Logique, Langage et Computation
Autour de la Logique Linéaire et ses Interfaces
IRPhiL, Lyon 3 – IXXI & MSH-LSE – ENS de Lyon
Lyon, France

Des statistiques à l'algèbre et au-delà

La logique des normes émergentes en langage naturel

Juan Luis Gastaldi et John Terilla

ETH zürich CUNY THE CITY UNIVERSITY OF NEW YORK

10 Novembre, 2023

Outline

Neural Word Embeddings

The Algebra behind Word Embeddings

Example: Wikipedia

The Structure behind the Algebra

Perspectives

Outline

Three Main Components of NLM To Be Explained

Subword Tokenization
(Sennrich et al., 2016)

Word Embeddings
(Mikolov, Sutskever, Chen, Corrado, and Dean, 2013)

Self-Attention
(Vaswani et al., 2017)

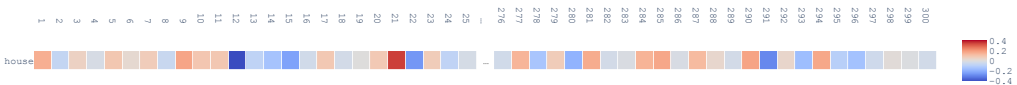
Three Main Components of NLM To Be Explained

Subword Tokenization
(Sennrich et al., 2016)

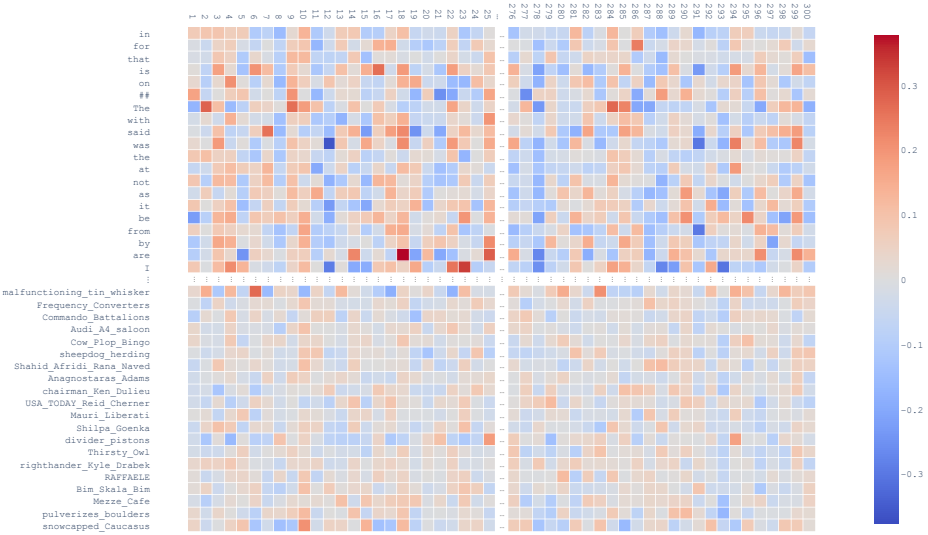
Word Embeddings
(Mikolov, Sutskever, Chen, Corrado, and Dean, 2013)

Self-Attention
(Vaswani et al., 2017)

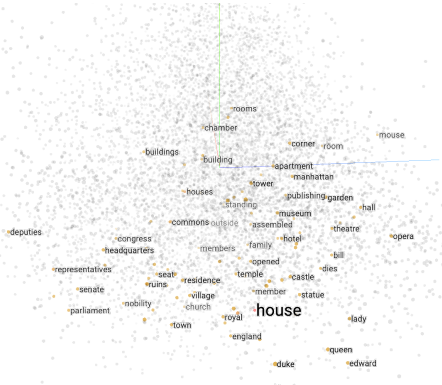
Word Embeddings: Vector



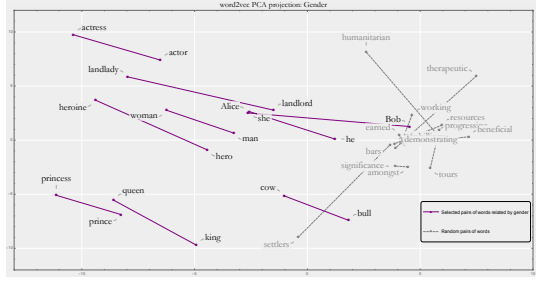
Word Embeddings: Matrix



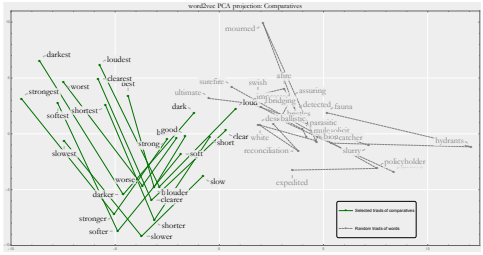
Embedding Space: Similarity and Analogy

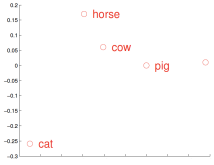


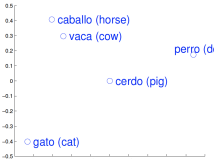
(https://projector.tensorflow.org)



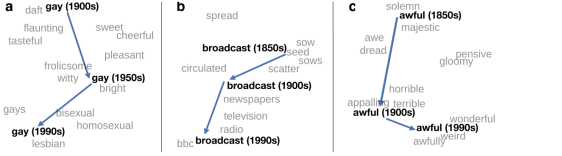
Embedding Space: Other Applications



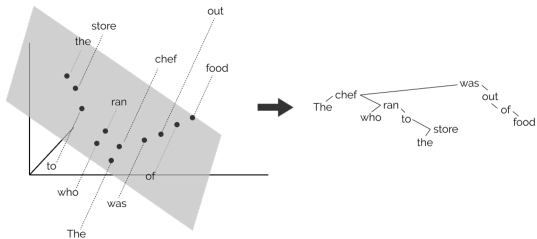




(Mikolov, Sutskever, Chen, Corrado, Dean, et al., 2013)

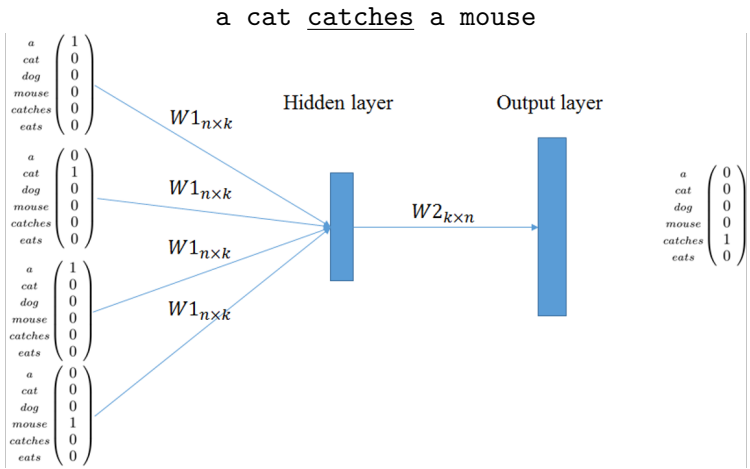


(Hamilton et al., 2016)



(https://nlp.stanford.edu/~johnhew/structural-probe.html)

word2vec Models

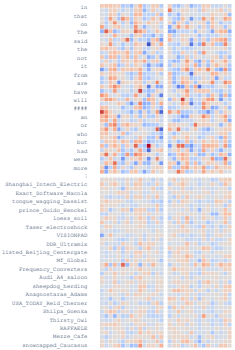


Credit: Ferrone et al., 2017

Outline

word2vec as Implicit Matrix Factorization

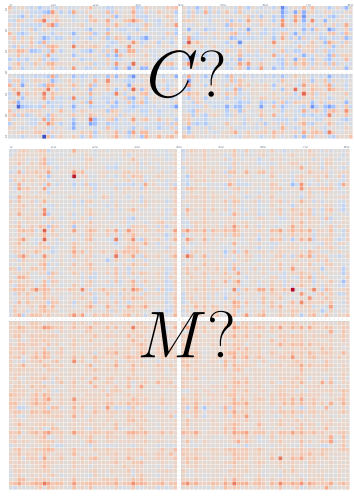
(Levy and Goldberg, 2014)



word2vec as Implicit Matrix Factorization

(Levy and Goldberg, 2014)

W



$C?$

$M?$

$$W \times C \approx M$$

word2vec Explained

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

word2vec Explained

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w, c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

word2vec Explained

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

Three results:

- $M = PMI(w,c) - \log k$ (Pointwise Mutual Information)

word2vec Explained

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

Three results:

- $M = PMI(w,c) - \log k$ (Pointwise Mutual Information)
- W is low dimensional

word2vec Explained

(Levy and Goldberg, 2014)

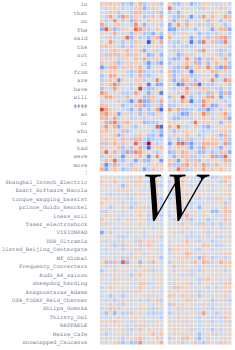
$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

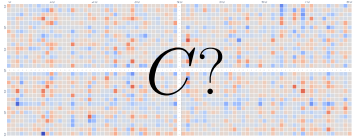
Three results:

- $M = PMI(w,c) - \log k$ (Pointwise Mutual Information)
- W is low dimensional
- The Singular Value Decomposition (SVD) provides an exact solution to find W

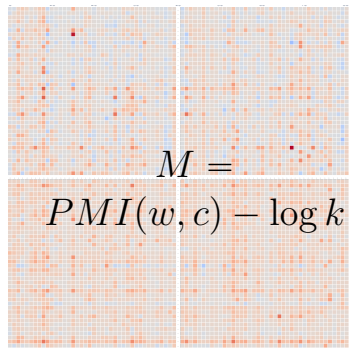
Pointwise Mutual Information (PMI)



W



$C?$



$M = PMI(w,c) - \log k$

$$PMI(w,c) = \log \frac{p(w,c)}{p(w)p(c)}$$

Singular Value Decomposition (SVD)

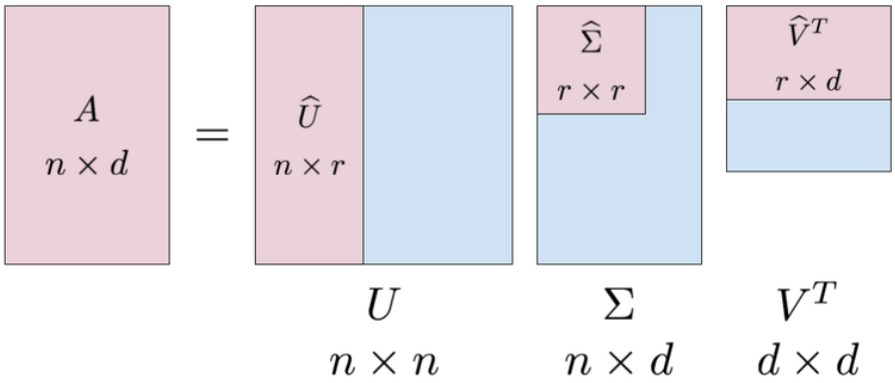
$$M = U\Sigma V^*$$

Where:

M = $m \times n$ (real or complex) matrix
U = $m \times m$ unitary matrix
Σ = $m \times n$ non-negative real rectangular diagonal matrix
V^* = conjugate transpose of V, a $n \times n$ unitary matrix

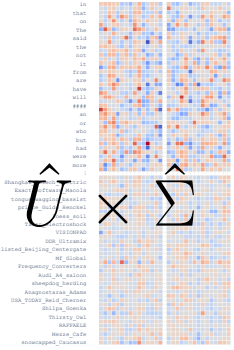
Truncated SVD

$$M = U\Sigma V^*$$



Credit: Angela Ju

Embeddings as Truncated SVD



$$\hat{U} \times \hat{\Sigma}$$

$$\hat{V}^*$$

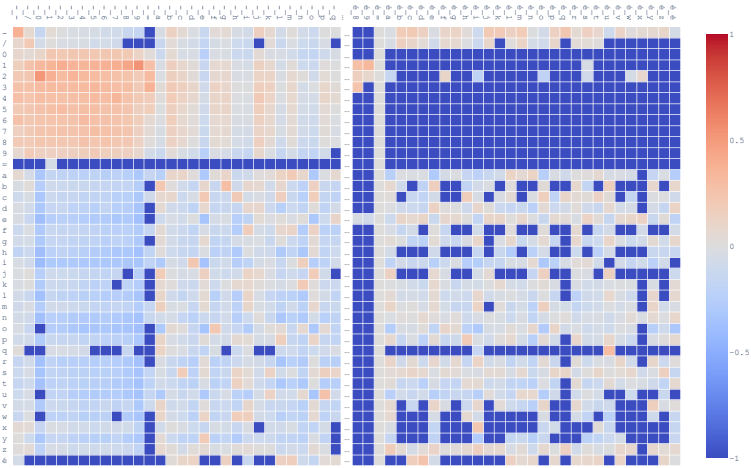
$$M = PMI(w, c) - \log k$$

$$M \approx \hat{U} \times \hat{\Sigma} \times \hat{V}^*$$

Outline

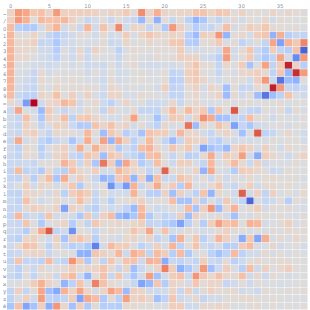
Example: Characters in Wikipedia

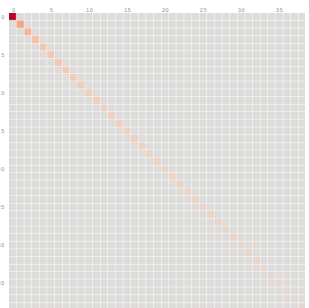
$$PMI(w, c) = \log \frac{p(w,c)}{p(w)p(c)}$$

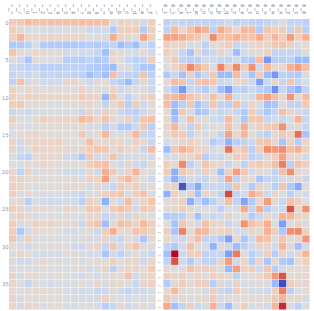


SVD of Wikipedia Character PMI Matrix

U Σ V^*

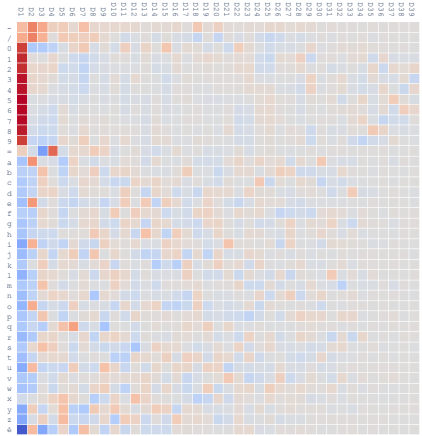






Truncate and Embed

$$U \times \Sigma$$

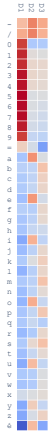


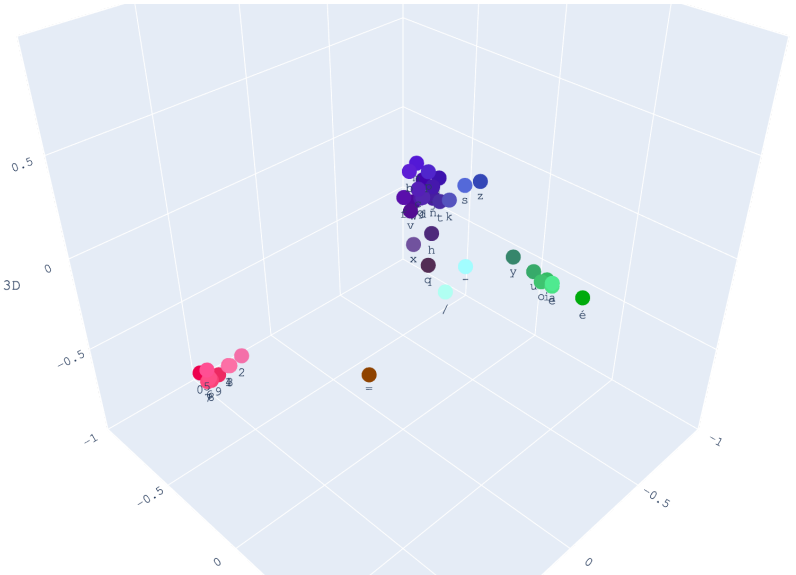
Truncate and Embed

$$\hat{U} \times \hat{\Sigma}$$

Truncate and Embed

$$\hat{U} \times \hat{\Sigma}$$





Outline

4 Why does this produce good word representations?

Good question. We don't really know.

The distributional hypothesis states that words in similar contexts have similar meanings. The objective above clearly tries to increase the quantity $v_w \cdot v_c$ for good word-context pairs, and decrease it for bad ones. Intuitively, this means that words that share many contexts will be similar to each other (note also that contexts sharing many words will also be similar to each other). This is, however, very hand-wavy.

Can we make this intuition more precise? We'd really like to see something more formal.

(Goldberg and Levy, 2014)

Singular Value Decomposition (SVD)

$$M = U\Sigma V^*$$

Where:

M = $m \times n$ (real or complex) matrix
U = $m \times m$ unitary matrix
Σ = $m \times n$ non-negative real rectangular diagonal matrix
V^* = conjugate transpose of V, a $n \times n$ unitary matrix

Singular Value Decomposition (SVD)

$$M = U\Sigma V^*$$

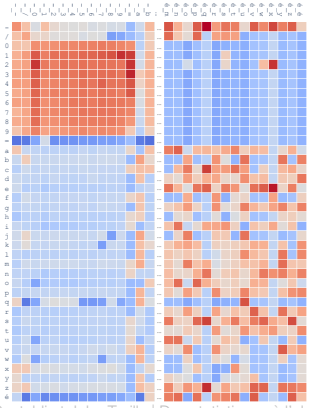
Where:

$$\begin{aligned} M &= m \times n \text{ (real or complex) matrix} \\ U &= m \times m \text{ unitary matrix} \\ \Sigma &= m \times n \text{ non-negative real rectangular diagonal matrix} \\ V^* &= \text{conjugate transpose of } V\text{, a } n \times n \text{ unitary matrix} \end{aligned}$$

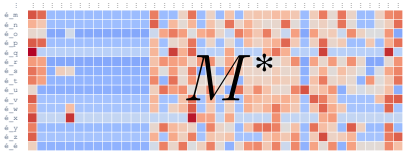
In particular:

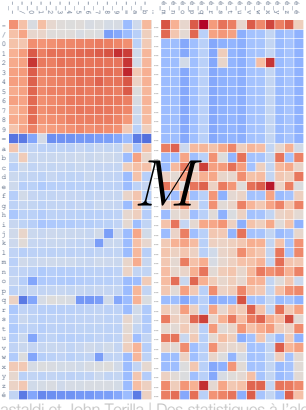
- The columns of U (left singular vectors) are eigenvectors of $M \times M^*$
- The rows of V^* (right singular values) are eigenvectors of $M^* \times M$
- The non-zero elements of Σ (non-zero singular values) are the square roots of the non-zero eigenvalues of $M \times M^*$ or $M^* \times M$

$M \times M^*$ as a Covariance Matrix

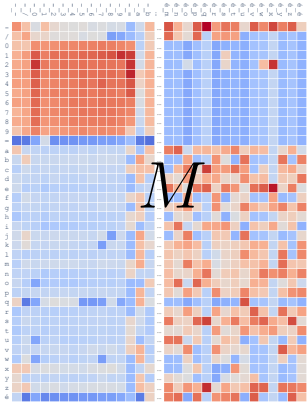


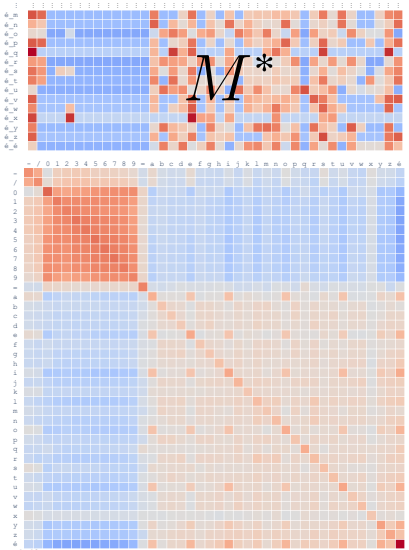
$M \times M^*$ as a Covariance Matrix



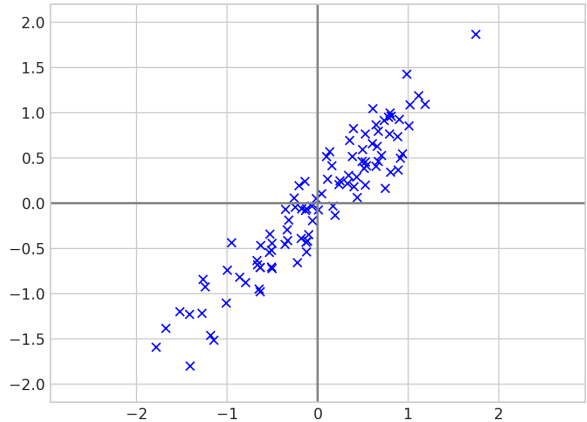


$M \times M^*$ as a Covariance Matrix



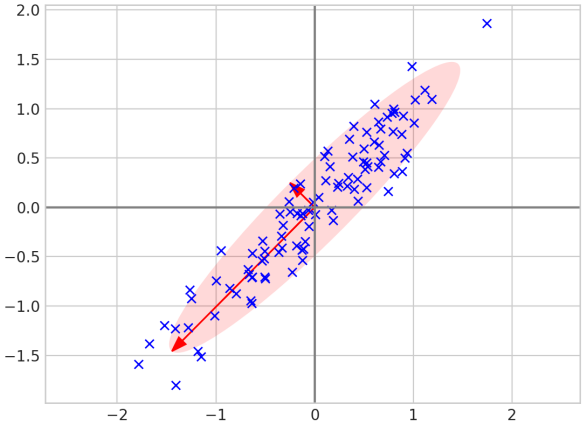


Eigenvectors and Eigenvalues



Credit: Joel Laity

Eigenvectors and Eigenvalues

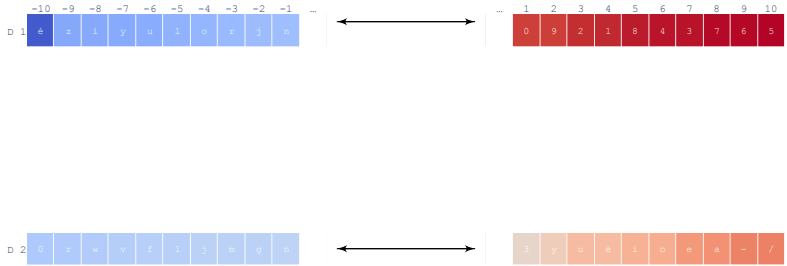


Credit: Joel Laity

Structural Features

Eigenvectors of $M \times M^*$:

“Typing” Information

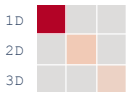


Eigenvectors of $M \times M^*$:

Structural Features

Eigenvectors of $M \times M^*$:

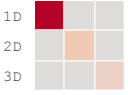
Eigenvalues of $M \times M^*$:



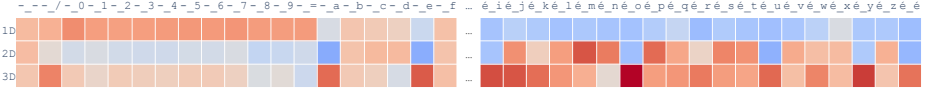
Structural Features

Eigenvectors of $M \times M^*$:

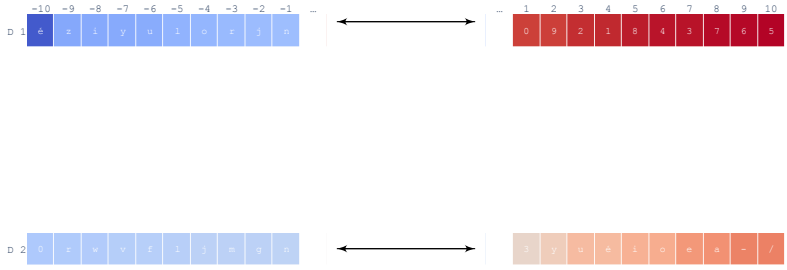
Eigenvalues of $M \times M^*$:



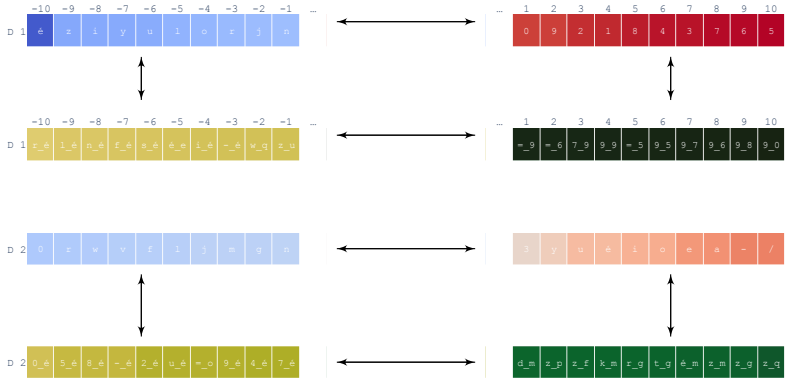
Eigenvectors of $M^* \times M$:



“Typing” Information

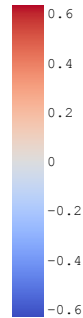


“Typing” Information



“Typing” Information (words)

	−1	−2	−3	−4	...	4	3	2	1
D1	church	university	field	house	...	used	offered	found	made
D2	use	leave	keep	buy	...	club	sun	uk	hotel
D3	show	boy	project	move	...	size	interests	activities	nature
D4	used	expected	made	considered	...	europe	scotland	france	england
D5	used	expected	food	water	...	couple	under	series	lot
D6	perhaps	indeed	under	during	...	bit	series	couple	lot
D7	difficult	hard	easy	necessary	...	gave	started	saw	took
D8	europe	scotland	england	france	...	want	needs	army	could
D9	wish	tried	seem	seemed	...	established	won	published	produced
D10	10	15	20	30	...	on	black	into	through

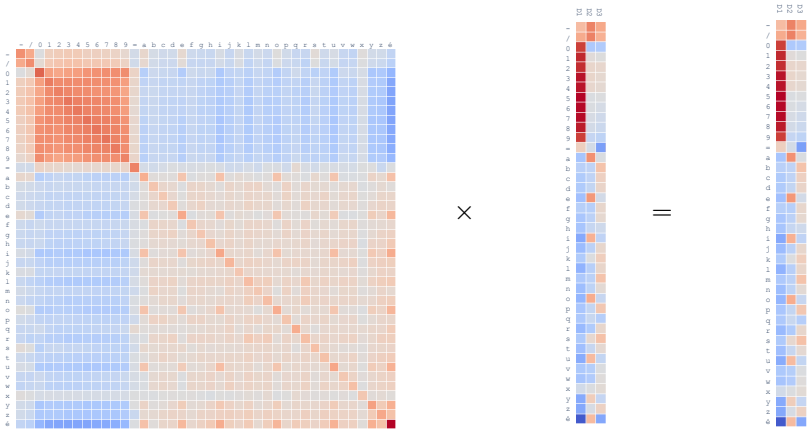


Eigenvectors as Fixed Points

$$(M \times M^*)v = \lambda v$$

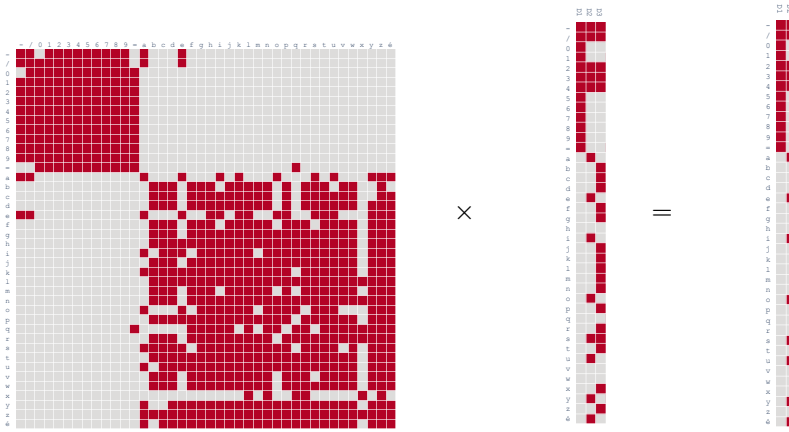
Eigenvectors as Fixed Points

$$(M \times M^*)v = \lambda v$$

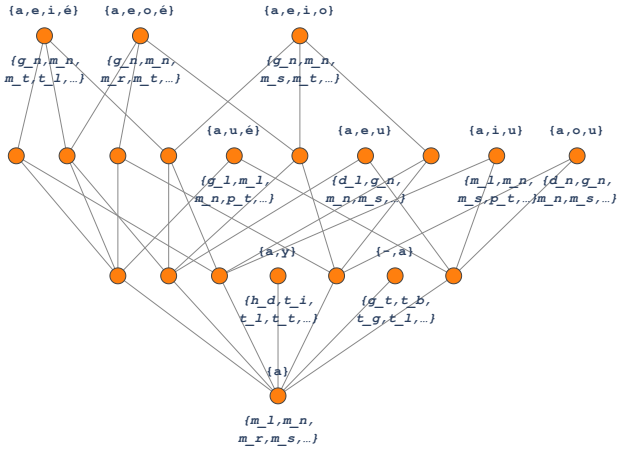


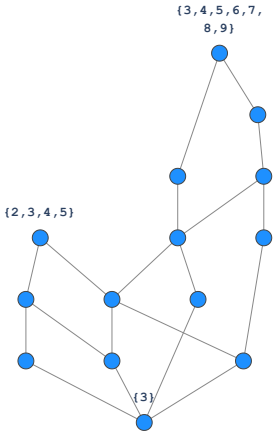
Binary Matrices: Formal Concepts

$$(M_t \star M_t^*) \star v = v$$

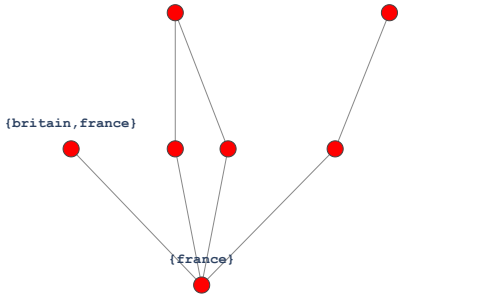


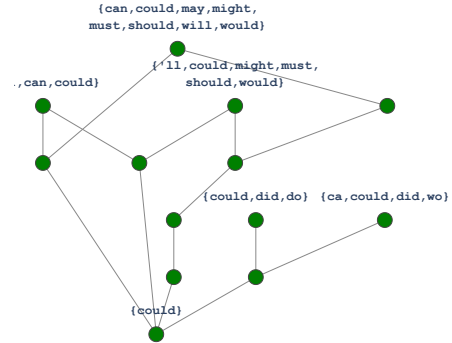
Formal Concepts



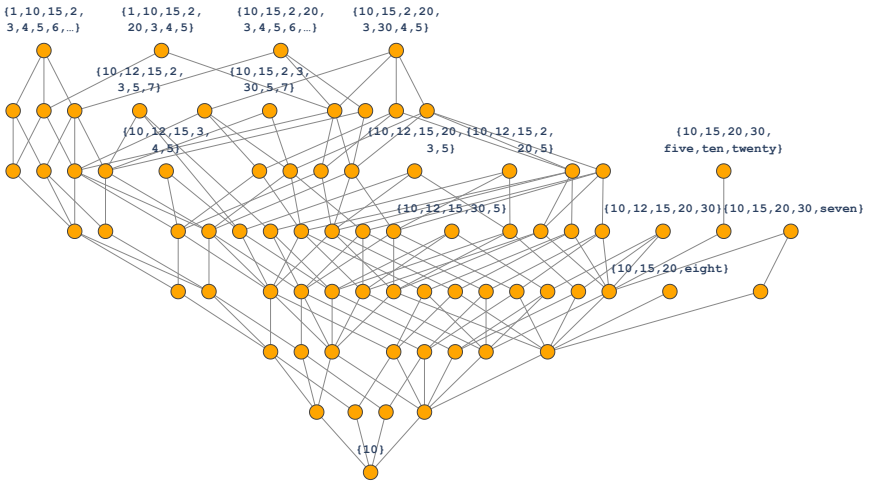


Formal Concepts (words)





Formal Concepts (words)



Outline

Perspectives

- Three possible lines of research:
  - Linguistically grounded decomposition of the PMI matrix
    Positive vs. Negative, Left vs. Right
  - Refinement through a categorical approach
    Nuclei of profunctors
  - Linear Logic
    (classic, probabilistic, Banach, quantum) coherent spaces
- Connection between all those perspectives
- Computationally: Tensor Networks
- Connection between word embeddings, subword tokenization (segmentation), and attention mechanisms (grammar)

Reference Papers

Reference Papers

- Bradley, T.-D., Gastaldi, J. L., & Terilla, J. (Forthcoming 2024a). The structure of meaning in language: parallel narratives in linear algebra and category theory. Notices of the AMS

- Pestun, V., Terilla, J., & Vlassopoulos, Y. (2017). Language as a matrix product state.
- Gastaldi, J. L. (2020). Why can computers understand natural language?: The structuralist image of language behind word embeddings. *Philosophy & Technology*
- Bradley, T.-D., Stoudenmire, E. M., & Terilla, J. (2020). Modeling sequences with quantum states: A look under the hood. *Machine Learning: Science and Technology*, *1*(3), 035008. https://doi.org/10.1088/2632-2153/ab8731
- Gastaldi, J. L., & Pellissier, L. (2021). The calculus of language: Explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews*. https://doi.org/10.1080/03080188.2021.1890484
- Bradley, T.-D., Terilla, J., & Vlassopoulos, Y. (2021). An enriched category theory of language: From syntax to semantics.

References I

Bradley, T.-D., Gastaldi, J. L., & Terilla, J. (Forthcoming 2024a). The structure of meaning in language: parallel narratives in linear algebra and category theory. *Notices of the AMS*.

Bradley, T.-D., Stoudenmire, E. M., & Terilla, J. (2020). Modeling sequences with quantum states: A look under the hood. *Machine Learning: Science and Technology*, *1*(3), 035008. https://doi.org/10.1088/2632-2153/ab8731

Bradley, T.-D., Terilla, J., & Vlassopoulos, Y. (2021). An enriched category theory of language: From syntax to semantics.

Gastaldi, J. L. (2020). Why can computers understand natural language?: The structuralist image of language behind word embeddings. *Philosophy & Technology*.

Gastaldi, J. L., & Pellissier, L. (2021). The calculus of language: Explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews*. https://doi.org/10.1080/03080188.2021.1890484

Goldberg, Y., & Levy, O. (2014). Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, *abs/1402.3722*.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *CoRR*, *abs/1605.09096*.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 2177–2185.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., Le, Q., & Strohmann, T. (2013). *Learning representations of text using neural networks. NIPS deep learning workshop 2013 slides*.

References II

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *CoRR*, *abs/1310.4546*.

Pestun, V., Terilla, J., & Vlassopoulos, Y. (2017). Language as a matrix product state.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the ACL*, 1715–1725.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need.

Logique, Langage et Computation
Autour de la Logique Linéaire et ses Interfaces
IRPhiL, Lyon 3 – IXXI & MSH-LSE – ENS de Lyon
Lyon, France

Des statistiques à l'algèbre et au-delà

La logique des normes émergentes en langage naturel

Juan Luis Gastaldi et John Terilla

ETH zürich CUNY THE CITY UNIVERSITY OF NEW YORK

10 Novembre, 2023