# Non-identifiability of word embeddings, and connections to shape analysis

#### Simon Preston, Rachel Carrington, Karthik Bharath simon.preston@nottingham.ac.uk

University of Nottingham

December 14, 2019





• "Shape": info invariant to translation, scaling, rotation (+ reflection)



"Configuration"

(Assume V is centred: 
$$V1 = 0$$
)

• "Shape": info invariant to translation, scaling, rotation (+ reflection)



"Configuration"

(Assume V is centred: V1 = 0)

• "Shape": info invariant to translation, scaling, rotation (+ reflection)

• Can identify shape as  $[V] = \{ c \mathbf{Q} V : c \in \mathbb{R}^+ ; \mathbf{Q} \in O(m) \}$ 



"Configuration"

(Assume V is centred: V1 = 0)

• "Shape": info invariant to translation, scaling, rotation (+ reflection)

- Can identify shape as  $[V] = \{ c \mathbf{Q} V : c \in \mathbb{R}^+ ; \mathbf{Q} \in O(m) \}$
- "Shape function":  $g(\cdot)$  such that  $g(\mathbf{V}) = g(c\mathbf{QV})$ .

	abide 0 0 	2 apsence 2 1 3	2 2 absent	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Re 1 0 0 0	prese	doc 1 doc 2 doc 3 doc 4 doc 5	, X
7				$\downarrow$	Wor	d em	bedding	g, <b>V</b>
	₹ - N -		coa players	ch				
		score match	refe	ree comp	outer <sup>screel</sup>	device app	s	
	4 -	-4	-2	0	2	4 6		

#### Text data

ALLER NON FEI INCOLUNT PROPER FEINER INNER CAPTER I THEFT IS THE TAXE WHEN ALL AND TAXE AND ADDRESS OF THE

Mr. Anal Wagers, So., London, Departul instant Autor Wilson tide, providing, we billion that unclearly appendixe.

wells) as time for soften, in the same 1 often losse 1 the system, and must denot tracks and losse all day, in the day of a same tracket body and in the close pill will offer any, fright, mean in a site lags, and off, which are there is a site in the set of the soften losses in deal gave the time will instance losses.

b. as both as any investor monty, as follow presentative and the holds, both because 11 houses any end of an element of a shared on a subsequence of an element of an element of the subsequence of the state of the state of the sub-former of the sub-former of a state of the state of the origination of the state of the state of the state of a state of the origination of the state of the origination of the state of the

#### the same an and the second of the second. National Social Social

INFI PERS

oter on buse, which a buserie after the addication of significant regions." They are, the post any deal which addicate is part, " they are on the set of the set of

#### Rise of Word Embedding Models

TITLE	CITED BY	YEAR
Distributed representations of words and phrases and their compositionality T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean Neural information processing systems	16538	2013
Efficient estimation of word representations in vector space T Mikolov, K Chen, G Corrado, J Dean arXiv preprint arXiv:1301.3781	13414	2013
Glove: Global vectors for word representation J Pennington, R Socher, C Manning Proceedings of the 2014 conference on empirical methods in natural language	10790	2014

#### Rise of Word Embedding Models

TITLE	CITED BY	YEAR	Cited by 16538
Distributed representations of words and phrases and their compositionality T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean Neural Information processing systems	16538	2013	2014 2015 2016 2017 2018 2019 Cited by 13414
Efficient estimation of word representations in vector space T Mikolov, K Chen, G Corrado, J Dean arXiv preprint arXiv:1301.3781	13414	2013	2014 2015 2016 2017 2018 2019
Glove: Global vectors for word representation J Pennington, R Socher, C Manning Proceedings of the 2014 conference on empirical methods in natural language	10790	2014	Cited by 10790

#### Rise of Word Embedding Models

TITLE	CITED BY	YEAR	Cited by 16538
Distributed representations of words and phrases and their compositionality T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean Neural Information processing systems	16538	2013	2014 2015 2016 2017 2018 2019 Cited by 13414
Efficient estimation of word representations in vector space T Mikolov, K Chen, G Corrado, J Dean arXiv preprint arXiv:1301.3781	13414	2013	2014 2015 2016 2017 2018 2016
Glove: Global vectors for word representation J Pennington, R Socher, C Manning Proceedings of the 2014 conference on empirical methods in natural language	10790	2014	Cited by 10790

#### Invariance and identifiability issues for word embeddings

2019

R Carrington, K Bharath, S Preston Advances in Neural Information Processing Systems, 15114-15123

- Word embedding **V** encodes word meaning.
- Used for, and evaluated on, word tasks.
- Word similarity: "given word A, how similar is word B?"
- Word association: "A is to B as C is to what?", e.g. Paris is to France as Madrid is to ...?
- Task performance measured by  $g(data, \mathbf{V})$ .

Simple word embedding model: take  ${\boldsymbol{\mathsf{V}}}$  as minimiser of

$$\|\mathbf{X} - \mathbf{U}\mathbf{V}\|^2 = \sum_{ij} \left( x_{ij} - \mathbf{u}_i^ op \mathbf{v}_j 
ight)^2$$

Simple word embedding model: take  $\boldsymbol{V}$  as minimiser of

$$\|\mathbf{X} - \mathbf{U}\mathbf{V}\|^2 = \sum_{ij} \left(x_{ij} - \mathbf{u}_i^{ op}\mathbf{v}_j
ight)^2$$

Solution for V: write SVD of  $\mathbf{X} = \mathbf{A} \mathbf{\Sigma} \mathbf{B}^{\top}$ . Then  $\|\mathbf{X} - \mathbf{X}_d\|$  is minimised by  $\mathbf{X}_d = \mathbf{A}_d \mathbf{\Sigma}_d \mathbf{B}_d^{\top}$ , so take

$$\mathbf{U}^* = \mathbf{A}_d, \quad \mathbf{V}^* = \mathbf{\Sigma}_d \mathbf{B}_d^\top.$$

Simple word embedding model: take  ${\boldsymbol{\mathsf{V}}}$  as minimiser of

$$\|\mathbf{X} - \mathbf{U}\mathbf{V}\|^2 = \sum_{ij} \left(x_{ij} - \mathbf{u}_i^{ op}\mathbf{v}_j
ight)^2$$

Solution for V: write SVD of  $\mathbf{X} = \mathbf{A} \mathbf{\Sigma} \mathbf{B}^{\top}$ . Then  $\|\mathbf{X} - \mathbf{X}_d\|$  is minimised by  $\mathbf{X}_d = \mathbf{A}_d \mathbf{\Sigma}_d \mathbf{B}_d^{\top}$ , so take

$$\mathbf{U}^* = \mathbf{A}_d, \quad \mathbf{V}^* = \mathbf{\Sigma}_d \mathbf{B}_d^\top.$$

... or  $\mathbf{V}^* = \mathbf{\Sigma}_d^{1-lpha} \mathbf{B}_d^{ op}$  ?

Simple word embedding model: take  ${\boldsymbol{\mathsf{V}}}$  as minimiser of

$$\|\mathbf{X} - \mathbf{U}\mathbf{V}\|^2 = \sum_{ij} \left(x_{ij} - \mathbf{u}_i^{ op}\mathbf{v}_j
ight)^2$$

Solution for V: write SVD of  $\mathbf{X} = \mathbf{A} \mathbf{\Sigma} \mathbf{B}^{\top}$ . Then  $\|\mathbf{X} - \mathbf{X}_d\|$  is minimised by  $\mathbf{X}_d = \mathbf{A}_d \mathbf{\Sigma}_d \mathbf{B}_d^{\top}$ , so take

$$\mathbf{U}^* = \mathbf{A}_d, \quad \mathbf{V}^* = \mathbf{\Sigma}_d \mathbf{B}_d^\top.$$

 $\ldots$  or  $\mathbf{V}^* = \mathbf{\Sigma}_d^{1-lpha} \mathbf{B}_d^{ op}$  ?  $\ldots$  or  $\mathbf{V}^* = \mathbf{C} \mathbf{B}_d^{ op}$  for any  $\mathbf{C} \in \mathit{GL}(d)$  ?!

Simon Preston (University of Nottingham)

Simple word embedding model: take  ${\boldsymbol{\mathsf{V}}}$  as minimiser of

$$\|\mathbf{X} - \mathbf{U}\mathbf{V}\|^2 = \sum_{ij} (x_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)^2 = f(\mathbf{X}, \mathbf{U}\mathbf{V})$$

Solution for V: write SVD of  $\mathbf{X} = \mathbf{A} \mathbf{\Sigma} \mathbf{B}^{\top}$ . Then  $\|\mathbf{X} - \mathbf{X}_d\|$  is minimised by  $\mathbf{X}_d = \mathbf{A}_d \mathbf{\Sigma}_d \mathbf{B}_d^{\top}$ , so take

$$\mathbf{U}^* = \mathbf{A}_d, \quad \mathbf{V}^* = \mathbf{\Sigma}_d \mathbf{B}_d^\top.$$

... or  $\mathbf{V}^* = \mathbf{\Sigma}_d^{1-lpha} \mathbf{B}_d^{ op}$  ?

... or 
$$\mathbf{V}^* = \mathbf{C}\mathbf{B}_d^ op$$
 for any  $\mathbf{C} \in GL(d)$  ?!

Non-identifiability:  $f(\mathbf{X}, \mathbf{UV}) = f(\mathbf{X}, \mathbf{UC}^{-1}\mathbf{CV})$ 

Simon Preston (University of Nottingham)

#### Non-identifiability of different embedding models

$$\begin{aligned} \mathbf{LSA:} & \sum_{ij} \left( \mathbf{x}_{ij} - \mathbf{u}_i^{\top} \mathbf{v}_j \right)^2 \\ \text{word2vec:} & \sum_{ij} \log \left( \sigma \left( \mathbf{u}_i^{T} \mathbf{v}_j \right) \right) + k \cdot \frac{\sum_l x_{il} \sum_m x_{mj}}{\sum_{ij} x_{ij}} \log \left( \sigma \left( -\mathbf{u}_i^{T} \mathbf{v}_j \right) \right) \\ \mathbf{GloVe:} & \sum_{ij} h\left( x_{ij} \right) \left( \mathbf{u}_i^{T} \mathbf{v}_j - h_1\left( x_{ij} \right) \right)^2 \end{aligned}$$

#### Non-identifiability of different embedding models

$$\begin{aligned} \mathbf{LSA:} & \sum_{ij} \left( x_{ij} - \mathbf{u}_i^{\top} \mathbf{v}_j \right)^2 \\ \text{word2vec:} & \sum_{ij} \log \left( \sigma \left( \mathbf{u}_i^{T} \mathbf{v}_j \right) \right) + k \cdot \frac{\sum_l x_{il} \sum_m x_{mj}}{\sum_{ij} x_{ij}} \log \left( \sigma \left( -\mathbf{u}_i^{T} \mathbf{v}_j \right) \right) \\ \mathbf{GloVe:} & \sum_{ij} h\left( x_{ij} \right) \left( \mathbf{u}_i^{T} \mathbf{v}_j - h_1\left( x_{ij} \right) \right)^2 \end{aligned}$$

Each is such that for any particular solution

$$\mathbf{V}^* = \mathop{\arg\min}\limits_{\mathbf{V}} f(\mathbf{X}, \mathbf{UV})$$

a general solution set is

$$\{\mathbf{V}: \mathbf{V} = \mathbf{C}\mathbf{V}^*, \mathbf{C} \in \mathrm{GL}(r)\}$$

Simon Preston (University of Nottingham)

### Non-identifiability - implications? (1)

#### Word similarity task:



$$\cos(\mathbf{v}_i,\mathbf{v}_j) = \mathbf{v}_i^\top \mathbf{v}_j / (\|\mathbf{v}_i\| \cdot \|\mathbf{v}_j\|)$$

## Non-identifiability - implications? (1)

#### Word similarity task:



$$\cos(\mathbf{v}_i, \mathbf{v}_j) = \mathbf{v}_i^\top \mathbf{v}_j / (\|\mathbf{v}_i\| \cdot \|\mathbf{v}_j\|)$$

... then measure "embedding performance" by

$$g(\mathsf{data}, \mathbf{V}) = \mathsf{corr}(\{y_i\}, \{z_i\})$$

### Non-identifiability - implications? (2)

Word analogy task:

• Paris is to France as Madrid is to ...? Given **V** solve

 $\arg\max_{i} \left[ \cos(\mathbf{v}_{i}, \mathbf{v}_{\text{"France"}}) - \cos(\mathbf{v}_{i}, \mathbf{v}_{\text{"Paris"}}) + \cos(\mathbf{v}_{i}, \mathbf{v}_{\text{"Madrid"}}) \right]$ 

- The data are a set of human-chosen analogies.
- Performance metric  $g(data, \mathbf{V})$  is the proportion correct.

### Non-identifiability - implications? (2)

Word analogy task:

 $\bullet\,$  Paris is to France as Madrid is to  $\ldots ?\,$  Given  ${\bf V}$  solve

 $\arg\max_{i} \left[ \cos(\mathbf{v}_{i}, \mathbf{v}_{\text{"France"}}) - \cos(\mathbf{v}_{i}, \mathbf{v}_{\text{"Paris"}}) + \cos(\mathbf{v}_{i}, \mathbf{v}_{\text{"Madrid"}}) \right]$ 

- The data are a set of human-chosen analogies.
- Performance metric  $g(\text{data}, \mathbf{V})$  is the proportion correct.

For both similarity and analogy, g depends on  $\mathbf{V}$  only via  $\cos(\mathbf{v}_i, \mathbf{v}_j)$ . Hence  $g(\text{data}, \mathbf{V}) = g(\text{data}, c\mathbf{Q}\mathbf{V}) \Rightarrow g$  is a shape function.

#### Mis-match of invariances

Training objective  $f(\mathbf{X}, \mathbf{UV})$  invariant to  $\mathbf{V} \mapsto \mathbf{CV}$ . Test objective  $g(\text{data}, \mathbf{V})$  invariant to  $\mathbf{V} \mapsto c\mathbf{QV}$ 

$$\begin{split} f\left(\boldsymbol{X},\boldsymbol{U}\boldsymbol{V}\right) &= f\left(\boldsymbol{X},\boldsymbol{U}\boldsymbol{C}^{-1}\boldsymbol{C}\boldsymbol{V}\right), \qquad \boldsymbol{C} \in \mathrm{GL}(r) \\ g\left(\boldsymbol{D},\boldsymbol{V}\right) &= g\left(\boldsymbol{D},c\boldsymbol{Q}\boldsymbol{V}\right), \qquad \boldsymbol{Q} \in \mathrm{O}(d), c \in \mathbb{R} \end{split}$$

What is the set  $\mathcal{F}_d \subset GL(d)$  which leaves f invariant but not g?

#### Mis-match of invariances

Training objective  $f(\mathbf{X}, \mathbf{UV})$  invariant to  $\mathbf{V} \mapsto \mathbf{CV}$ . Test objective  $g(\text{data}, \mathbf{V})$  invariant to  $\mathbf{V} \mapsto c\mathbf{QV}$ 

$$\begin{split} f\left(\boldsymbol{X},\boldsymbol{U}\boldsymbol{V}\right) &= f\left(\boldsymbol{X},\boldsymbol{U}\boldsymbol{C}^{-1}\boldsymbol{C}\boldsymbol{V}\right), \qquad \boldsymbol{C} \in \mathrm{GL}(r) \\ g\left(\boldsymbol{D},\boldsymbol{V}\right) &= g\left(\boldsymbol{D},c\boldsymbol{Q}\boldsymbol{V}\right), \qquad \boldsymbol{Q} \in \mathrm{O}(d), c \in \mathbb{R} \end{split}$$

What is the set  $\mathcal{F}_d \subset GL(d)$  which leaves f invariant but not g?

Write  $\mathcal{F}_d = \tilde{\mathcal{F}}_d - c\mathcal{I}$ , where

- *˜*<sub>d</sub> = GL(d)\O(d), and can be identified with UT(d), upper tringular matrices with +ve diag elements. (Intuition: QR decomposition of C)
- $c\mathcal{I} = \{cl_d : c \in \mathbb{R}\}$  is set of scale transformations
- dimension of  $\mathcal{F}_d$  is d(d-1)/2 1.

• "When all methods are allowed to tune a similar set of hyperparameters their performance is largely comparable"<sup>1</sup>

<sup>1</sup>Levy, Goldberg, Dagan, *Trans. Assoc. Comput. Ling.*, 2015 Simon Preston (University of Nottingham)

- "When all methods are allowed to tune a similar set of hyperparameters their performance is largely comparable"<sup>1</sup>
- Some hyperparameters index different elements of solution set *f*, chosen for performance in *g*, e.g. V<sup>\*</sup> = Σ<sup>1-α</sup><sub>d</sub>B<sup>T</sup><sub>d</sub>
- *f* typically optimised by Monte Carlo (different solns explained as local optima but also due to non-identifiability) then soln chosen for *g*.

<sup>&</sup>lt;sup>1</sup>Levy, Goldberg, Dagan, *Trans. Assoc. Comput. Ling.*, 2015 Simon Preston (University of Nottingham)

- "When all methods are allowed to tune a similar set of hyperparameters their performance is largely comparable"<sup>1</sup>
- Some hyperparameters index different elements of solution set *f*, chosen for performance in *g*, e.g. V<sup>\*</sup> = Σ<sup>1-α</sup><sub>d</sub>B<sup>T</sup><sub>d</sub>
- *f* typically optimised by Monte Carlo (different solns explained as local optima but also due to non-identifiability) then soln chosen for *g*.
- Both are (implicitly) supervised approaches.

<sup>1</sup>Levy, Goldberg, Dagan, *Trans. Assoc. Comput. Ling.*, 2015 Simon Preston (University of Nottingham)  $\operatorname{arg\,min}_{\mathbf{V}} f(\mathbf{X}, \mathbf{UV})$ 

Identifying a solution unique up to orthogonal transformations:

Impose constraint VV<sup>⊤</sup> = I, then for any solution V\* any other solution CV\* for C ∈ GL(d) satisfies g(data, CV\*) = g(data, V\*).

 $\operatorname{arg\,min}_{\mathbf{V}} f(\mathbf{X}, \mathbf{UV})$ 

Identifying a solution unique up to orthogonal transformations:

Impose constraint VV<sup>⊤</sup> = I, then for any solution V\* any other solution CV\* for C ∈ GL(d) satisfies g(data, CV\*) = g(data, V\*).

Identifying a unique solution:

Additionally impose: (i) U<sup>⊤</sup>U = I, (ii) diag(U<sup>⊤</sup>U) decreasing, (iii) positive first non-zero elements of each col of U.

#### Sensitivity to particular solution



#### Sensitivity to particular solution



#### Outlook: dynamic embedding



#### Text data



The track ray or table among transmission we does not convert on a shear of the table of the table of the table of the table beam of the table of the does not be table of the table of the table of the table of the does not be table of the table of the table of the table of the of the table of the table of the table of the table of the does not be table of the table of the table of the table of the does not be table of the table of the table of the table of the source of the table of the table of the table of table of table parts of the table of the table of the table of table of table of the table of the table of the table of the table of the table of table of

Rep 12, 167. Josef Wopper, Boy, 2133, A.M. Departul, Colomology, Multi-Molecul Holico, presslep, We Sciences restations contends agreed to:-

View Has assumption has based and, with failuring of presented approximation, we considerable approximation press instructional for mean failured that a standard and approximate approximation approximation for approximation of the standard of the standard approximation for approximation approximation approximation approximation approximation for approximation approximation approximation approximation approximation for approximation approximation

That while this meaning is dealy sensible of the schemages which easy access to the cases of scippes, free the schemages is show they have any address or these the two the source of sensitive of sensitivity, they, ALAMALL, and Alama Spectra Sciences and Sciences, they can be bettern.



Simple is a general () is to take for withing, by the same () where bottom is the set service, and want about 1400 states and () dec. for some dataset for the services, and want about 1400 states 1400 states () dec. bottom () and () an

It have follow presentially april that hadds, while humans it is forward and approximate and approximate present approximate of approximate in the distance of a forward spectrum of them applies to its sectorizations and above of branch spectrum of them applies to the approximate have been applied and the spectrum coupled in the laptic of a single of the original spectrum of approximate have coupled in the laptic of a single of the its spectrum of applies the spectrum of the spectrum of the spectrum of a spectrum of the spectrum.

The constant particle is not 0.5%. The mass resting the families of the set of the seto



t<sub>2</sub>

## **Invariance and identifiability issues for word embeddings.** Rachel Carrington, Karthik Bharath & Simon Preston. *NeurIPS*, 2019.



