Seminar “Understanding language”
Copenhagen Linguistic Circle
University of Copenhagen
Copenhagen, Denmark

Understanding Language (Models)

AI Language Models From a Structuralist Perspective

Juan Luis Gastaldi

ETH zürich

May 2nd, 2023

This project has received funding from the
European Union’s Horizon 2020 research and innovation programme
under grant agreement No 839730

Outline

Outline

Philosophy and NLP

- Philosophy seems absent from current developments in neural language models (NLM)
  - Usually summoned to do “ethics of AI”

Philosophy and NLP

- Philosophy seems absent from current developments in neural language models (NLM)
  - Usually summoned to do “ethics of AI”
- However, there are many epistemological challenges where philosophy could intervene
  - What do NLM teach us about language?

Philosophy and NLP

- Philosophy seems absent from current developments in neural language models (NLM)
  - Usually summoned to do "ethics of AI"
- However, there are many epistemological challenges where philosophy could intervene
  - What do NLM teach us about language?
  - One question in this direction: How can meaning emerge from (linguistic) form?
    - The current debate seems to miss the point about the nature of language

Philosophy and NLP

- Philosophy seems absent from current developments in neural language models (NLM)
  - Usually summoned to do "ethics of AI"
- However, there are many epistemological challenges where philosophy could intervene
  - What do NLM teach us about language?
  - One question in this direction: How can meaning emerge from (linguistic) form?
    - The current debate seems to miss the point about the nature of language
  - The only beginning of an answer is given by the distributional hypothesis

The Distributional Hypothesis

- “You shall know a word by the company it keeps!” (Firth, 1957)
- “Words which are similar in meaning occur in similar contexts” (Rubenstein & Goodenough 1965)
- “Words with similar meanings will occur with similar neighbors if enough text material is available” (Schütze & Pedersen 1995)
- “A representation that captures much of how words are used in natural context will capture much of what we mean by meaning” (Landauer & Dumais 1997)
- “Words that occur in the same contexts tend to have similar meanings” (Pantel 2005)
- “The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear” (Lenci, 2010)

The Structuralist Hypothesis: Hjelmslev

"A priori it would seem to be a generally valid thesis that for every process there is a corresponding system, by which the process can be analyzed and described by means of a limited number of premises. It must be assumed that any process, can be analyzed into a limited number of elements recurring in various combinations. Then, on the basis of this analysis, it should be possible to order these elements into classes according to their possibilities of combination. And it should be further possible to set up a general and exhaustive calculus of the possible combinations."

(Hjelmslev, 1953, p. 9)

The Structuralist Hypothesis

- Meaning is the effect of structure

- Distributional properties convey meaning only through the action of a latent structure determining possible semantic values, and which is inseparable from the principles of identification of the elementary units of language, since meaning is the effect of discriminating operations performed through segmentation procedures of which the units of language keep the trace

- Linguistic content is the effect of a virtual structure of classes and dependencies at multiple levels underlying (and derivable from) the mass of things said or written in a given language

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)

Outline

Philosophy and NLP

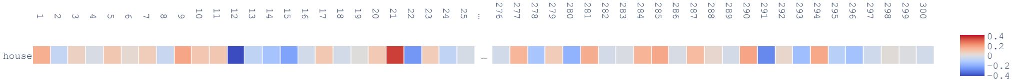
Word Embeddings

Example

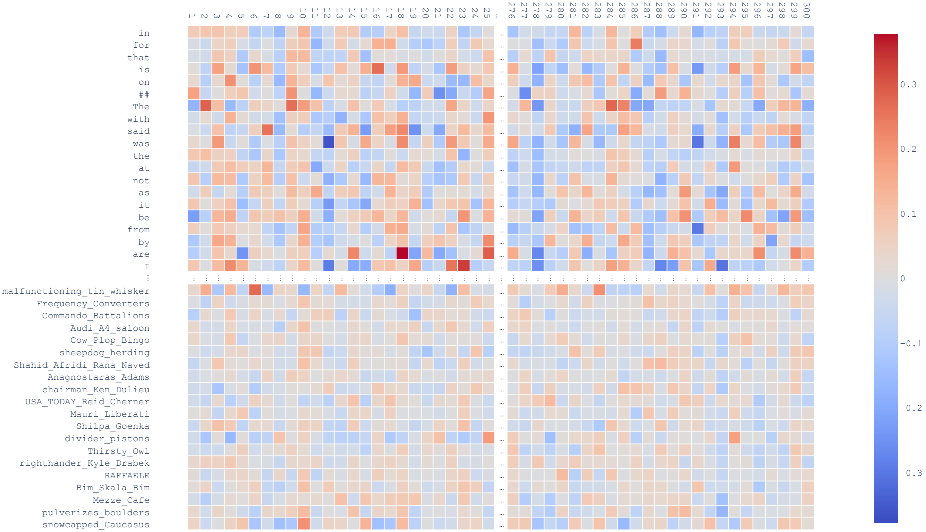
The Structure Behind Embeddings

Conclusions and Challenges

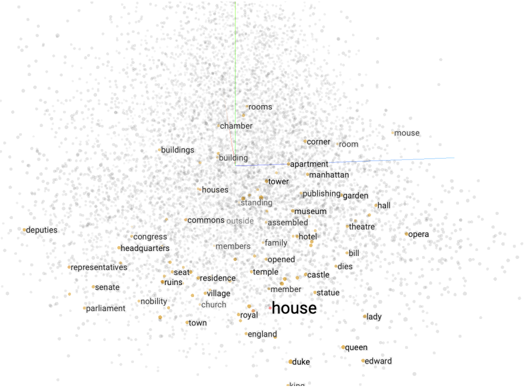
Word Embeddings: Vector



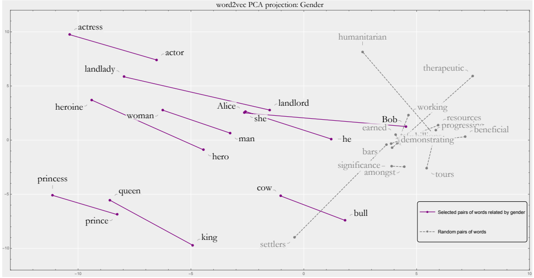
Word Embeddings: Matrix



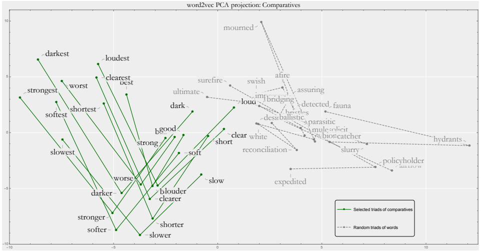
Embedding Space: Similarity and Analogy

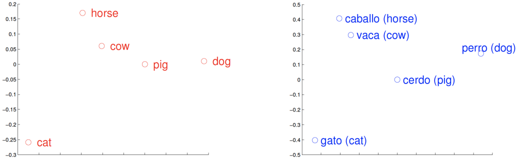


(https://projector.tensorflow.org)

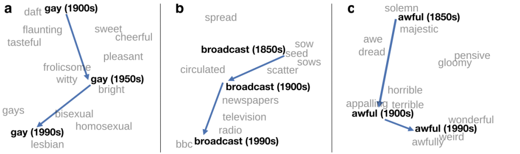


Embedding Space: Other Applications

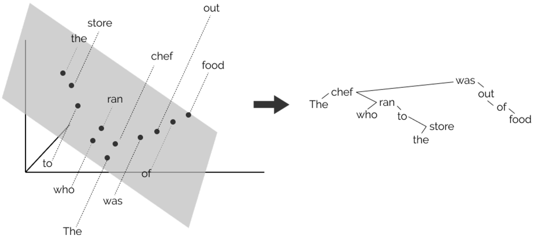




(Mikolov, Sutskever, Chen, Corrado, Dean, et al., 2013)

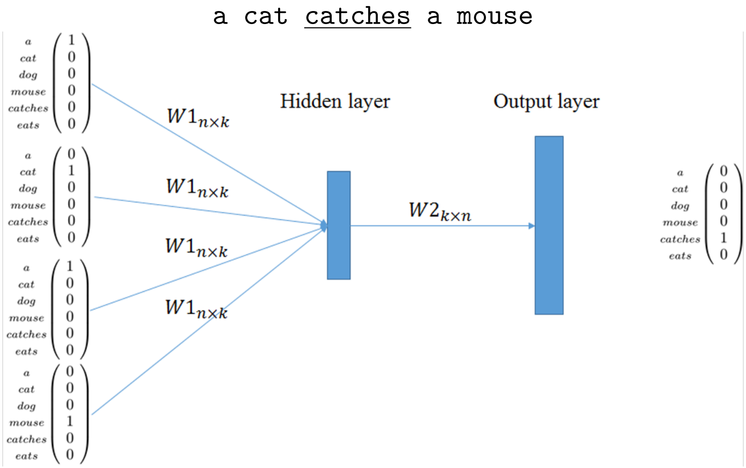


(Hamilton et al., 2016)



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

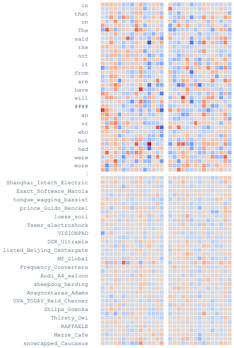
word2vec Models



Credit: Ferrone et al., 2017

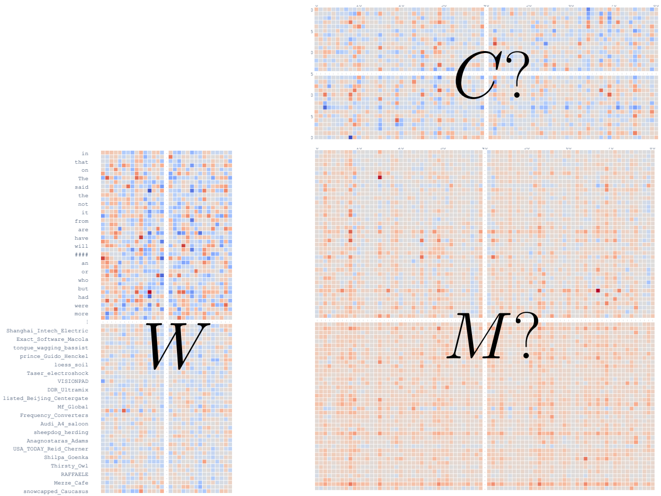
word2vec as Implicit Matrix Factorization

(Levy and Goldberg, 2014)



word2vec as Implicit Matrix Factorization

(Levy and Goldberg, 2014)



$$W \times C \approx M$$

word2vec Explained

(Levy and Goldberg, 2014)

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

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

Three results:

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

Three results:

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

word2vec Explained

(Levy and Goldberg, 2014)

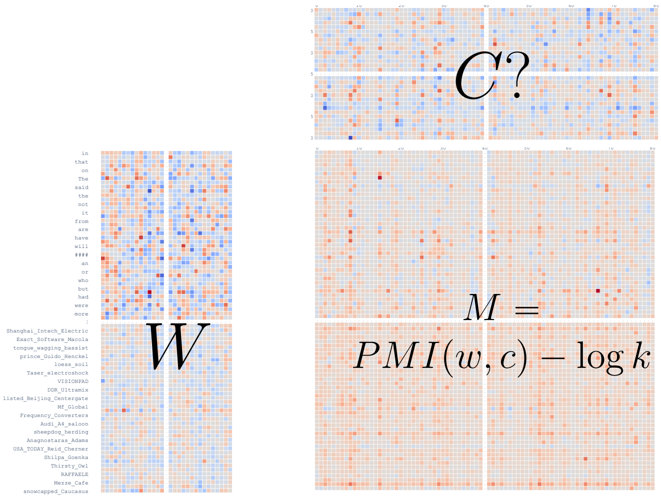
$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)\left(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)]\right)$$

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

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)



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

Singular Value Decomposition (SVD)

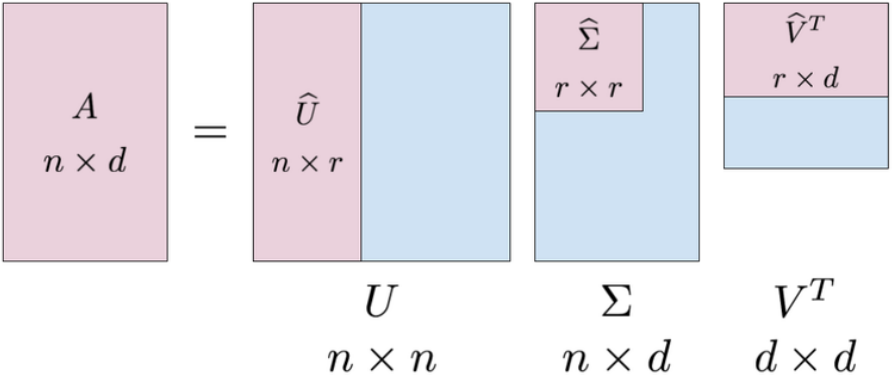
$$M = U\Sigma V^*$$

Where:

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

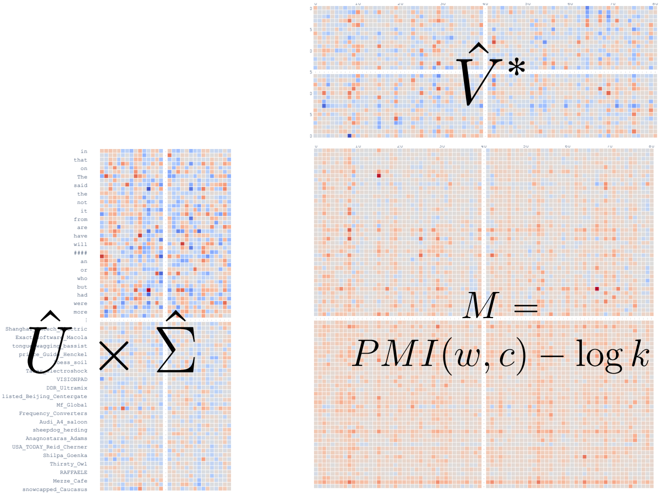
Truncated SVD

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



Credit: Angela Ju

Embeddings as Truncated SVD

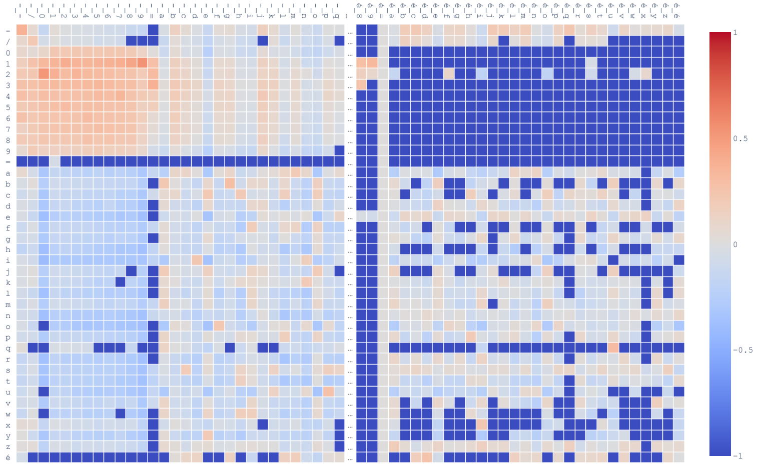


$$M \approx \hat{U} \times \hat{\Sigma} \times \hat{V}^*$$

Outline

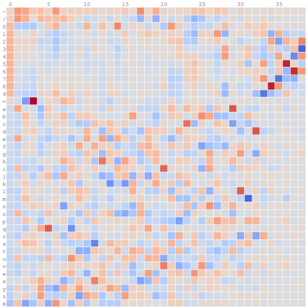
Example: Characters in Wikipedia

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

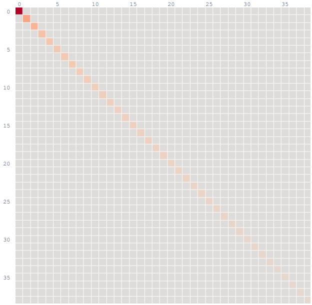


SVD of Wikipedia Character PMI Matrix

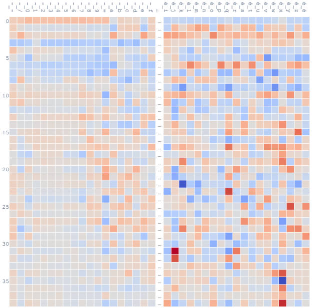
U



Σ

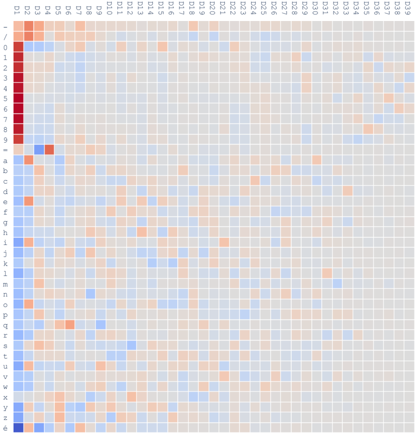


V^*



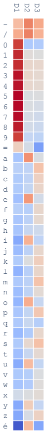
Truncate and Embed

$$U \times \Sigma$$



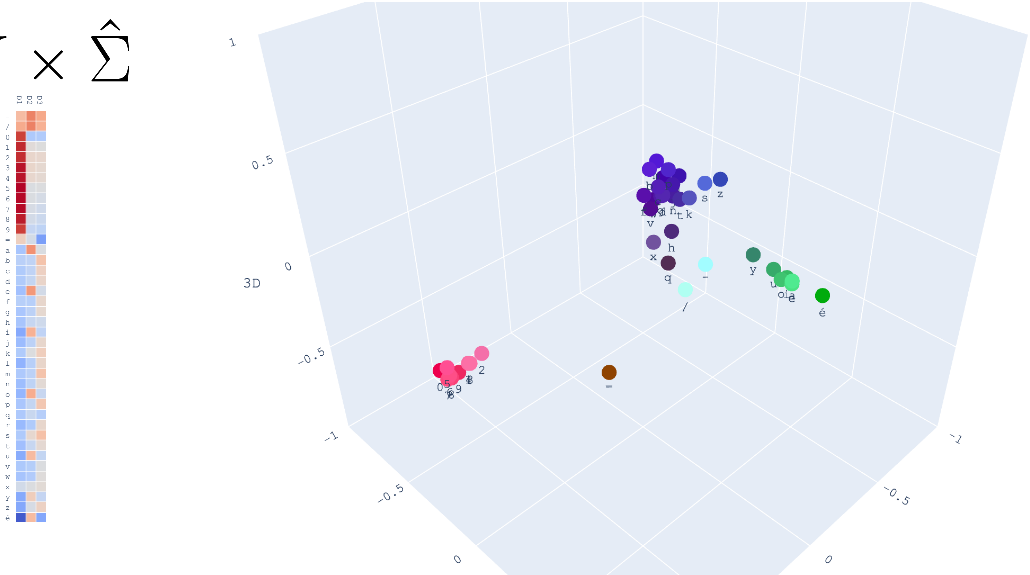
Truncate and Embed

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



Truncate and Embed

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



Outline

But Why?

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:

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

Singular Value Decomposition (SVD)

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

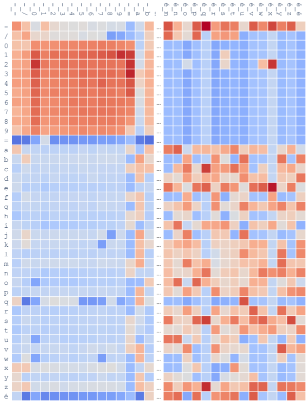
Where:

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

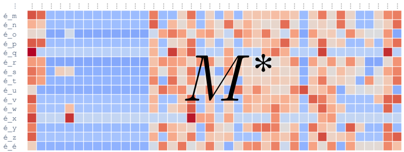
In particular:

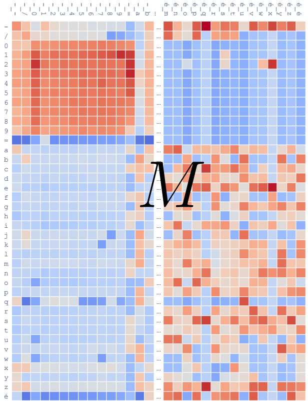
- The columns of U (left singular vectors) are eigenvectors of $M \times M^*$
- The rows of V^* (right singular values) are eigenvectors of $M^* \times M$
- The non-zero elements of Σ (non-zero singular values) are the square roots of the non-zero eigenvalues of $M \times M^*$ or $M^* \times M$

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

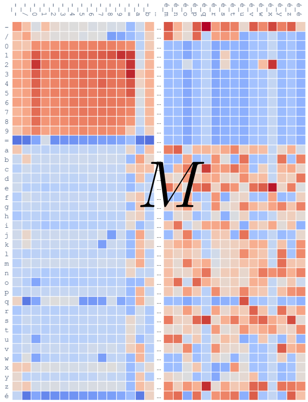


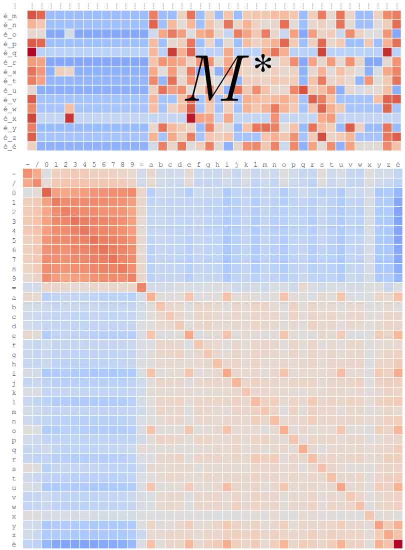
$M \times M^*$ as A Covariance Matrix



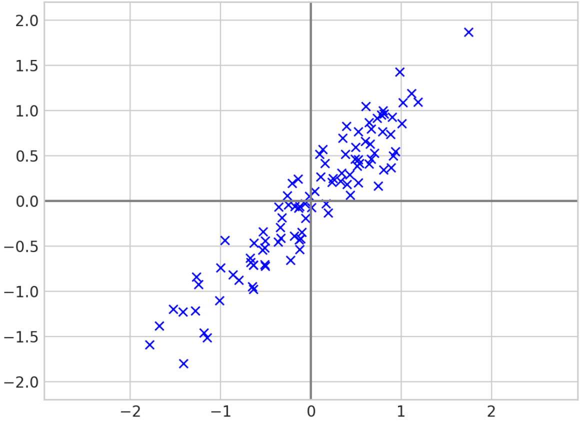


$M \times M^*$ as A Covariance Matrix



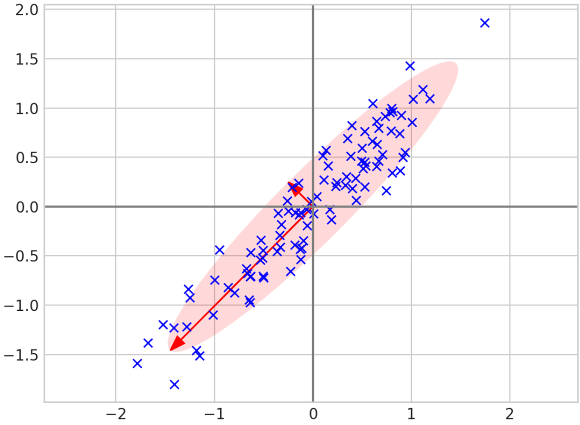


Eigenvectors and Eigenvalues



Credit: Joel Laity

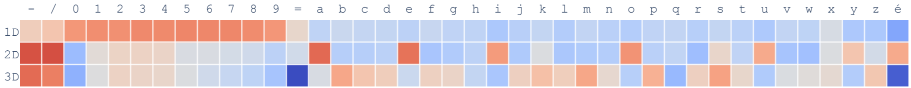
Eigenvectors and Eigenvalues



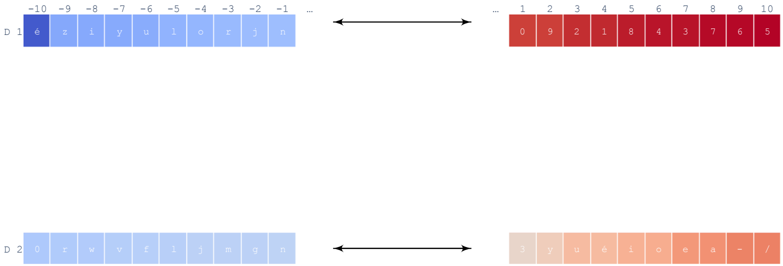
Credit: Joel Laity

"Eigenstructure"

Eigenvectors of $M \times M^*$:

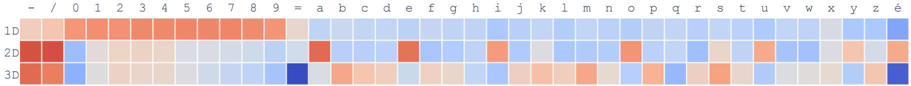


Commutation



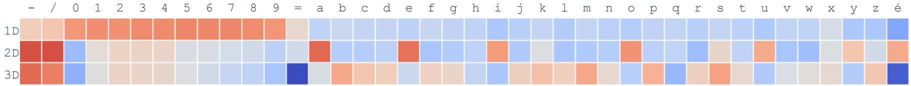
"Eigenstructure"

Eigenvectors of $M \times M^*$:

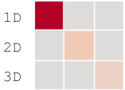


“Eigenstructure”

Eigenvectors of $M \times M^*$:

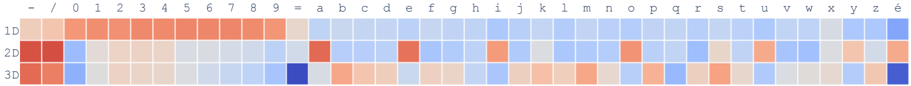


Eigenvalues of $M \times M^*$:



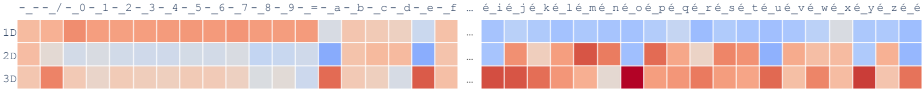
“Eigenstructure”

Eigenvectors of $M \times M^*$:

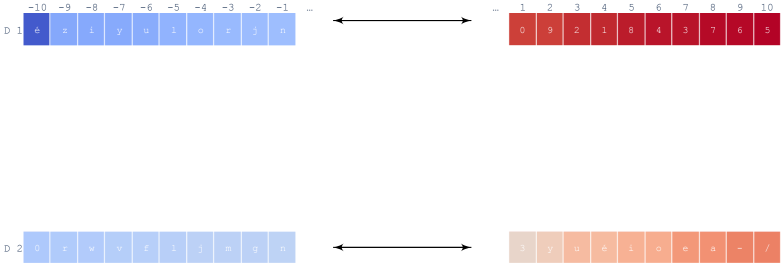


Eigenvalues of $M \times M^*$:

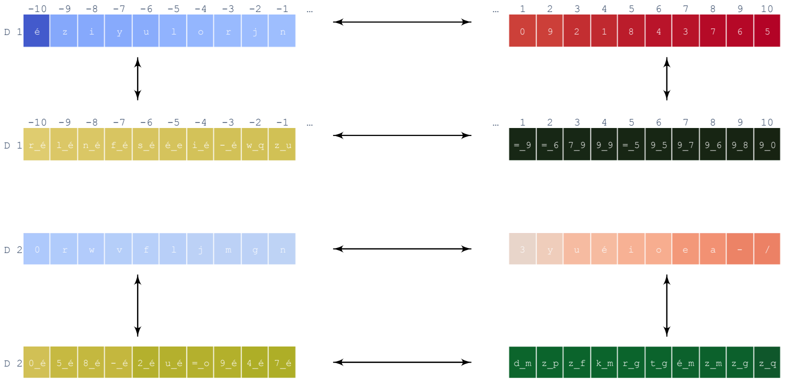
Eigenvectors of $M^* \times M$:



Commutation



Commutation

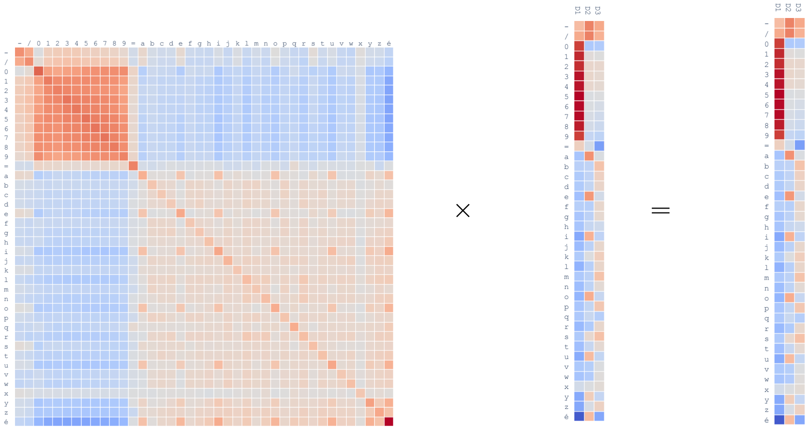


Eigenvectors as Fixed Points

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

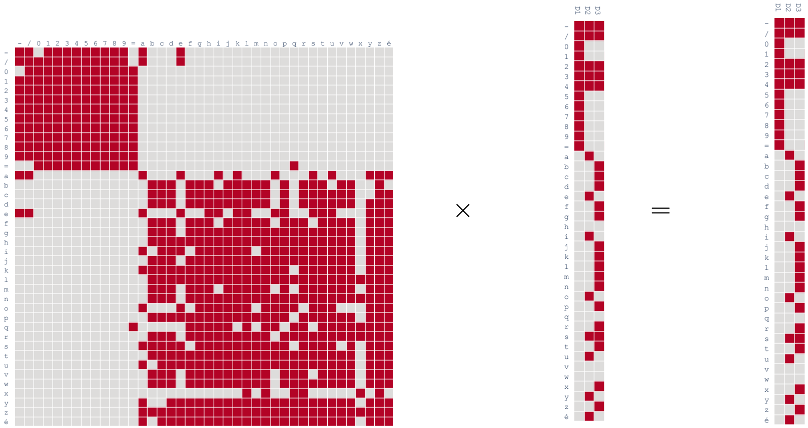
Eigenvectors as Fixed Points

$$(M \times M^{*})\textcolor{red}{v} = \lambda \textcolor{red}{v}$$

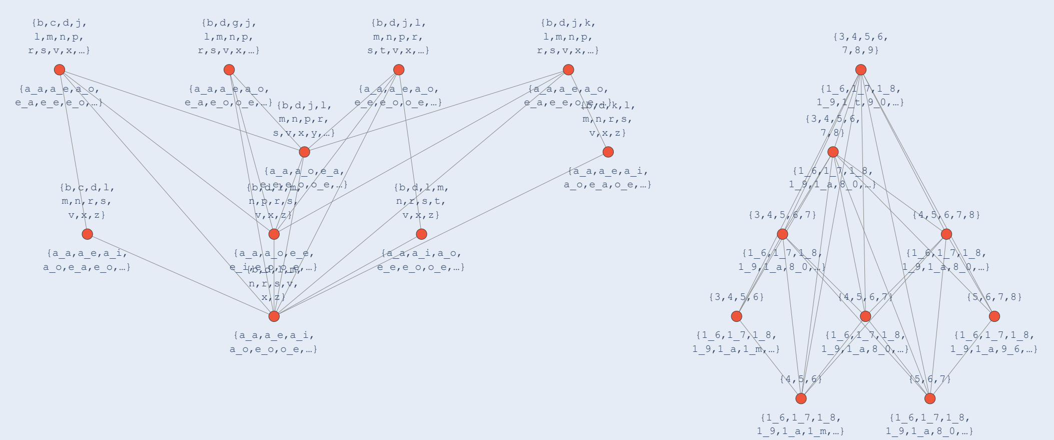


Binary: Formal Concepts

$$(M \times M^*)\textcolor{red}{v} = \lambda \textcolor{red}{v}$$



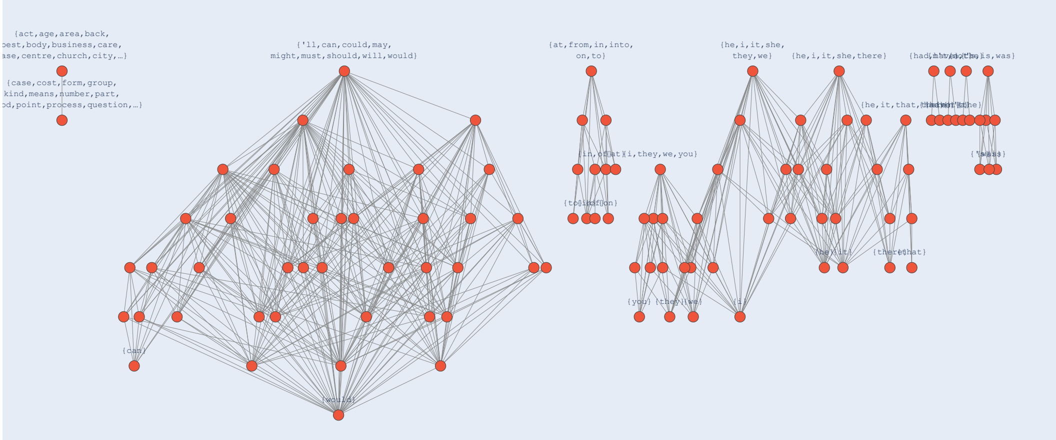
Formal Concepts



Words

	-5	-4	-3	-2	-1	0	1	2	3	4	5
D 1	church	university	field	house	centre	...	held	used	offered	found	made
D 2	use	leave	keep	buy	meet	...	boy	club	sun	uk	hotel
D 3	show	boy	project	move	play	...	production	size	interests	activities	nature
D 4	used	expected	made	considered	allowed	...	london	europe	scotland	france	england
D 5	used	expected	food	water	england	...	during	couple	under	series	lot
D 6	perhaps	indeed	under	during	in	...	cup	bit	series	couple	lot
D 7	difficult	hard	easy	necessary	close	...	won	gave	started	saw	took
D 8	europe	scotland	england	france	lot	...	middle	want	needs	army	could
D 9	wish	tried	seem	seemed	began	...	received	established	won	published	produced
D 10	10	15	20	30	3	...	from	on	black	into	through

Formal Concepts Words



Outline

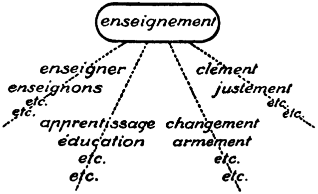
Conclusions

- Neural language models (NLMs) are the implicit implementation of a theory of language

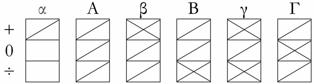
Conclusions

- Neural language models (NLMs) are the implicit implementation of a theory of language
- Starting from an operational treatment of syntagmatic relations, NLMs infer paradigmatic relations, and explore commutation properties identifying paradigmatic and syntagmatic units at different levels, and dependencies between them.

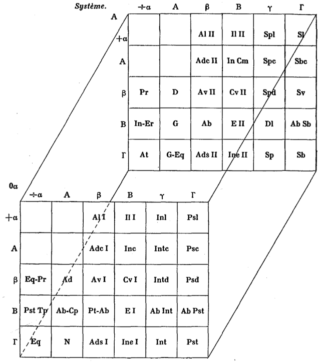
Structuralist Tools



(Saussure, 1980)



(Hjelmslev, 1975)



(Hjelmslev, 1935)

SEGMENTS	ENVIRONMENTS										
	#— r̥	#—r	#—l	e i —C	æ—C	a o u —C	s— e i	s—æ	s— a o u	...	t̥ —	C³—
t̥	√											
t		√		√	√	√	√	√	√			
K						√			√			
k		√	√		√			√				
ᴋ				√			√					
G						√						
g		√	√		√							
ɢ				√								
r				√	√	√						√
r̥											√	

(Harris, 1960)

Structuralist Tools

	a	b	d	e	f	g	h	i
a	aa	ab	ad		af	ag	ah	
b	ba							bi
d	da			de				di
e		eb	ed			eg		
f				fe				
g								gi
h	ha							hi
i			id				ih	i

Diagram 1.

	b	d	f	g	h	a	e	i
f						**fa**	fe	
h						ha		hi
g						**ga**	**ge**	gi
b						ba	**be**	bi
d						da	de	di
a	ab	ad	af	ag	ah	aa		
e	eb	ed	**ef**	eg				
i	**ib**	id		**ig**	ih			i

Diagram 3.

		I (p r s t)	II (i)	II (o)	II (u)	III (y)	IV (&)
I	p r s t	— / —	+	+	+	—	—
II	i	+					
	o	+		+		+	
	u	+					
III	y	—		+		+	
IV	&	—					+

Diagram 2.

(SpangHanssen1959)

Structuralist Tools

Table 8.
Vowel × binary final cluster (cf. sect. 84).

	ft	gt	ks	ds	vn	vl	dr!	mp	nk	ng	nd	nt	ns	lk	ld	lt	rk	rd	rt	rn	S	T	jC	
a	5	10	6	3	9	8	6	8	16	20	14	9	6	9	8	11	7	1	9	3	168	281	3	a
e	–	–	3	1	3	2	2	1	–	4	7	5	6	–	3	5	–	1	3	3	49	95	33	e
i	7	6	9	5	–	1	2	4	13	11	20	8	3	2	11	6	6	1	1	–	116	171	–	i
o	3	2	2	5	4	2	1	1	1	2	3	2	–	4	13	3	6	9	10	4	77	120	–	o
u	2	9	5	4	–	–	6	12	8	4	12	3	2	4	8	4	4	–	2	–	89	143	–	u
y	–	2	–	2	–	–	1	2	4	7	6	2	–	1	6	6	3	2	1	–	45	56	–	y
æ	4	11	1	–	4	4	2	2	9	11	8	1	3	2	11	4	6	6	6	4	99	145	–	æ
ø	5	2	–	–	1	4	–	–	–	–	1	2	3	–	–	–	3	–	1	6	28	47	10	ø
aa	–	–	–	1	–	–	1	–	–	–	4	–	–	–	–	–	–	2	–	1	9	11	–	aa
	26	42	26	21	21	21	21	30	51	59	75	32	23	22	60	39	35	22	33	21	680	1069	46	

(SpangHanssen1959)

Conclusions

- Neural language models (NLMs) are the implicit implementation of a theory of language
- Starting from an operational treatment of syntagmatic relations, NLMs infer paradigmatic relations, and explore commutation properties identifying paradigmatic and syntagmatic units at different levels, and dependencies between them.

Conclusions

- Neural language models (NLMs) are the implicit implementation of a theory of language
- Starting from an operational treatment of syntagmatic relations, NLMs infer paradigmatic relations, and explore commutation properties identifying paradigmatic and syntagmatic units at different levels, and dependencies between them.
- If NLMs can account for meaning, it is not because they are intelligent, but because meaning is the effect of (linguistic) structure and NLMs perform a structural analysis of language

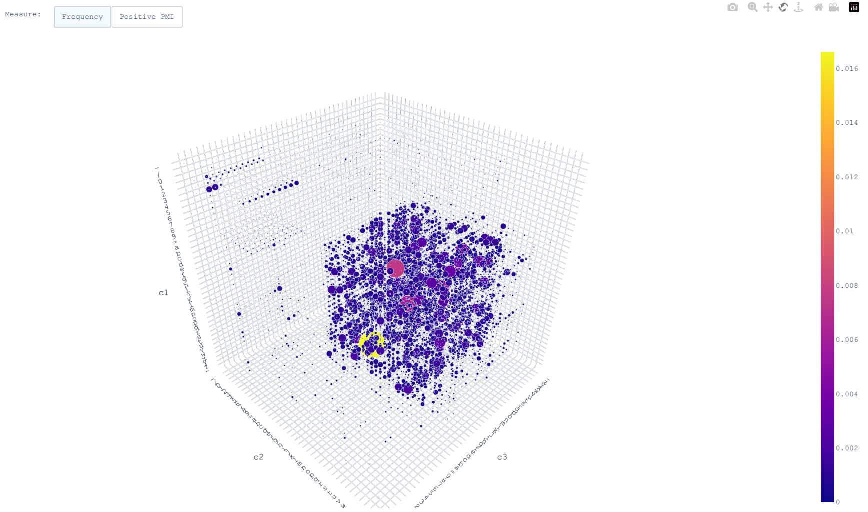
Conclusions

- Neural language models (NLMs) are the implicit implementation of a theory of language
- Starting from an operational treatment of syntagmatic relations, NLMs infer paradigmatic relations, and explore commutation properties identifying paradigmatic and syntagmatic units at different levels, and dependencies between them.
- If NLMs can account for meaning, it is not because they are intelligent, but because meaning is the effect of (linguistic) structure and NLMs perform a structural analysis of language
- Restituting the implicit structuralist grounding can provide interpretability and a reorientation of the research field

Challenges

- Politics of the corpus
- Non-cognitive philosophy and theory of language
- Integrated treatment of tokenization, embedding and attention
- Connection between distributional and structural features
- Treatment of long term dependencies
- Computability, tractability
- Generalization to non-linguistic corpora (semiology)

Generalization



Reference Papers

- J. L. Gastaldi. Why Can Computers Understand Natural Language?
  In: *Philosophy & Technology* 34.1 (2021), pp. 149–214.

- J. L. Gastaldi and L. Pellissier. The calculus of language: explicit representation of emergent linguistic structure through type-theoretical paradigms
  In: *Interdisciplinary Science Reviews* 46.4 (2021), pp. 569–590.

- T.-D. Bradley, J. L. Gastaldi, J. Terilla, The Structure of Meaning in Language: Moving from Linear Algebra to Category Theory
  Under review.

References I

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*.

Harris, Z. (1960). *Structural linguistics*. University of Chicago Press.

Hjelmslev, L. (1935). *La catégorie des cas*. Wilhelm Fink Verlag.

Hjelmslev, L. (1953). *Prolegomena to a theory of language*. Wawerly Press.

Hjelmslev, L. (1975). *Résumé of a Theory of Language*. Nordisk Sprog-og Kulturforlag.

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*.

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

Saussure. (1980). *Cours de linguistique générale*. Payot.

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.

Seminar “Understanding language”
Copenhagen Linguistic Circle
University of Copenhagen
Copenhagen, Denmark

Understanding Language (Models)

AI Language Models From a Structuralist Perspective

Juan Luis Gastaldi

ETH zürich

May 2nd, 2023

This project has received funding from the
European Union’s Horizon 2020 research and innovation programme
under grant agreement No 839730