Static models: A (relevant) message from the past

Haim Dubossarsky, h.dubossarsky@qmul.ac.uk

Queen Mary University of London

The Alan Turing Institute

UNIVERSITY OF CAMBRIDGE

Hello again!

CHANGE IS KEY.ORG

Haim Dubossarsky

Researcher @ Change is Key!
Queen Mary University of London (QMUL)
Language Technology Lab, University of Cambridge
The Alan Turing Institute
h.dubossarsky@qmul.ac.uk

Outline & takehomes

old ≠ not useful

- Brief recap on static models (an introduction to some?)
  - **Static models are not contextualized models**
  - Explicit models: Count-based, PPMI (interpretable dimensions)
  - Predictive models: word2vec and its likes (“opaque” dimensions)
- Doing **semantic change** with static models: features, pros and cons
  - 1-word : all meanings – polysemy
  - Measure change in meaning via cosine-distance
  - Can work well with small corpora
  - Some models provide much more detailed report of change
- Does not cover all models in this overview: e.g., topic models

Count-based models

Simple co-occurrence models within a context window

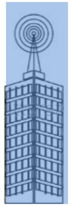
Very sparse

CONTEXT	ALL	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
COMPUTER	171						1	75	67	28
BIG	170	2	26	8	2	7	31	43	29	22
PIE	148	4	3	22	18	16	29	31	17	8
TREE	94	6	18	16	17	15	9	9	1	3
ADAM	59	8	12	9	7	9	8	3	3	
JOBS	57						1	14	23	19
TREES	51	4	10	18	5	6	3	4	1	
JUICE	45	1	2		3	5	9	7	11	7
MACINTOSH	43							24	13	6
IBM	41							24	17	

Taken from https://corpus.byu.edu/

Count-based models

Years



Context words

CONTEXT	ALL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
RADIO	454									1			15	42	98	69	52	50	54	38	35
TELEVISION	161												4	3	6	28	15	24	33	29	19
NEWS	161					1			1			3	3	7	27			30	19	18	20
STATION	115												4	11	35			10	9	9	5
STATIONS	111												5	13	23			7	12	10	12
TV	79																	7	11	28	19
LIVE	77																	15	18	16	18
NETWORK	61													5	14			6	5	10	14
SPEECH	61												2	18	16			4	4	2	2
PROGRAM	60												1	15	15			7	3	3	1
SCATTERED	44				3	5			7	4	7	5	2	1	1			1			
TONIGHT	43												3	5	15			1	1	3	1
ADDRESS	41												4	11	15			1	4		1
THROUGHOUT	37				1	2					2		7	6	1			3	2	2	3
LOCAL	37										1		4					5	3	8	5
NBC	36												1	8	5			5	4	8	4
MEDIA	36																	4	6	10	11
SUNDAY	36													5	14			1	3	1	1
WEEKLY	34												1	4	14			1	2	3	1
SPREAD	34				2				3	2	7	12	4					1			1
MOSCOW	31												1	3	10	9	4	2	1		1
MATERIAL	31												1	1	20	1	1	4		3	
CABLE	30												1						6	12	11
MESSAGE	30													4	7	5	3	6	3		2
SOWN	29			1	2	2	7	1	2	7	2	1	2		1	1					

$\vec{w_i}$=broadcast

$\vec{w_i}$=broadcast

Count-based models

- Very rare: Most cases will not be so clear

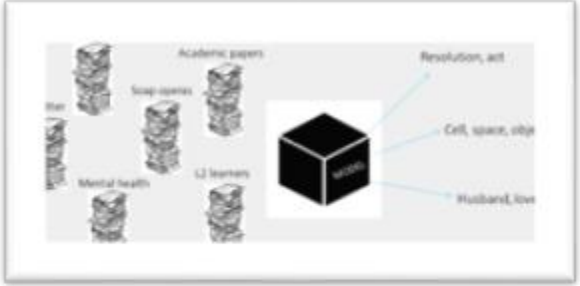
Years

Context words

CONTEXT	ALL	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
COMPUTER	171						1	75	67	28
BIG	170	2	26	8	2	7	31	43	29	22
PIE	148	4	3	22	18	16	29	31	17	8
TREE	94	6	18	16	17	15	9	9	1	3
ADAM	59	8	12	9	7	9	8	3	3	
JOBS	57						1	14	23	19
TREES	51	4	10	18	5	6	3	4	1	
JUICE	45	1	2		3	5	9	7	11	7
MACINTOSH	43							24	13	6
IBM	41							24	17	

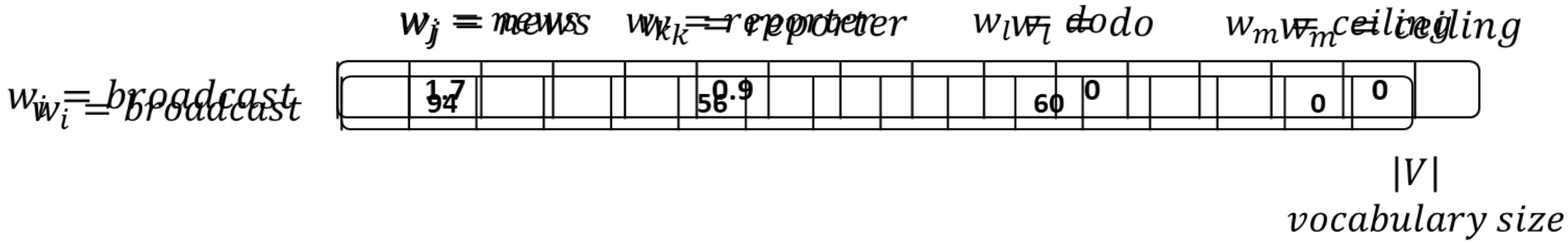
- Highly dependent and reflects the meaning of the corpus/domain
  - True for **all static & contextualized** models
  - More apparent here as static models are not “pre-trained”

- **Problem**: Highly skewed for frequent collocates
  - Prepositions, function words (stopwords)
  - Solution: ????



Positive Pointwise Mutual Information (PPMI)

- Co-occurrence models within a context window **with a twist**
  - **Twist**: Mutual information measures the strength of association between the target word and its co-occurring words



- **Learn associativity by informativity**

Positive Pointwise Mutual Information (PPMI)

- Only “strong” co-occurring words are retained, hence “positive” PMI

$$\mathrm{PPMI}(w,c) = \max(\log_2 \frac{P(w,c)}{P(w)P(c)}, 0)$$

Count(w,context)

	computer	data	pinch	result	sugar
apricot					
apple					
digital					
informatio					

p(w,context)

	computer	data	pinch	result	sugar	p(w)
apricot						
apple						
digital						
informatio						
p(context)						

PPMI(w,context)

	computer	data	pinch	result	sugar
appricot	-	-	2.25	-	2.25
apple	-	-	2.25	-	2.25
digital	1.66	0.00	-	0.00	-
information	0.00	0.57	-	0.47	-

From Speech and Language Processing (3rd ed.)

Advantages of explicit models (count-based & PPMI)

$w_j = news$ $w_k = reporter$ $w_l = do$ $w_m = ceiling$

$w_i = broadcast$

	1.7				0.9					0				0	

- Enables a finer analysis of change (association level)
  - Used in research: Stefanowitsch & Gries Collostructions Analysis (2003)

PPMI for ***prime***^{1990}	
ministe	
suspe	
cut	
numbe	

PPMI for ***prime***^{2010}	
minister	11.26
numbers	9.51
cut	10.1

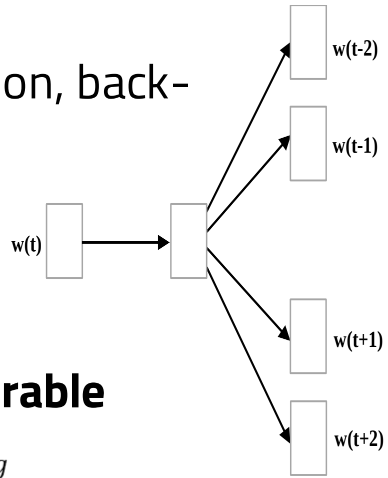
PPMI for ***heart***^{medical}	
attack	
chest	
pacemaker	

PPMI for ***heart***^{standard}	
attack	13.4
emotion	**4.9**
central	**4.5**
warmth	**3.2**

Predictive models (word2vec)

- Word2vec (Mikolov et al. 2013) is a Neural Network model
  - Shallow network: 1 layer
  - Uses known NN machinery: MLM, objective function, back-propogation, SGD, etc.

$$J_\theta = \frac{1}{T} \sum_{t=1}^{T} \sum_{-n \leq j \leq n, \neq 0} \log p\left(w_{t+j} \mid w_t\right)$$



- Vectors are now opaque & **vector spaces are incomparable**

		$w_j = news$				$w_k = reporter$					$w_l = do$				$w_m = ceiling$	
$w_i = broadcast$		1.7				0.9					0				0	$\lvert V \rvert$

		$?$			$??$		$???$		
$w_i = broadcast$	...	...	...	...	...	...	...	...	$\lvert d \rvert$

Predictive models (word2vec)

Even worse models are sometimes better

- Why word2vec is more popular than PPMI?
  - Easier and more efficient implementation
  - PR: Nice demonstration of abilities (analogy solving etc.)
  - Simply because of sheer numbers of users

- Is word2vec better than PPMI? Sometimes, but often not.

Method	WordSim Similarity	WordSim Relatedness	Bruni et al. MEN	Radinsky et al. M. Turk	Luong et al. Rare Words	Hill et al. SimLex	Google Add / Mul	MSR Add / Mul
PPMI	.755	**.697**	.745	.686	.462	.393	.553 / .679	.306 / .535
SVD	**.793**	.691	**.778**	.666	**.514**	.432	.554 / .591	.408 / .468
SGNS	**.793**	.685	.774	**.693**	.470	**.438**	.676 / **.688**	.618 / **.645**

From Levy et. al. 2015

  - Word2vec-like models are mathematically equivalent to PPMI (Levy et. al., 2014, 2015)

- Not the right question: **Is word2vec better for Semantic change?**

Reminder: measuring change computationaly

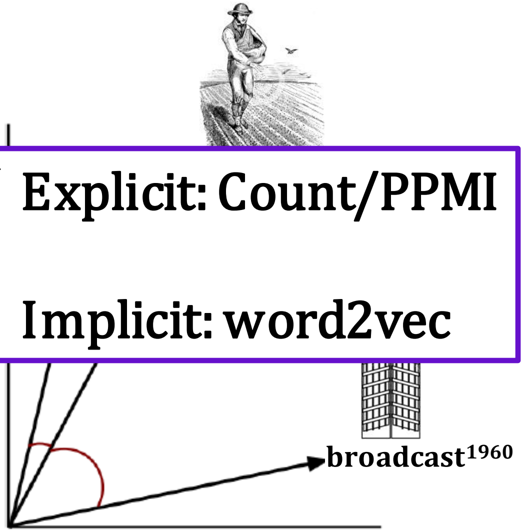
$w_i = field$

X^{1880}

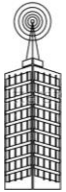
$w_j = broadcast$

$w_k = news$

$|\mathbb{V}| \times |\mathbb{V}|$



- ✓ Explicit: Count/PPMI
- × Implicit: word2vec



$w_i = field$

Y^{1960}

$w_j = broadcast$

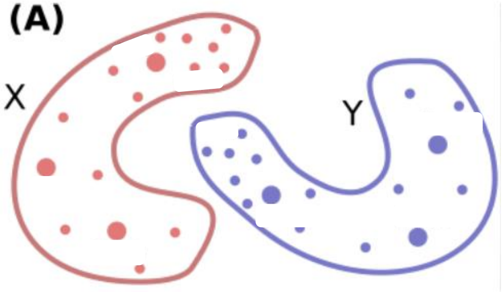
$w_k = news$

$|\mathbb{V}| \times |\mathbb{V}|$

$$\boldsymbol{Change} = \frac{\overrightarrow{\boldsymbol{w}}^{t1} \cdot \overrightarrow{\boldsymbol{w}}^{t2}}{||\overrightarrow{\boldsymbol{w}}^{t1}|| \cdot ||\overrightarrow{\boldsymbol{w}}^{t2}||}$$

Lexical semantic change with w2v-like models

- Word2vec models are initiated with random parameters. Hence, **if we don't do something about it**, their vectors lie in difference spaces, and are incomparible.



From Conneau et al. 2018

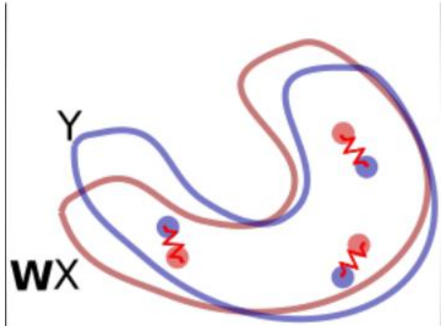
- Solutions:
  - Aligning the vector spaces prior to comparison
  - Avoiding the need for alignment

Aligning vector spaces

We need to find $\boldsymbol{\varphi}(\boldsymbol{X}) \rightarrow \boldsymbol{Y}$

$W^* = \ argmin_w \|WX \ - Y\|_2$

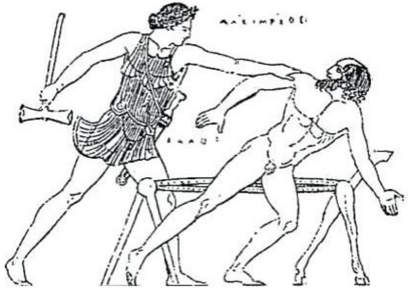
$X, Y \epsilon R^d$



Under orthogonal constraint ($W^T W = I$) the solutions is:

$U\Sigma V^T = \ SVD(YX^T)$
$W = UV^T$

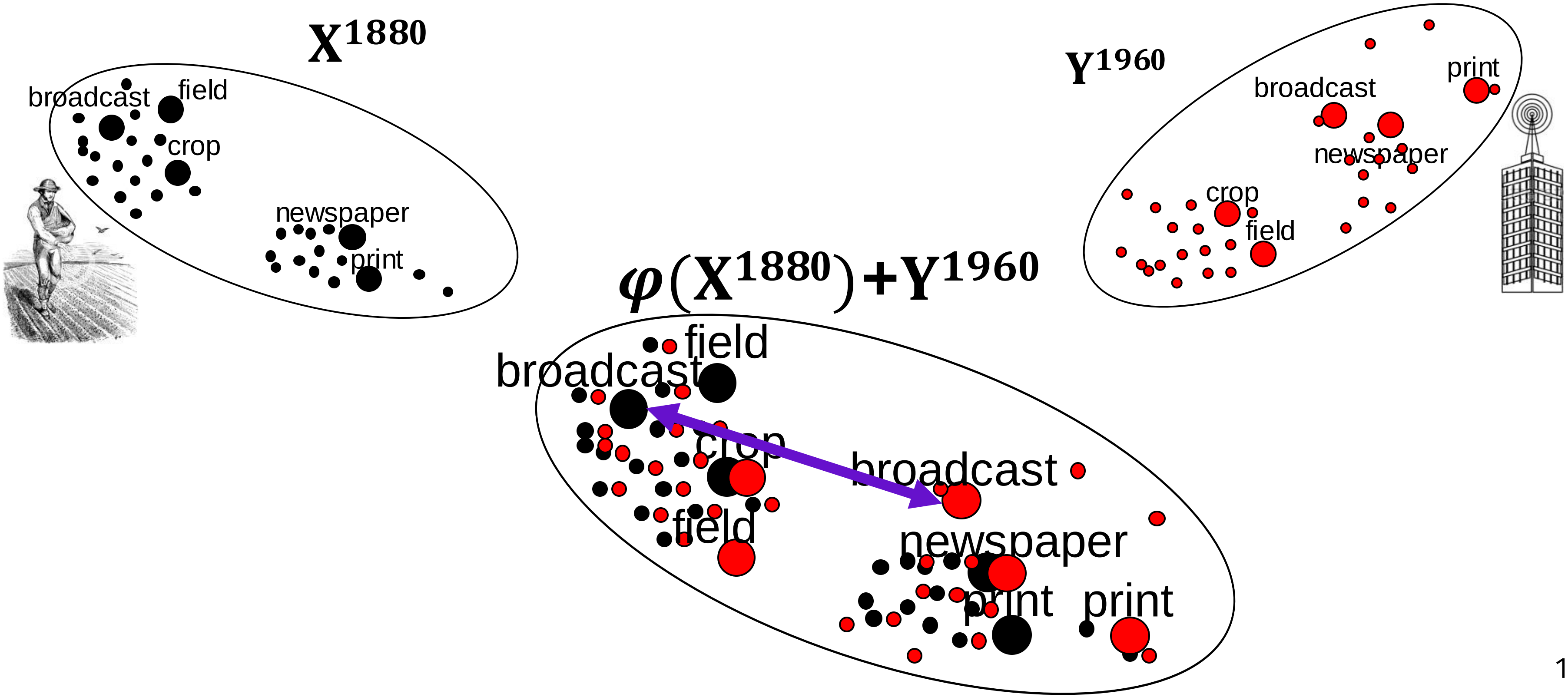
Procrustes



<u>Assumptions</u>
We have 1:1 mapping (dictionaries)
Vector spaces are <u>comparable</u> (isometric)

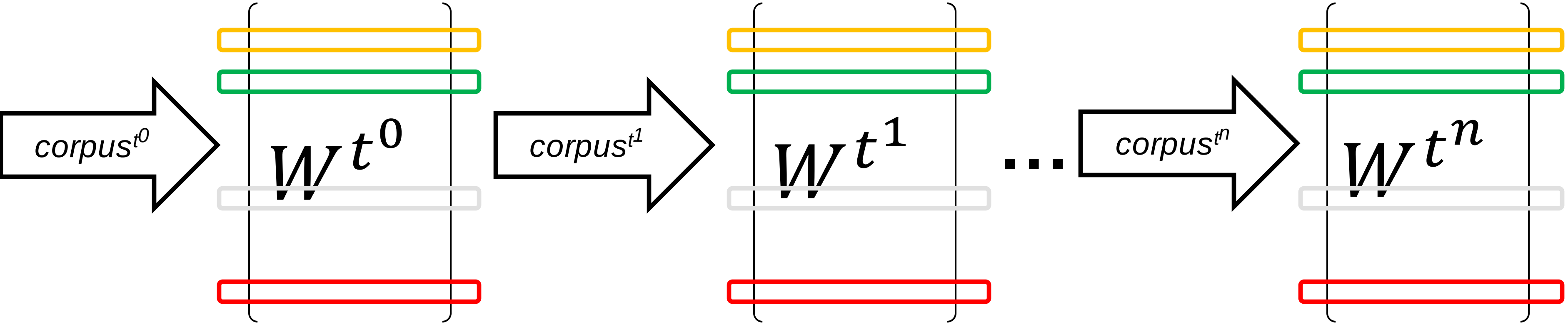
Aligning vector spaces

Alignment is not noise free.
What is the nature of the noise?



Avoiding alignment I

- Incremental training (Kim et al., 2014)
  - For every time step, model is initiated with the parameters of the trained model from the previous step.
  - Causes drift (noise) for the entire vector space



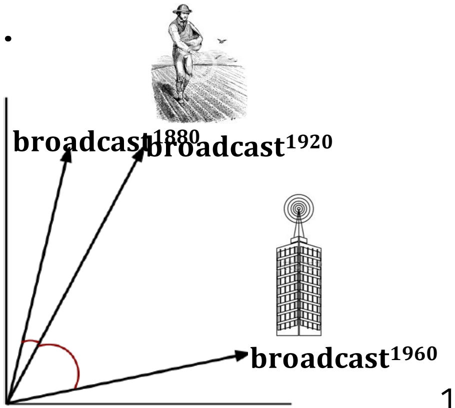
Avoiding alignment II

- Temporal referencing (Dubossarsky et al. 2019)
  - Words are tagged according to the time of corpus
  - Observed the least amount of noise

<u>Example</u>

`Silken cauliflowers sown broadcast`^{1870} `over the land.`

`The dramatic broadcast`^{1970} `stunned the nation.`



Take homes

old ≠ not useful

- Brief recap on static models (an introduction to some?)
  - **Static models are not contextualized models**
  - Explicit models: Count-based, PPMI (interpretable dimensions)
  - Predictive models: word2vec and its likes (“opaque” dimensions)
- Doing **semantic change** with static models: features, pros and cons
  - 1-word : all meanings – polysemy
  - Measure change in meaning via cosine-distance
  - Can work well with small corpora
  - Some models provide much more detailed report of change