Semantică distribuțională pentru studiul poeziilor în engleză

Adrian Manea

510, SLA

Scopul și metoda

„Sensul unui cuvînt este utilizarea lui în limbaj“ (L. Wittgenstein, 1922)

Scopul și metoda

„Sensul unui cuvînt este utilizarea lui în limbaj“ (L. Wittgenstein, 1922)

Ideea: Putem ghici sensul cuvintelor din context.

Scopul și metoda

„Sensul unui cuvînt este utilizarea lui în limbaj“ (L. Wittgenstein, 1922)

Ideea: Putem ghici sensul cuvintelor din context.

Co-ocurența repetată generează o distribuție relevantă pentru semantică

Modelul matematic

Obiectul central: spațiile vectoriale (semantice)

Modelul matematic

Obiectul central: spațiile vectoriale (semantice)

Dimensiunea = fereastra de context relevantă, centrată în cuvîntul-cheie (cc)

Modelul matematic

Obiectul central: spațiile vectoriale (semantice)

Dimensiunea = fereastra de context relevantă, centrată în cuvîntul-cheie (cc)

Filtre ulterioare: ignoră prepozițiile și punctuația, ignoră cuvintele prea generale (e.g. *foarte, mare, mic*), păstrează doar părțile de vorbire relevante (e.g. cc = substantiv $\Rightarrow$ adjectiv, verb, ~~adverb~~)

Modelul matematic

Obiectul central: spațiile vectoriale (semantice)

Dimensiunea = fereastra de context relevantă, centrată în cuvîntul-cheie (cc)

Filtre ulterioare: ignoră prepozițiile și punctuația, ignoră cuvintele prea generale (e.g. *foarte, mare, mic*), păstrează doar părțile de vorbire relevante (e.g. cc = substantiv $\Rightarrow$ adjectiv, verb, ~~adverb~~)

Prelucrări matematice: scor de similaritate între cc (cosinus euclidian), SVD pentru reducerea dimensiunii spațiului semantic

Problema semanticii

Frege (1892), Tarski (1944): semantică formală, bazată pe mulțimi

Problema semanticii

Frege (1892), Tarski (1944): semantică formală, bazată pe mulțimi

Wittgenstein (1922): „Meaning is use“

Problema semanticii

Frege (1892), Tarski (1944): semantică formală, bazată pe mulțimi

Wittgenstein (1922): „Meaning is use“

Wheelwright vs. Miles cca. 1940: semantica poeziilor

Problema semanticii

Frege (1892), Tarski (1944): semantică formală, bazată pe mulțimi

Wittgenstein (1922): „Meaning is use“

Wheelwright vs. Miles cca. 1940: semantica poeziilor

- “Music is the exquisite knocking of the blood” (R. Brooke)

Problema semanticii

Frege (1892), Tarski (1944): semantică formală, bazată pe mulțimi

Wittgenstein (1922): „Meaning is use“

Wheelwright vs. Miles cca. 1940: semantica poeziilor

- “Music is the exquisite knocking of the blood” (R. Brooke)
- “Your huge mortgage of hope” (T. Hughes)

Problema semanticii

Frege (1892), Tarski (1944): semantică formală, bazată pe mulțimi

Wittgenstein (1922): „Meaning is use“

Wheelwright vs. Miles cca. 1940: semantica poeziilor

- “Music is the exquisite knocking of the blood” (R. Brooke)
- “Your huge mortgage of hope” (T. Hughes)
- “Skelleton bells of trees” (A. Slater)

Problema semanticii

Are poezia semantică?

Problema semanticii

Are poezia semantică?

Dacă da, este ea consonantă cu cea a limbajului comun?

Problema semanticii

Are poezia semantică?

Dacă da, este ea consonantă cu cea a limbajului comun?

Dacă nu, de ce nu?

Problema semanticii

Are poezia semantică?

Dacă da, este ea consonantă cu cea a limbajului comun?

Dacă nu, de ce nu?

Bruns (2005): *poezia este alcătuită din limbaj, dar nu este o utilizare a acestuia.*

Problema semanticii

Are poezia semantică?

Dacă da, este ea consonantă cu cea a limbajului comun?

Dacă nu, de ce nu?

Bruns (2005): *poezia este alcătuită din limbaj, dar nu este o utilizare a acestuia.*

Miles: Fără cunoașterea limbajului comun, poeziile nu au sens.

Aplicație: Coerența semantică în poezia modernă

Teza: poeziile folosesc o structură similară a limbii cu limbajul comun.

Aplicație: Coerența semantică în poezia modernă

Teza: poeziile folosesc o structură similară a limbii cu limbajul comun.

Topic coherence: { scaun, masă, birou } > { scaun, elefant, nor }

Aplicație: Coerența semantică în poezia modernă

Teza: poeziile folosesc o structură similară a limbii cu limbajul comun.

Topic coherence: { scaun, masă, birou } > { scaun, elefant, nor }

$W = \{w_1, \ldots, w_n\}$ cuvinte, $n = 2000$ (relevant).

Aplicație: Coerența semantică în poezia modernă

Teza: poeziile folosesc o structură similară a limbii cu limbajul comun.

Topic coherence: { scaun, masă, birou } > { scaun, elefant, nor }

$W = \{w_1, \ldots, w_n\}$ cuvinte, $n = 2000$ (relevant).

Similaritatea(W) = $\text{avg}\{\text{Sim}(w_i, w_j) \mid 1 \leq i < j \neq n\}$

Aplicație: Coerența semantică în poezia modernă

Teza: poeziile folosesc o structură similară a limbii cu limbajul comun.

Topic coherence: { scaun, masă, birou } > { scaun, elefant, nor }

$W = \{w_1, \ldots, w_n\}$ cuvinte, $n = 2000$ (relevant).

Similaritatea(W) = avg$\{\text{Sim}(w_i, w_j) \mid 1 \leq i < j \neq n\}$

$$\text{Sim}(w_i, w_j) = \cos(w_i, w_j) = \frac{\sum_k w_i^k \cdot w_j^k}{\sqrt{\sum_k (w_i^k)^2 \cdot (w_j^k)^2}}$$

Aplicație: Coerența semantică în poezia modernă

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

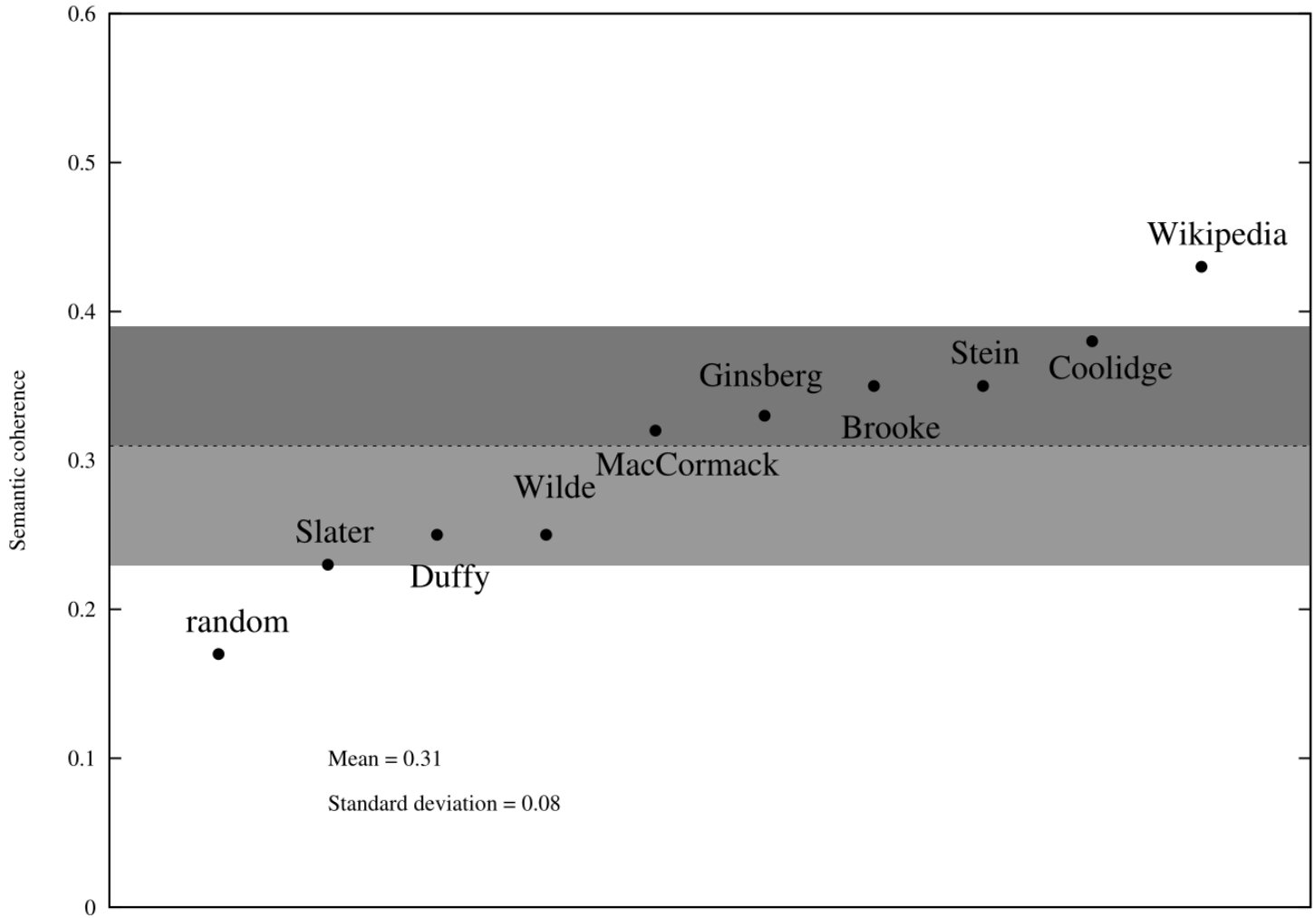
Ilustrație: Poeziile alese în experimentul pentru coerență din [Herbelot, 2015]

Aplicație: Coerența semantică în poezia modernă

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	1

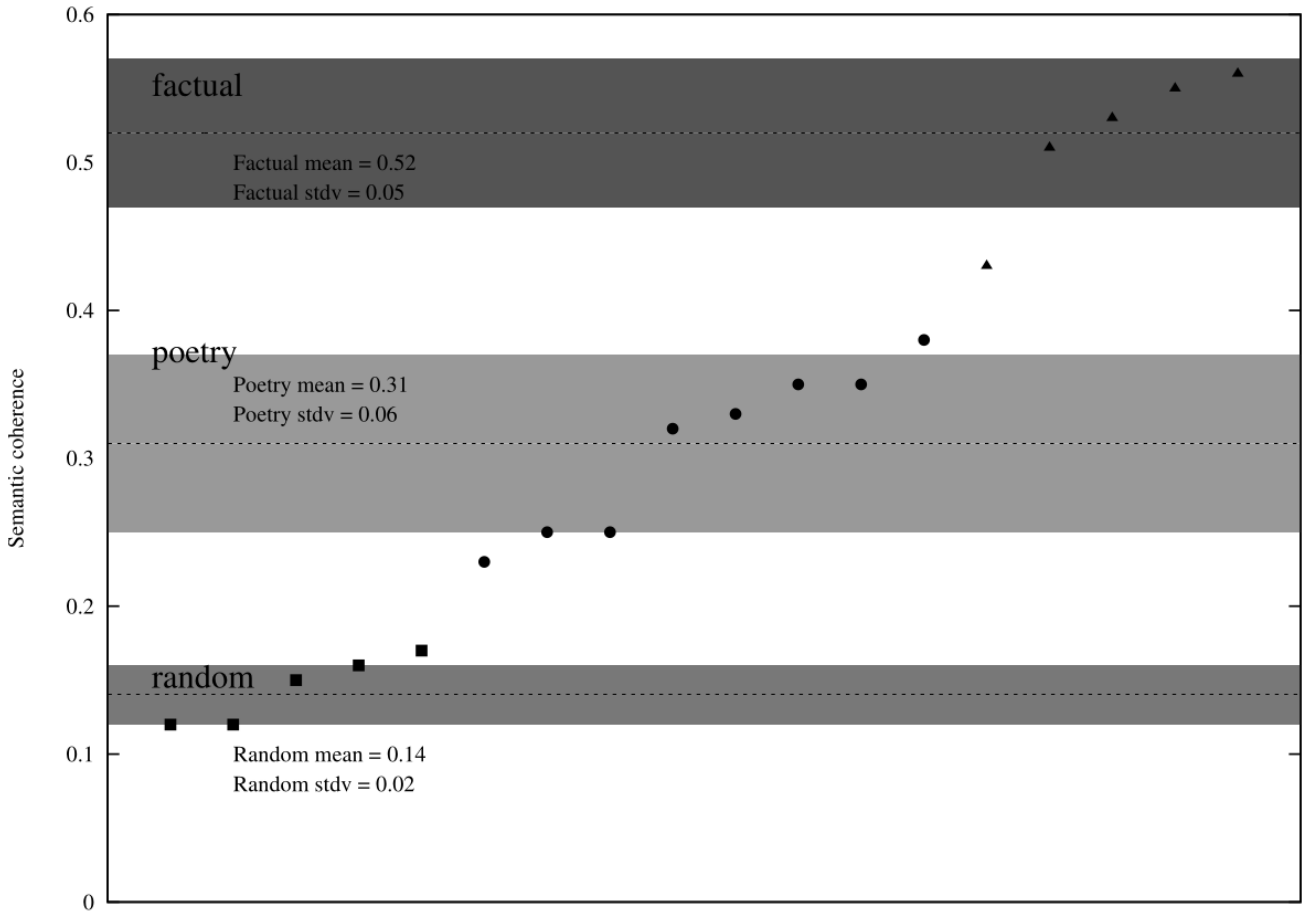
Ilustrație: Scorurile de dificultate atribuite textelor alese în [Herbelot, 2015]

Aplicație: Coerența semantică în poezia modernă



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

Aplicație: Coerența semantică în poezia modernă



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

Aplicație: Coerența semantică în poezia modernă

Concluzii:

;

Aplicație: Coerența semantică în poezia modernă

Concluzii:

- Folosind SD, se poate vedea o relație între limbajul „obișnuit“ și cel „neobișnuit“ (al poeziei);

Aplicație: Coerența semantică în poezia modernă

Concluzii:

- Folosind SD, se poate vedea o relație între limbajul „obișnuit“ și cel „neobișnuit“ (al poeziei);
- Distincție clară între texte umane vs. generate aleatoriu

Aplicație: Coerența semantică în poezia modernă

Concluzii:

- Folosind SD, se poate vedea o relație între limbajul „obișnuit“ și cel „neobișnuit“ (al poeziei);
- Distincție clară între texte umane vs. generate aleatoriu
- Coerența poeziilor este *între* texte aleatorii și texte științifice

Aplicație: Coerența semantică în poezia modernă

Concluzii:

- Folosind SD, se poate vedea o relație între limbajul „obișnuit“ și cel „neobișnuit“ (al poeziei);
- Distincție clară între texte umane vs. generate aleatoriu
- Coerența poeziilor este *între* texte aleatorii și texte științifice
- Coerența poeziilor nu (prea) depinde de dificultatea textului

Aplicație: Lord Byron vs. Thomas Moore 1813-1817

Teza: Moore s-a inspirat de la Byron în poeziile din curentul *orientalismului romantic*

Aplicație: Lord Byron vs. Thomas Moore 1813-1817

Teza: Moore s-a inspirat de la Byron în poeziile din curentul *orientalismului romantic*

Anti-teza: Genul literar are un vocabular și idei specifice, limitate $\Rightarrow$ similaritatea este inevitabilă

Aplicație: Lord Byron vs. Thomas Moore 1813-1817

Metoda: ESA (*analiză semantică explicită*)

Aplicație: Lord Byron vs. Thomas Moore 1813-1817

Metoda: ESA (*analiză semantică explicită*)

Ideea: „Se antrenează“ semantic pe Wikipedia, apoi analizează textele date

Aplicație: Lord Byron vs. Thomas Moore 1813-1817

Metoda: ESA (*analiză semantică explicită*)

Ideea: „Se antrenează“ semantic pe Wikipedia, apoi analizează textele date

Experimentul: cîte 4 poeme narative (mii de versuri), împărțite în grupuri de cîte $\sim$ 200 versuri fiecare și se calculează scorul ESA

Aplicație: Lord Byron vs. Thomas Moore 1813-1817

Metoda: ESA (*analiză semantică explicită*)

Ideea: „Se antrenează“ semantic pe Wikipedia, apoi analizează textele date

Experimentul: cîte 4 poeme narative (mii de versuri), împărțite în grupuri de cîte $\sim$ 200 versuri fiecare și se calculează scorul ESA

Model suplimentar: „antrenare“ pe 892 poeme narative

Aplicație: Lord Byron vs. Thomas Moore 1813-1817

Concluzii:

Aplicație: Lord Byron vs. Thomas Moore 1813-1817

Concluzii:

- Aproximativ 1000 perechi de versuri „foarte legate“

Aplicație: Lord Byron vs. Thomas Moore 1813-1817

Concluzii:

- Aproximativ 1000 perechi de versuri „foarte legate“
- S-au analizat uman 15 perechi cu metoda Wikipedia și 15 perechi din modelul suplimentar

Aplicație: Lord Byron vs. Thomas Moore 1813-1817

Concluzii:

- Aproximativ 1000 perechi de versuri „foarte legate“
- S-au analizat uman 15 perechi cu metoda Wikipedia și 15 perechi din modelul suplimentar
- Similaritățile sînt exact în zonele relevate de critici literari: personaje, sentimente și decoruri

Aplicație: Lord Byron vs. Thomas Moore 1813-1817

Concluzii:

- Aproximativ 1000 perechi de versuri „foarte legate“
- S-au analizat uman 15 perechi cu metoda Wikipedia și 15 perechi din modelul suplimentar
- Similaritățile sînt exact în zonele relevate de critici literari: personaje, sentimente și decoruri
- *Dacă* a existat, inspirația dintre cei doi s-a manifestat unde s-a preconizat, DAR...

Aplicație: Lord Byron vs. Thomas Moore 1813-1817

Concluzii:

- Aproximativ 1000 perechi de versuri „foarte legate“
- S-au analizat uman 15 perechi cu metoda Wikipedia și 15 perechi din modelul suplimentar
- Similaritățile sînt exact în zonele relevate de critici literari: personaje, sentimente și decoruri
- *Dacă* a existat, inspirația dintre cei doi s-a manifestat unde s-a preconizat, DAR...
- Corpusul și textele nu sînt reprezentative; autorii recomandă rafinarea modelelor ESA și a corpusului înainte de concluzii clare

Aplicație: Lord Byron vs. Thomas Moore 1813-1817

Concluzii:

- Aproximativ 1000 perechi de versuri „foarte legate“
- S-au analizat uman 15 perechi cu metoda Wikipedia și 15 perechi din modelul suplimentar
- Similaritățile sînt exact în zonele relevate de critici literari: personaje, sentimente și decoruri
- *Dacă* a existat, inspirația dintre cei doi s-a manifestat unde s-a preconizat, DAR...
- Corpusul și textele nu sînt reprezentative; autorii recomandă rafinarea modelelor ESA și a corpusului înainte de concluzii clare
- *Posibil* ca genul în sine să limiteze, să impună clișee

Bibliografie și lecturi suplimentare

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

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

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

Bibliografie și lecturi suplimentare II


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


Gabrilovich, E. and 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, A. (2015).
The semantics of poetry: A distributional reading.
Digit. Scholarsh. Humanit., 30(4):516–531.


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


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


Newman, D., Lau, J. H., Grieser, K., and 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, P. D. and Pantel, P. (2010).
From frequency to meaning: Vector space models of semantics.
J. Artif. Intell. Res., 37:141–188.