărul de co-ocurențe asociate cuvîntului-cheie. Putem fie să folosim toate cele 8 dimensiuni în cazul de mai sus, sau să păstrăm doar 6, după o filtrare a celor relevante.

Ulterior, se pot aplica filtrări suplimentare, care pot avea diverse roluri care să facă mai relevante cifrele calculate. În plus, una dintre cele mai importante transformări ulterioare este calculul *coeficientului de similaritate* între doi vectori corespunzători la două cuvinte. Acesta se calculează ca fiind cosinusul unghiului dintre vectorii asociați celor două cuvinte, folosind produsul scalar euclidian.

Aplicație: Coerența în poezia modernă și contemporană

Punctul de pornire a acestei aplicații este critica adusă poeziei, conform căreia acest tip de scriere nu are o semantică, întrucît cuvintele sînt folosite cu sensuri mult diferite față de sensurile de dicționar. De aceea, dacă se încearcă analiza *coerenței subiectelor*, adică extragerea cuvintelor care arată despre ce este vorba în poezia analizată, mult prea rar se

constată o coerență, adică se găsesc cuvinte care aparțin aceleiași familii lexicale. De exemplu, o mulțime de subiecte de forma { scaun, masă, birou, echipă } are o coerență mai mare decît { scaun, rece, elefant, nor }.

Se definește *coerența unei mulțimi de cuvinte* $W = \{w_i \mid 1 \le i \le n\}$ prin:

$$\overline{S}(W) = \overline{\{\mathrm{Sim}(w_i, w_j) \mid 1 \le i < j \le n\}},$$

unde bara superioară notează media, iar Sim reprezintă similaritatea, calculată cu ajutorul cosinusului:

$$\mathrm{Sim}(A, B) = \frac{\sum A_i \cdot B_i}{\sqrt{\sum A_i^2 \cdot B_i^2}},$$

iar mărimea ferestrei de conținut (mai precis, dimensiunea spațiului semantic) este $n = 2000$, număr care s-a dovedit relevant și în alte studii.

În experiment, au fost alese 8 texte, dintre care 6 sînt poezii moderne și contemporane, iar 2 au fost de control: un text generat aleatoriu și un text de pe Wikipedia. Textelor s-au atribuit scoruri de dificultate a înțelegerii, iar rezultatul studiului coerenței este cel din figura 1.

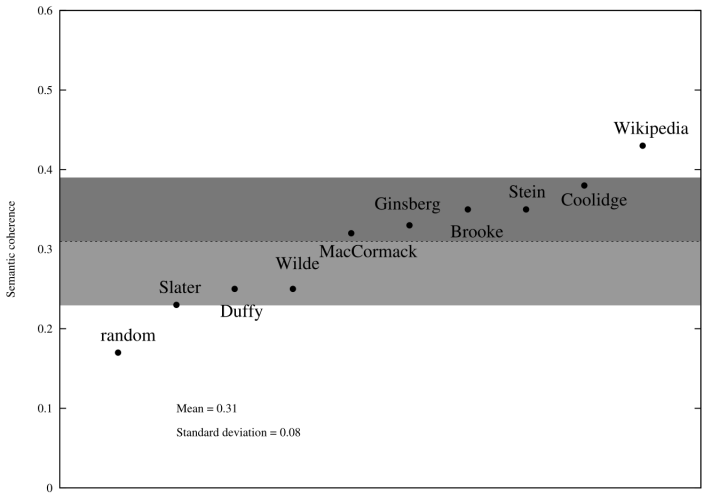

Figure 1: Coerența subiectelor din poezii moderne și contemporane

Concluziile studiului au fost că nu se evidențiază discrepanța pe care unii critici au remarcat-o între coerența subiectelor din limbajul comun și cea a limbajului poetic. În plus, putem constata din grafic faptul că poeziile se situează undeva la mijloc între un text aleatoriu (deloc coerent) și unul științific (cu coerența maximă). Și mai mult decît atît, dificultatea de înțelegere a textului nu este legată de coerența subiectelor. Astfel, textul lui MacCormarck a fost apreciat drept cel mai greu de înțeles (după cel generat aleatoriu), dar observăm că el are un scor mediu de coerență. Pe de altă parte, textul lui Stein a fost apreciat ca fiind mediu de înțeles, iar coerența sa este foarte mare.

Pentru o și mai bună claritate, s-au adăugat texte suplimentare de control, unele generate aleatoriu și unele factuale (științifice). Tendința de a se situa la mijloc se vede mai accentuat, conform figurii 2.

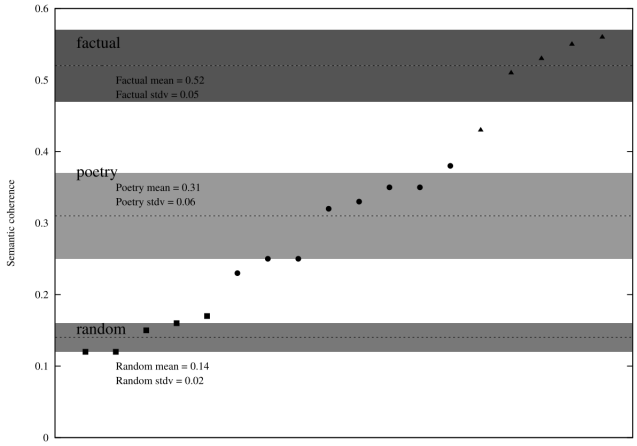

Figure 2: Coerența poeziilor, cu texte de control suplimentare

Bibliografie selectivă

(1) Erk, K. – *Vector space models of word meaning and phrase meaning: A survey*, Language and Linguistics Compass, 2012;

(2) Herbelot, A. – *The Semantics of Poetry: A Distributional Reading*, Digit. Scholarsh. Humanit., 2015;

(3) Turney, P., Pantel, P. – *From frequency to meaning: Vector space models of semantics*, J. Artif. Intell. Res., 2010.