Semantic Spaces

Matilde Marcolli

MAT1509HS: Mathematical and Computational Linguistics

University of Toronto, Winter 2019, T 4-6 and W 4, BA6180

Reference

- Yuri I. Manin, Matilde Marcolli, *Semantic spaces*, Math. Comput. Sci. Vol.10 (2016) N.4, 459–477.

Lexemes and semic axes

• P. Guiraud. *The semic matrices of meaning.* Social Science Information, 7(2), 1968, pp. 131–139.

- core "meanings" assigned not to "words" but to "lexemes"
- lexeme as equivalence class of all word forms that differ only by inflectional significations
- example: lexeme $\text{TAKE}_{(V)}$ includes lexical items: *take, takes, took, taking, . . . , have taken, has taken, . . . , have been taken,. . .*
- tag *(V)* meaning word "take" understood as *verb* not *noun*
- encoding of meaning: "sense" of lexemes
- list of "semes" such as *animate, inanimate, actor, process* etc.
- encoding of meaning of lexemes is specified by listing a subset of semes

Geometric picture

- N semes represented by basis vectors e_i, $i = 1, \ldots, N$ of $\mathbb{R}^N$
- meanings represented by subsets of vertices of the unit cube $[0, 1]^N$
- "bisemic" description: subsets of vertices of $[-1, 1]^N$ with sign for complementarity relations like animate/inanimate etc.
- allow for points in $\mathbb{R}^N$ not localized at vertices of unit cube to describe certain associations and combinations of meanings

• elaboration of this idea of geometric space of meanings in

- P. Gärdenfors. *Geometry of Meaning: Semantics Based on Conceptual Spaces.* Cambridge, Mass. MIT Press, 2014

Vector Space Models of Semantics

- VSM takes a large corpus of natural language texts and produces a matrix of frequences
- intermediate steps: (i) linguistic and (ii) statistical
- linguistic: creation of the relevant *vocabulary of lexemes*
- statistical: matching of text words to lexemes and normalized counting of occurrences
- matrix entries characterise correlations between lexemes and text/contexts

- D vocabulary with $M = \#D$ number of lexemes
- text T with a set of subtexts *contexts*: $C(T) = \{c_1, \dots, c_N\}$ (sentences, paragraphs, or windows of certain length around each word)

- two settings:

1. *large vocabulary case*: size of vocabulary of lexemes large compared to number of contexts in the texts: goal selecting from large dictionary words that best represent the given contexts semantically
2. *information retrieval case*: vocabulary small compared to number of contexts (e.g. words used in a query) goal selecting among contexts in a given corpus best match semantically for query words

Matrix of frequencies

- $N \times M$ matrix of frequencies $P = P(T)$ with M lexemes (words) N contexts
- entries p_{ij} estimated probability (frequency) of occurrence of word $w_i \in D$ in context $c_j \in C(T)$
- matrix $X = X(T)$ with entries $X = (x_{ij})$

$$x_{ij} = \max\{0, \log\left(\frac{p_{ij}}{p_{i\star}p_{\star j}}\right)\}$$

- with $p_{i\star} = \sum_j p_{ij}$ estimated probability of the word $w_i \in D$ and $p_{\star j} = \sum_i p_{ij}$ estimated probability of the context $c_j \in C(T)$
- condition $p_{ij} = p_{i\star}p_{\star j}$ statistical independence of word w_i and context c_j
- condition $p_{ij} > p_{i\star}p_{\star j}$ signals presence of a semantic relation between them

- case of overlapping contexts (windows around words): if a word in the intersection of two adjacent contexts j and $j+1$ it affects the counting in both p_{ij} and $p_{i,j+1}$
- example of overlapping contexts: Shannon 3-gram model

Large Vocabulary case $\quad N \leq M$

- *Statistical Semantics Hypothesis*: thing that occur together signify together
- statistical patterns of word usage in texts determine their semantical meaning
- (parts of) text that have similar vectors in the frequency matrices also have similar meanings
- $r = rank(P)$ be the rank of the matrix $P(T)$ ($r \leq N$ for large dictionary) measures largest number of words and contexts that the text T can disambiguate semantically
- linear dependence of frequency vectors interpreted as revealing underlying semantic relations

Semantics on Grassmannians

- when $r = N$ the matrix $P(T)$ of text T determines a point $p(T)$ in the real Grassmannian $Gr(N, M)$ of N-planes in $\mathbb{R}^M$
- similarly for $rank(X(T)) = N$ the matrix $X(T)$ determines a point $x(T) \in Gr(N, M)$

Matroids

- finite set $E = \{1, \dots, M\}$ and set $\mathcal{I} \subseteq 2^E$ subsets of E with
  1. $\emptyset \in \mathcal{I}$
  2. if $T \in \mathcal{I}$ and $S \subseteq T$ then $S \in \mathcal{I}$
  3. if $S, T \in \mathcal{I}$ and $\#T > \#S$ then there is $t \in T \smallsetminus S$ such that $S \cup \{t\} \in \mathcal{I}$
- matroid structure describes linear independence relations

- Example of matroid

column vectors of the complex (or real) matrix

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}.$$

Describe this matroid in terms of subsets of [5].

Solution: The matroid consists of $\{1,2,4\}$, $\{1,2,5\}$, $\{1,3,4\}$, $\{1,3,5\}$, $\{1,4,5\}$, $\{2,3,4\}$, $\{2,3,5\}$ and $\{2,4,5\}$ and all of their subsets.

Matroid strata in the Grassmannian

- a point P in $Gr(N,M)$ determines a matroid $\mathcal{M}_P$ with bases the N subsets $\{1,\ldots,M\}$ with nonzero determinant of corresponding minor
- given a matrod $\mathcal{M}$ the matroid stratum $\mathcal{S}_{\mathcal{M}} \subset Gr(N,M)$ is $\mathcal{S}_{\mathcal{M}} = \{P \,|\, \mathcal{M}_P = \mathcal{M}\}$
- $\mathcal{M} = \mathcal{M}(T)$ set of subsets of $\{1,\ldots,M\}$ of cardinality N, such that determinant of corresponding minor is $\Delta_I(P(T)) \neq 0$
- $\mathcal{M}$ determines a matroid stratum $\mathcal{S}_{\mathcal{M}} \subset Gr(N,M)$ with $P(T) \in \mathcal{M}$

- order of appearance of words in text T relevant to semantic interpretation: can order $D(T)$ listing words in order of appearance in T
- set of subsets $I = \{i_1, \ldots, i_N\}$ of $[M] := \{1, \ldots, M\}$ with $i_1 < i_2 < \ldots < i_N$
- correspond to choices of words $w_{i_1}, \ldots, w_{i_N}$ in $D(T)$, such that order of appearance in text T respected
- frequency vectors $P_{i_k} := (p_{i_k,j})_j$ occurrence of word w_{i_k} in context c_j
- Gale ordering;
  - two subsets $I = \{i_1, \ldots, i_N\}$ and $J = \{j_1, \ldots, j_N\}$ as above with $i_1 < i_2 < \cdots < i_N$ and $j_1 < j_2 < \cdots < j_N$
  - $I \leq_G J$ iff $i_1 \leq j_1$, $i_2 \leq j_2$, ..., $i_N \leq j_N$
  - Gale ordering corresponds to relative position of words w_{i_k} and w_{j_k} in dictionary $D(T)$ according to first appearance in T
- original dictionary also has an ordering different from order of appearance in text T
- some permutation $\sigma \in S_M$, such that Gale ordering is $I \leq_\sigma J$

$$\sigma^{-1} I \leq_G \sigma^{-1} J$$

Matroid strata and semantic relations

- if for subset I minor of the matrix $P(T)$ has vanishing determinant $\Delta_I(P(T)) = 0$ there is a linear dependence between vectors P_{i_k} hence (Statistical Semantics Hypothesis) a semantic relation between the w_{i_k}
- matroid stratum $\mathcal{S}_{\mathcal{M}} \subset Gr(N, M)$ containing point $p(T) \in \mathcal{S}_{\mathcal{M}}$ describes all choices of words w_{i_k}, $k = 1, \dots, N$ in dictionary for which the semantic vectors P_{i_k} are independent
- maximal amount of semantic information that can be extracted from the text and its contexts

Positroid Cell

- positive (or totally non-negative) Grassmannian $Gr_{\geq 0}(N, M)$ is the subset $Gr_{\geq 0}(N, M) \subset Gr(N, M)$ of matrices A such that all $\Delta_I(A) \geq 0$ for I as above
- positroid cell $\mathcal{S}_{\mathcal{M}}^{\geq 0}$: all $\Delta_I(A) > 0$, for all $I \in \mathcal{M}$
- $p(T)$ lies in the positroid cell $\mathcal{S}_{\mathcal{M}}^{\geq 0}$ iff there is a continuous paths γ_I, for each $I \in \mathcal{M}$, where
  - $\gamma_I(0) = P(T)$ and
  - $\gamma_I(1)$ matrix with I-minor the identity
  - for all $t \in [0, 1]$ have $\gamma_I(t) \in \mathcal{S}_{\mathcal{M}}^{\geq 0}$
- this condition expresses fact that choice of words $w_{i_1}, \dots, w_{i_N}$ for contexts $c_1, \dots, c_N$ of the text T contains maximal amount of semantic information
- case with minor equal identity corresponds to word w_{i_k} entirely specified semantically by context c_k no contribution from c_j with $j \neq k$

Information Retrieval case $N \geq M$

- list of words in a query, locate texts or contexts semantically most relevant to that query
- setting similar as before with matrix $P(T)$ point in Grassmannian $Gr(M, N)$
- minors $I = \{i_1, \ldots, i_M\}$ correspond to choices of contexts c_{i_k} in T in response to query by the words w_k
- condition $\Delta_I(P(T)) > 0$: assignments of a context to each word of the query that best match it semantically

Paths in Projective Spaces

- text subdivided into contexts: collection of points, path in a projective space, rather than single point in a Grassmannian
- fixed large vocabulary D of lexemes, $M = \#D$
- subdivide the text into contexts c_k but number N of contexts need not be smaller than M
- semantic vectors $P_k(T) = (p_{ik})_{i \in D}$ of probability (frequency) of word $w_i \in D$ in context c_k of T
- each P_k determines a point p_k in the projective space $\mathbb{P}^{M-1} \simeq Gr(1, M)$ (normalization so up to scale)
- a text T corresponds to an ordered N-tuple of points in $\mathbb{P}^{M-1}$
- an oriented path $\Gamma(T)$ by drawing geodesic arcs between consecutive points

Semantic comparison via path distance

- for different texts T and T' comparison of semantics of contexts by distance between paths $\Gamma(T)$ and $\Gamma(T')$ in ambient $\mathbb{P}^{M-1}$
- $d_{FS}(x, y)$ Fubini-Study metric on $\mathbb{P}^{M-1}$
- Fréchet distance between the two polygonal curves

$$\delta(\Gamma(T), \Gamma(T')) = \inf_{\gamma,\gamma'} \max_{t\in[0,1]} d_{FS}(\gamma(t), \gamma'(t)),$$

  with $\gamma : [0,1] \to \Gamma(T)$ and $\gamma' : [0,1] \to \Gamma(T')$ parameterizations by $[0,1]$
- infimum over reparameterizations by $[0,1]$ of maximum over $t \in [0,1]$ of distance between corresponding points
- Fréchet distance for polygonal curves computable in polynomial time

Flag Varieties

- to keep track of semantic interpretation changes when more contexts are considered in the linear ordering of a text
- $P_k(T) = (p_{ik})_{i\in D}$ semantic vectors of the text
- considers vector spaces $V_k = span\{P_j : j = 1, \dots, k\}$
- spaces $V_1 \subset V_2 \subset \cdots \subset V_N$ form a flag in $\mathbb{R}^M$
- $F(d_1, \dots, d_\ell)$ the flag variety of flags $W_1 \subset \cdots W_\ell$ with $\dim(W_k/W_{k-1}) = d_k$
- associate to a text T the point of the corresponding flag variety $F(1, \dots, 1, M-N)$ determined by the flag $V_1 \subset V_2 \subset \cdots \subset V_N$ with $V_k = span\{P_j : j = 1, \dots, k\}$

Semantic comparison in Flag varieties

- Fubini–Study metric on projective spaces has analog for Grassmannians and flag varieties obtained from curvature form of first Chern class of determinant line bundle of a hermitian vector bundle
- equivalently obtained by realizing as quotients of $SU(n)$ by subgroups, with the metric induced from the bivariant metric of $SU(n)$
- compare texts in Grassmannians or flag varieties by measuring distance in this metric

Semantic Dictionaries

- instead of "lexemes dictionary" D of words passing to a "semantic dictionary" S where lexemes are grouped together according to some semantic description
- this can be done in two ways
  1. Supervised Learning: "sense tagging", lexemes grouped together in semantic categories by assigning tagging; when semantic vectors retains correct information when passing to quotient semantic categories?
  2. Unsupervised Learning: "sense discrimination" by grouping together words into unlabelled groups using information contained in semantic vectors; resulting grouping in terms of *persistent topology*

Supervised Learning

- associate to texts T points $p(T)$ in a Grassmannian (either $Gr(N,M)$ for $N < M$ (case $N > M$ similar)
- operation of passing from the lexemes in D to the semantic categories in S as projection

$$\pi_{M,M'} : Gr(N,M) \twoheadrightarrow Gr(N,M')$$

  with $M' \leq M$ size of set of semantic categories $M' = \#S$
- a corpus $\mathcal{C} = \{T\}$ of texts T with fixed number of contexts and size of dictionary D
- view corpus $\mathcal{C} = \{T\}$ as a discrete sampling of a subvariety of the Grassmannian $Gr(N,M)$

Projectability in Grassmannians

- $\Pi_{\mathcal{C}} = \{p(T)\}_{T\in\mathcal{C}}$ finite set of points in $Gr(N,M)$ corresponding to texts in a corpus
- possible algebraic subvarieties $X_{\mathcal{C}} \subset Gr(N,M)$ that interpolate the points $p(T) \in \Pi_{\mathcal{C}}$: algebraic subvarieties $X_{\mathcal{C}}$ of $Gr(N,M)$ with $\Pi_{\mathcal{C}} \subset X_{\mathcal{C}}$
- projectability in Grassmannians: subvariety $X \subset Gr(N,M)$ is k–projectable, for some $0 \leq k \leq N-1$, under $\pi_{M,M'} : Gr(N,M) \twoheadrightarrow Gr(N,M')$ if any two N–planes in the image of X only meet along linear spaces of dimension less than k
- case $k = N$ corresponds to X being isomorphically projectable to $Gr(N,M')$
- k–projectability implies no two N–planes in X can intersect in dimension $\geq k$

- variety $X_{\mathcal{C}}$ associated to a corpus $\mathcal{C}$ of texts k–projectable to $Gr(N, M')$ means N-planes $\pi_{M,M'}(p(T))$ and $\pi_{M,M'}(p(T'))$ of $p(T)$, $p(T')$, with $T, T' \in \mathcal{C}$, intersect in at most a $(k-1)$-dimensional space
- size of intersection between N-planes of T and T' measures of dependence between semantic vectors, hence semantic relatedness of texts
- if for $X_{\mathcal{C}}$ every two N–planes intersect in dimension less than k but $X_{\mathcal{C}}$ not k-projectable under $\pi_{M,M'} : Gr(N, M) \twoheadrightarrow Gr(N, M')$ there is *loss of semantic information* in passing to semantic categories
- strong algebro-geometric constraints on projectability: example Veronese embedding of $\mathbb{P}^n$ only variety in $Gr(d-1, dn+d-1)$ that can be projected to $Gr(d-1, n+2d-3)$ so that any two $(d-1)$-planes meet in at most one point

Projectability and Paths in Projective Spaces

- polygonal paths in projective space $\mathbb{P}^{M-1}$ with $M = \#D$ size of dictionary
- question about k–projectable subvarieties in projective spaces
- algebraic subvarieties $X_{\mathcal{C}}$ of $\mathbb{P}^{M-1}$ containing points $\Pi_{\mathcal{C}} = \{p_k(T) \,:\, T \in \mathcal{C},\, k = 1, \ldots, N(T)\}$
- if geodesically complete contains also paths $\Gamma(T)$
- consider projection $\pi_{M,M'} : \mathbb{P}^{M-1} \twoheadrightarrow \mathbb{P}^{M'-1}$ corresponding to identification of lexemes according into semantic categories
- problem of projecting isomorphically subvariety $X_{\mathcal{C}}$ of $\mathbb{P}^{M-1}$ containing points $\Pi_{\mathcal{C}}$ to quotient $\mathbb{P}^{M'-1}$
- strong restrictions on isomorphically projectable subvarieties: example only n-dimensional variety that can be isomorphically projected from $Gr(1, 2n+1)$ to $Gr(1, n)$ Veronese variety embedding of $\mathbb{P}^n$ in $Gr(1, 2n+1)$ via $O_{\mathbb{P}^n}(1)^{\oplus d}$
- if not isomorphically projectable from $\mathbb{P}^{M-1}$ to $\mathbb{P}^{M'-1}$ then some loss of information in semantic vectors

Topologically detecting semantic relatedness

- set Π of points in a metric space, Vietoris–Rips complexes $R(\Pi, \epsilon)$ at scale $\epsilon > 0$
- $R_n(\Pi, \epsilon)$ spanned by unordered $(n+1)$-tuples of points in Π with all pairwise distances $\leq \epsilon$
- maps in homology $H_n(R(\Pi, \epsilon_1)) \to H_n(R(\Pi, \epsilon_2))$: persistent homology
- $\Pi_{\mathcal{C}} = \{p(T)\}_{T \in \mathcal{C}}$ points in $Gr(N, M)$ corresponding to texts in a large corpus
- semantic similarity detected by persistent H_0 persistent connected components
- additional structure of relatedness in higher persistent H_k
- under projection $\pi_{M,M'} : Gr(N, M) \twoheadrightarrow Gr(N, M')$ measure change in semantic relatedness by change in persistent topology

Unsupervised Learning

- ‘sense discrimination” obtained solely from position of semantic vectors without external tagging by semantic categories
- search for semantic relatedness in unsupervised context by frequent co-occurrences within same context
- many co-occurrences just for syntactic reasons, but those for words in different parts of speech, while semantic co-occurrences more often found between same part of speech
- vectors $P_k(T) = (p_{ik})_{i \in D}$ associated to contexts c_k in a text depend on both syntactic and semantic information

- how to make semantic vectors more syntax independent?
- possible way if have matched training corpus of different language translations of same texts and average semantic vectors $P_k(T, L)$ over set of languages L
- by considering points $p_k(T, L)$ for languages L in fixed ambient $\mathbb{P}^{M-1}$ and replace by barycenter $\bar{p}_k(T)$ in Fubini-Study metric
- reduces syntactic dependence if the set of languages has sufficiently different syntactic parameters (but cannot entirely decouple semantics from syntax)

Latent Semantics

- word–document semantic matrices are typically very sparse
- perform dimensional reduction based on a singular value decomposition
- semantic matrix P as product $U\Sigma V^\tau$ with U an $M \times M$ and V an $N \times N$ unitary and Σ $N \times M$ with the singular values on diagonal of rank $r = \mathrm{rank}(P)$
- truncations of matrix $U\Sigma V^\tau$ to rank $k < r$ approximation $U_k \Sigma_k V_k^\tau$ by k largest singular values
- creating low-dimensional linear mapping between words and contexts improving estimates of semantic similarity
- symmetric matrix $A = P^\tau P$ "term co-occurrence matrix" and its spectral decomposition
- "semantic spectrum": truncations obtained by applying power methods to separate span of eigenvectors of largest k eigenvalues of $A = P^\tau P$ from complementary space
- these operations also have natural geometric interpretation in terms of geometry of projective spaces and Grassmannians

Perron–Frobenius case and Riccati equation

- when only one top eigenvalue: Perron–Frobenius
- $Sp(A) = \{\lambda_1, \ldots, \lambda_N\}$ with $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_N|$
- $x_0 \in \mathbb{P}^{N-1}$ and sequence $x_m = A^m x_0$ converges to point in $\mathbb{P}^{N-1}$ line spanned by Perron–Frobenius eigenvector
- local chart vectors with first component equal to one:

$$A : x_m = \begin{pmatrix} 1 \\ y_m \end{pmatrix} \mapsto x_{m+1} = Ax_m = \begin{pmatrix} 1 \\ y_{m+1} \end{pmatrix}$$

$$y_{m+1} = \frac{A_3 + A_4 y_m}{A_1 + A_2 y_m}$$

where

$$A = \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix}$$

where A_4 is an $(N-1) \times (N-1)$-matrix and A_1 a number

- recursion relation of sequence y_m given by

$$y_{m+1} - y_m = (A_3 + A_4 y_m - y_m A_1 - y_m A_2 y_m)(A_1 + A_2 y_m)^{-1}$$

- discretization of the matrix Riccati equation

$$\frac{d}{dt} y(t) = A_3 + A_4 y(t) - y(t) A_1 - y(t) A_2 y(t)$$

- both equations have same stationary points given by solutions to

$$A_3 + A_4 y - y A_1 - y A_2 y = 0$$

- to find limit $x = \lim_m x_m$ stationary solution $y_{m+1} = y_m$ of difference equation consider Riccati flow to same fixed point
- Perron–Frobenius theory as matrix Riccati equation in projective space (Ammar–Martin formulation)

Latent Semantics and flows on Grassmannians

- similarly selection of span of eigenvectors of k largest eigenvalues of matrix $A = P^\tau P$ performed dynamically by a Riccati flow on Grassmannian $G(k, N)$
- k-dimensional vector space $V \in G(k, N)$ and matrix $A \in GL_N$ with $AV \in G(k, N)$ given by $AV = \{Av \, : \, v \in V\}$
- initial point $V_0 \in G(k, N)$ and sequence $V_{m+1} = AV_m$
- if U span of eigenvectors of largest λ_i with $i = 1, \dots, k$, sequence of points V_m in $G(k, N)$ converge point given by space U

- complementary subspaces $U \in G(k, N)$ and $W \in G(N-k, N)$, and morphism $L \in Hom(U, W)$, consider $U_L \in G(k, N)$ given by

$$U_L = \{ \begin{pmatrix} u \\ Lu \end{pmatrix} \,|\, u \in U \} \subset U \oplus W$$

- have in local chart on the Grassmannian for matrix A in decomposition $U \oplus W$ given by

$$A = \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix}$$

$$AU_L = U_{(A_3 + A_4 L)(A_1 + A_2 L)^{-1}}$$

- sequence

$$L_{m+1} = (A_3 + A_4 L_m)(A_1 + A_2 L_m)^{-1}$$

can be written as a difference equation

$$L_{m+1} - L_m = (A_3 + A_4 L_m - L_m A_1 - L_m A_2 L_m)(A_1 + A_2 L_m)^{-1}$$

- stationary solutions of difference equation

$$L_{m+1} - L_m = (A_3 + A_4 L_m - L_m A_1 - L_m A_2 L_m)(A_1 + A_2 L_m)^{-1}$$

as stationary points of matrix Riccati flow (Ammar–Martin formulation)

$$\frac{d}{dt} L(t) = A_3 + A_4 L(t) - L(t) A_1 - L(t) A_2 L(t)$$

- latent semantics method based on singular value decomposition and truncation to the top k singular values for P can be obtained via a geometric flow on a Grassmannian