Course Pack: Meaningful Text Analysis with Word Embeddings, Digital Humanities Summer Institute, 2021

Jonathan Reeve

April 27, 2021

Contents

Chapter 1

About this Course Pack

This course pack contains papers and readings that may be of interest to students in the course, "Meaningful Text Analysis with Word Embeddings," a one-week course at the Digital Humanities Summer Institute, during the summer of 2021. Many of these are originally meant for students in STEM fields, so you may encounter notation or terminology that will seem alien. Don't let that discourage you! In this course, we will not need most of the mathematical details you see in these papers, since we will be using libraries and software packages that have already implemented these details for us. That said, if you're curious about the inner workings of the algorithms we'll be using, or have a background in computer science, these details can be useful!

Contained here is a chapter, and accompanying slides, from the classic textbook from Jurafski and Martin, *Speech and Language Processing*. Following that is Mikolov et al's seminal paper describing Word2Vec, one of the most well-known embedding libraries. Finally, there are two papers which deal with potential bias and ethics issues associated with word embeddings analysis.

Find all these readings, and more, at the course website: https://dhsi2021.jonreeve.com, which is the canonical location for course materials.

Chapter 2

Jurafsky's Slides

Slides for Chapter 6 of Jurafski, Dan, and James H. Martin. *Speech and Language Processing*. Third edition draft. https://web.stanford.edu/~jurafsky/slp3/

Vector Semantics & Embeddings

Word Meaning

What do words mean?

Introductory logic classes:

- The meaning of "dog" is DOG; cat is CAT

$$\forall x \, \text{DOG}(x) \longrightarrow \text{MAMMAL}(x)$$

Old joke by Barbara Partee:

- Q: What's the meaning of life?
- A: LIFE

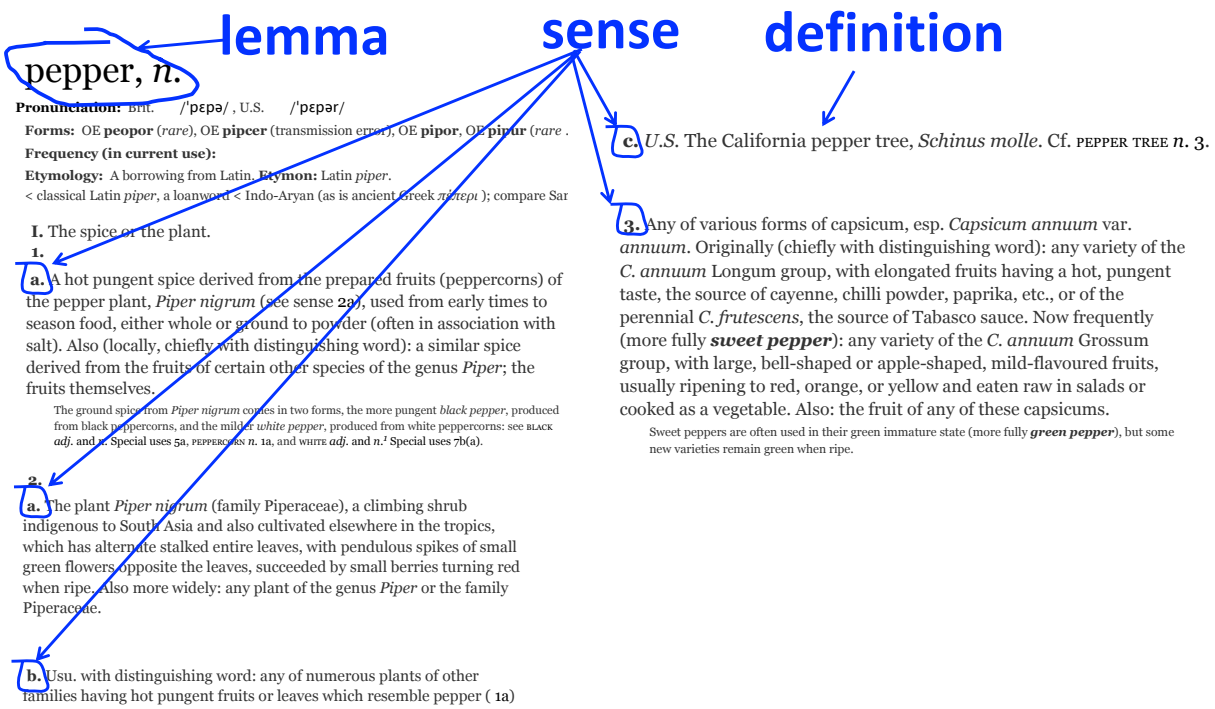
That seems unsatisfactory!

What do words mean?

Next thought: look in a dictionary

http://www.oed.com/

Words, Lemmas, Senses, Definitions



Lemma pepper

Sense 1: spice from pepper plant

Sense 2: the pepper plant itself

Sense 3: another similar plant (Jamaican pepper)

Sense 4: another plant with peppercorns (California pepper)

Sense 5: *capsicum* (i.e. chili, paprika, bell pepper, etc)

A sense or "concept" is the meaning component of a word

Relations between senses: Synonymy

Synonyms have the same meaning in some or all contexts.

- filbert / hazelnut
- couch / sofa
- big / large
- automobile / car
- vomit / throw up
- water / H_20

Relation: Synonymy

Note that there are probably no examples of perfect synonymy.

- Even if many aspects of meaning are identical
- Still may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.

Relation: Synonymy?

water/H_20
big/large
brave/courageous

The Linguistic Principle of Contrast

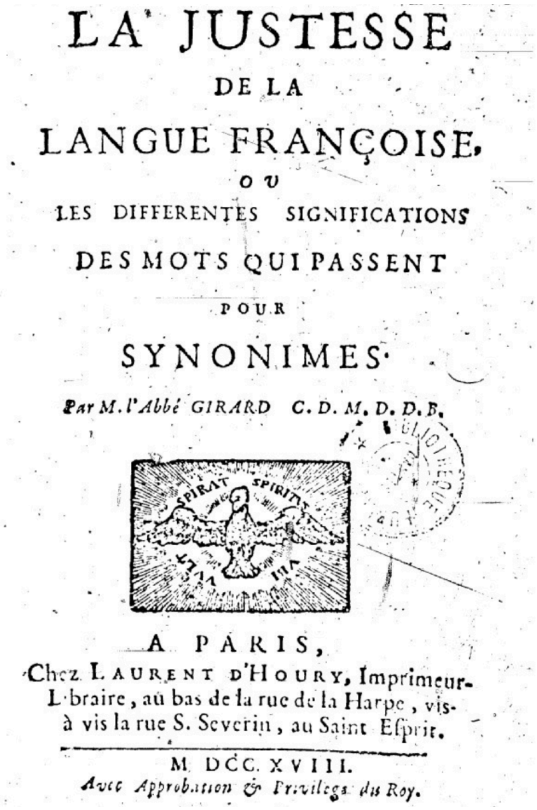
Difference in form → difference in meaning

Abbé Gabriel Girard 1718

Re: "exact" synonyms

"je ne crois pas qu'il y ait de mot ſynonime dans aucune Langue."

[I do not believe that there is a synonymous word in any language]



LA JUSTESSE
DE LA
LANGUE FRANÇOISE,
OU
LES DIFFERENTES SIGNIFICATIONS
DES MOTS QUI PASSENT
POUR
SYNONIMES.

Par M. l'Abbé GIRARD C. D. M. D. D. F.

A PARIS,
Chez LAURENT D'HOURY, Imprimeur-Libraire, au bas de la rue de la Harpe, vis-à-vis la rue S. Severin, au Saint Eſprit.

M. DCC. XVIII.
Avec Approbation & Privilege du Roy.

Relation: Similarity

Words with similar meanings. Not synonyms, but sharing some element of meaning

```
car, bicycle
cow, horse
```

Ask humans how similar 2 words are

word1	word2	similarity
vanish	disappear	9.8
behave	obey	7.3
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

SimLex-999 dataset (Hill et al., 2015)

Relation: Word relatedness

Also called "word association"

Words can be related in any way, perhaps via a semantic frame or field

- `car, bicycle`: **similar**
- `car, gasoline`: **related**, not similar

Semantic field

Words that
- cover a particular semantic domain
- bear structured relations with each other.

hospitals
surgeon, scalpel, nurse, anaesthetic, hospital
restaurants
waiter, menu, plate, food, menu, chef
houses
door, roof, kitchen, family, bed

Relation: Antonymy

Senses that are opposites with respect to only one feature of meaning

Otherwise, they are very similar!

`dark/light` `short/long` `fast/slow` `rise/fall`
`hot/cold` `up/down` `in/out`

More formally: antonyms can

- define a binary opposition or be at opposite ends of a scale
  - `long/short, fast/slow`
- Be *reversives*:
  - `rise/fall, up/down`

Relation: Superordinate/ subordinate

One sense is a **subordinate** of another if the first sense is more specific, denoting a subclass of the other

- *car* is a subordinate of *vehicle*
- *mango* is a subordinate of *fruit*

Conversely **superordinate**

- *vehicle* is a superordinate of *car*
- *fruit* is a subodinate of *mango*

Superordinate	vehicle	fruit	furniture
Subordinate	car	mango	chair

These levels are not symmetric

One level of category is distinguished from the others

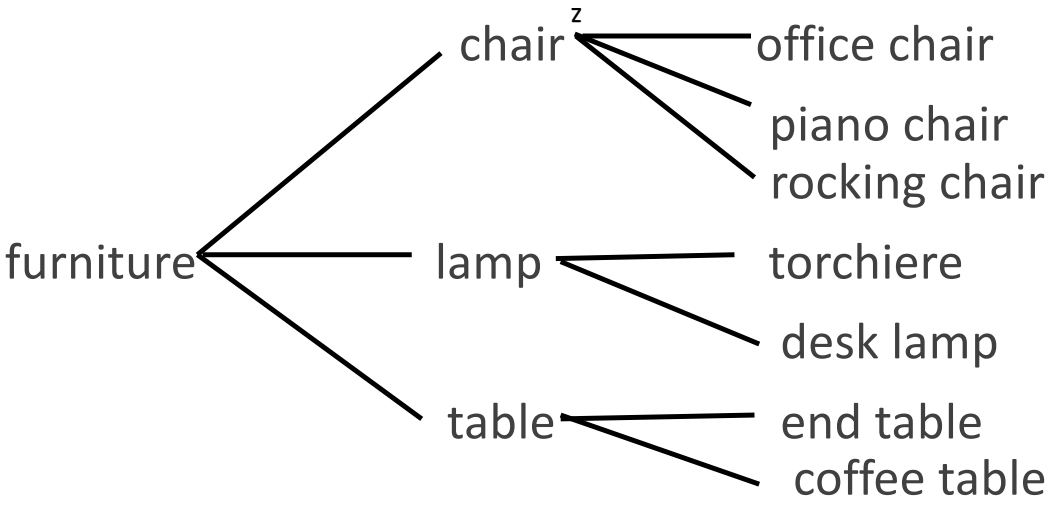
The "basic level"

Name these items





Superordinate **Basic** **Subordinate**



Cluster of Interactional Properties

Basic level things are "human-sized"

Consider chairs

- We know how to interact with a chair (sit)
- Not so clear for superordinate categories like furniture
  - "Imagine a furniture without thinking of a bed/table/chair/specific basic-level category"

The basic level

Distinctive actions

Learned earliest in childhood

Names are shortest

Names are most frequent

Connotation (sentiment)

Words have **affective** meanings

positive connotations (*happy*)

negative connotations (*sad*)

positive evaluation (*great*, *love*)

negative evaluation (*terrible*, *hate*).

Connotation

Osgood et al. (1957)

Words seem to vary along 3 affective dimensions:

- **valence**: the pleasantness of the stimulus (e.g.: *unhappy* vs. *happy*)
- **arousal**: the intensity of emotion provoked by the stimulus (*excited* vs. *calm*)
- **dominance**: the degree of control exerted by the stimulus (*controlling* vs. *awed*)

	Valence	Arousal	Dominance
courageous	8.05	5.5	7.38
music	7.67	5.57	6.5
heartbreak	2.45	5.65	3.58
cub	6.71	3.95	4.24

So far

Concepts or word senses

- Have a complex many-to-many association with **words** (homonymy, multiple senses)

Have relations with each other

- Synonymy
- Antonymy
- Similarity
- Relatedness
- Superordinate/subordinate, basic level
- Connotation

Vector Semantics & Embeddings

Word Meaning

Vector Semantics & Embeddings

It's hard to define a concept

But how to define a concept?

Classical ("Aristotelian") Theory of Concepts

The meaning of a word:

a concept defined by **necessary** and **sufficient** conditions

A **necessary** condition for being an X is a condition C that X must satisfy in order for it to be an X.

- If not C, then not X
- "Having four sides" is necessary to be a square.

A **sufficient** condition for being an X is condition such that if something satisfies condition C, then it must be an X.

- If and only if C, then X
- The following necessary conditions, jointly, are sufficient to be a square
  - x has (exactly) four sides
  - each of x's sides is straight
  - x is a closed figure
  - x lies in a plane
  - each of x's sides is equal in length to each of the others
  - each of x's interior angles is equal to the others (right angles)
  - the sides of x are joined at their ends

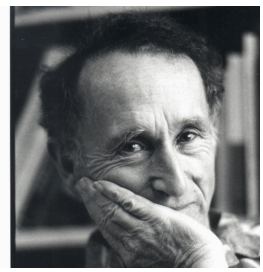
Example from Norman Swartz, SFU

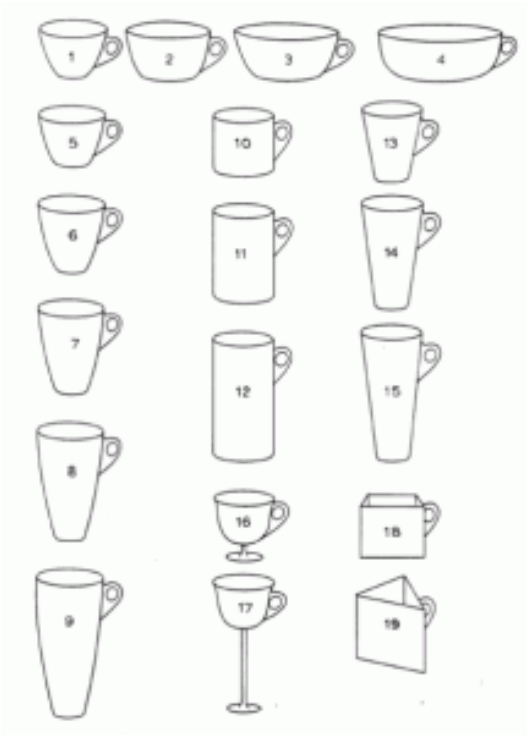
Problem 1: The features are complex & may be context-dependent

William Labov. 1975

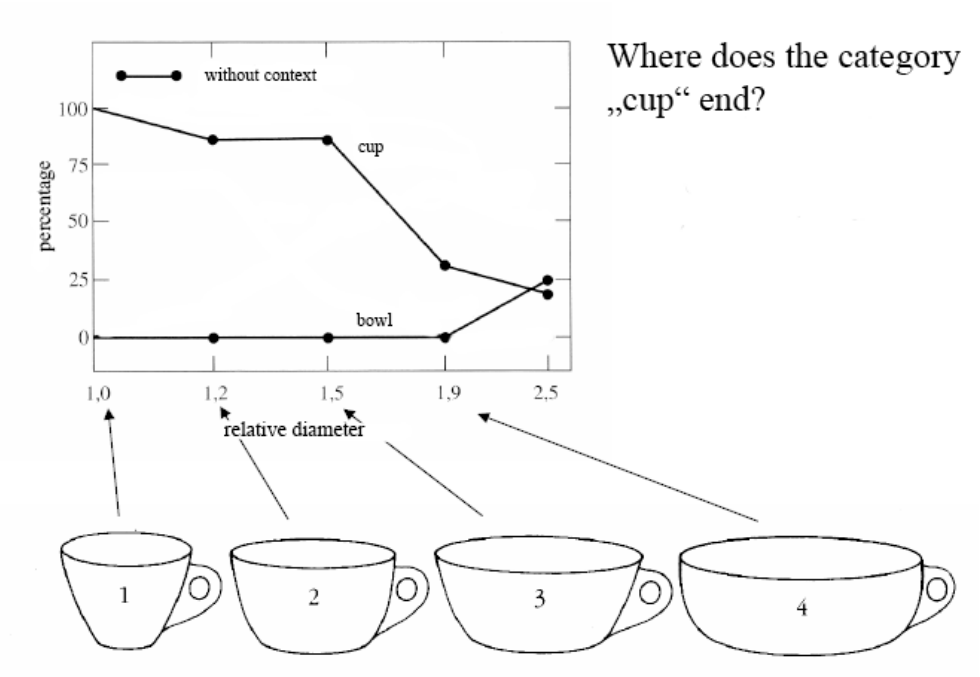
What are these?

Cup or bowl?



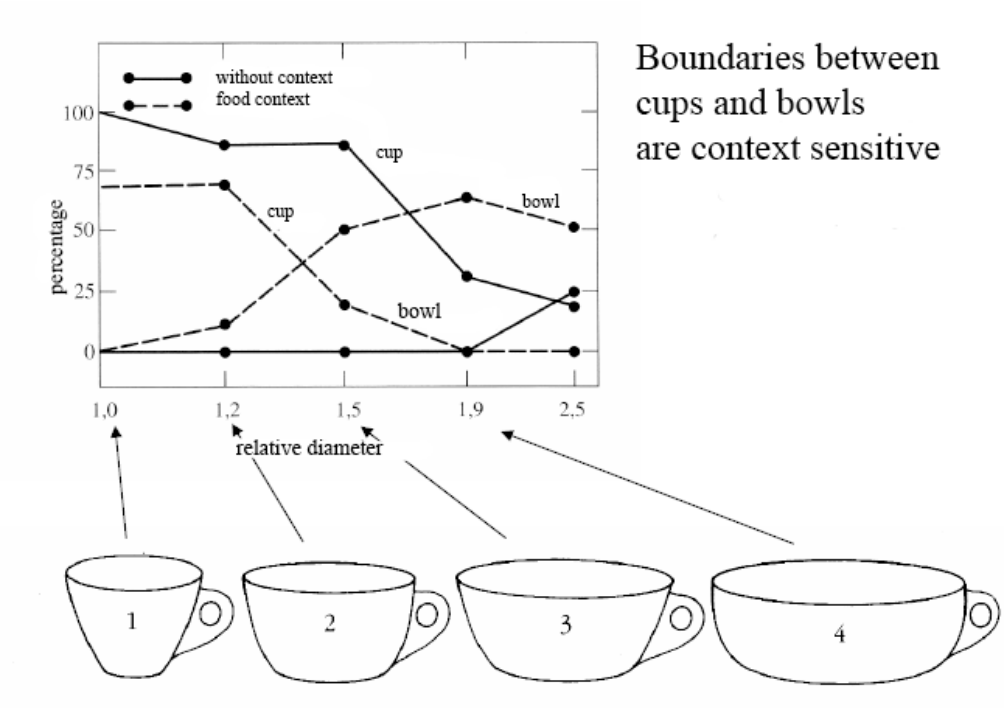


The category depends on complex features of the object (diameter, etc)



Where does the category „cup“ end?

The category depends on the context! (If there is food in it, it's a bowl)



Boundaries between cups and bowls are context sensitive

Labov's definition of cup

The term *cup* is used to denote round containers with a ratio of depth to width of $1 \pm r$ where $r \leq r_b$, and $r_b = \alpha_1 + \alpha_2 + \ldots \alpha_\upsilon$ and α_1 is a positive quality when the feature i is present and 0 otherwise.

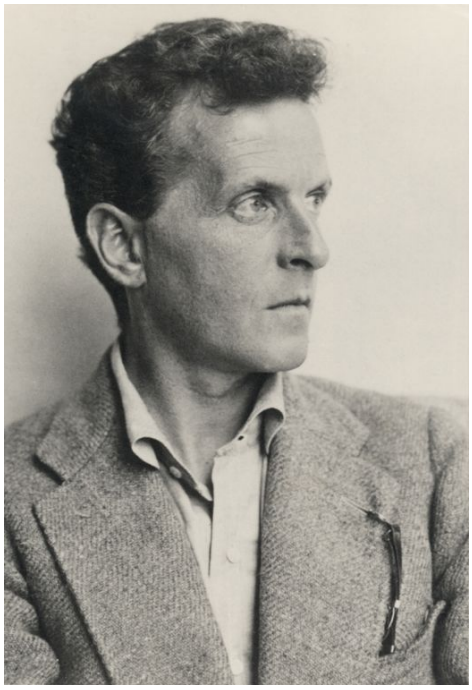
feature 1 = with one handle
2 = made of opaque vitreous material
3 = used for consumption of food
4 = used for the consumption of liquid food
5 = used for consumption of hot liquid food
6 = with a saucer
7 = tapering
8 = circular in cross-section

Cup is used variably to denote such containers with ratios width to depth $1 \pm r$ where $r_b \leq r \leq r_1$ with a probability of $r_1 - r / r_t - r_b$. The quantity $1 \pm r_b$ expresses the distance from the modal value of width to height.

Ludwig Wittgenstein (1889-1951)

Philosopher of language

In his late years, a proponent of studying "ordinary language"



Wittgenstein (1945) *Philosophical Investigations.* Paragraphs 66,67

66. Consider for example the proceedings that we call "games". I mean board-games, card-games, ball-games, Olympic games, and so on. What is common to them all?—Don't say: "There *must* be something common, or they would not be called 'games' "—but *look and see* whether there is anything common to all.—For if you look at them you will not see something that is common to *all*, but similarities, relationships, and a whole series of them at that. To repeat: don't think, but look!—Look for example at board-games, with their multifarious relationships. Now pass to card-games; here you find many correspondences with the first group, but many common features drop out, and others appear. When we pass next to ball-games, much that is common is retained, but much is lost.—Are they all 'amusing'? Compare chess with noughts and crosses. Or is there always winning and losing, or competition between players? Think of patience. In ball games there is winning and losing; but when a child throws his ball at the wall and catches it again, this feature has disappeared. Look at the parts played by skill and luck; and at the difference between skill in chess and skill in tennis. Think now of games like ring-a-ring-a-roses; here is the element of amusement, but how many other characteristic features have disappeared! And we can go through the many, many other groups of games in the same way; can see how similarities crop up and disappear.

And the result of this examination is: we see a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail.

67. I can think of no better expression to characterize these similarities than "family resemblances"; for the various resemblances between members of a family: build, features, colour of eyes, gait, temperament, etc. etc. overlap and criss-cross in the same way.—And I shall say: 'games' form a family.

And for instance the kinds of number form a family in the same way. Why do we call something a "number"? Well, perhaps because it has a—direct—relationship with several things that have hitherto been called number; and this can be said to give it an indirect relationship to other things we call the same name. And we extend our concept of number as in spinning a thread we twist fibre on fibre. And the strength of the thread does not reside in the fact that some one fibre runs through its whole length, but in the overlapping of many fibres.

But if someone wished to say: "There is something common to all these constructions—namely the disjunction of all their common properties"—I should reply: Now you are only playing with words. One might as well say: "Something runs through the whole thread—namely the continuous overlapping of those fibres".

What is a game?

Wittgenstein's thought experiment "What is a game":

PI #66:

"Don't say "there must be something common, or they would not be called `games'"—but *look and see* whether there is anything common to all"

- Is it amusing?
- Is there competition?
- Is there long-term strategy?
- Is skill required?
- Must luck play a role?
- Are there cards?
- Is there a ball?

Family Resemblance

Game 1	Game 2	Game 3	Game 4
ABC	BCD	ACD	ABD

"each item has at least one, and probably several, elements in common with one or more items, but no, or few, elements are common to all items" Rosch and Mervis

Vector Semantics & Embeddings

It's hard to define a concept

Vector Semantics & Embeddings

Vector Semantics

How about a radically different approach?

Ludwig Wittgenstein

PI #43:

"The meaning of a word is its use in the language"

Let's define words by their usages

One way to define "usage":

words are defined by their environments (the words around them)

Zellig Harris (1954):

If A and B have almost identical environments we say that they are synonyms.

What does recent English borrowing *ongchoi* mean?

Suppose you see these sentences:

- Ong choi is delicious **sautéed with garlic**.
- Ong choi is superb **over rice**
- Ong choi **leaves** with salty sauces

And you've also seen these:

- …spinach **sautéed with garlic over rice**
- Chard stems and **leaves** are **delicious**
- Collard greens and other **salty** leafy greens

Conclusion:

- Ongchoi is a leafy green like spinach, chard, or collard greens

Ongchoi: *Ipomoea aquatica "Water Spinach"*

空心菜
kangkong
rau muống
…

Yamaguchi, Wikimedia Commons, public domain

A new model of meaning focusing on distributional similarity

Each word = a vector

- Not just "word" or word45.

Similar words are "nearby in space"

We define a word as a vector

Called an "embedding" because it's embedded into a space

The standard way to represent meaning in NLP

Every modern NLP algorithm uses embeddings as the representation of word meaning

Fine-grained model of meaning for similarity

Intuition: why vectors?

Consider sentiment analysis:

- With **words**, a feature is a word identity
  - Feature 5: 'The previous word was "terrible"'
  - requires **exact same word** to be in training and test
- With **embeddings**:
  - Feature is a word vector
  - 'The previous word was vector [35,22,17…]
  - Now in the test set we might see a similar vector [34,21,14]
  - We can generalize to **similar but unseen** words!!!

We'll discuss 2 kinds of embeddings

tf-idf

- Information Retrieval workhorse!
- A common baseline model
- **Sparse** vectors
- Words are represented by (a simple function of) the **counts** of nearby words

Word2vec

- **Dense** vectors
- Representation is created by training a classifier to **predict** whether a word is likely to appear nearby
- In later chapters we'll discuss extensions called **contextual embeddings**

Vector Semantics & Embeddings

Vector Semantics

Vector Semantics & Embeddings

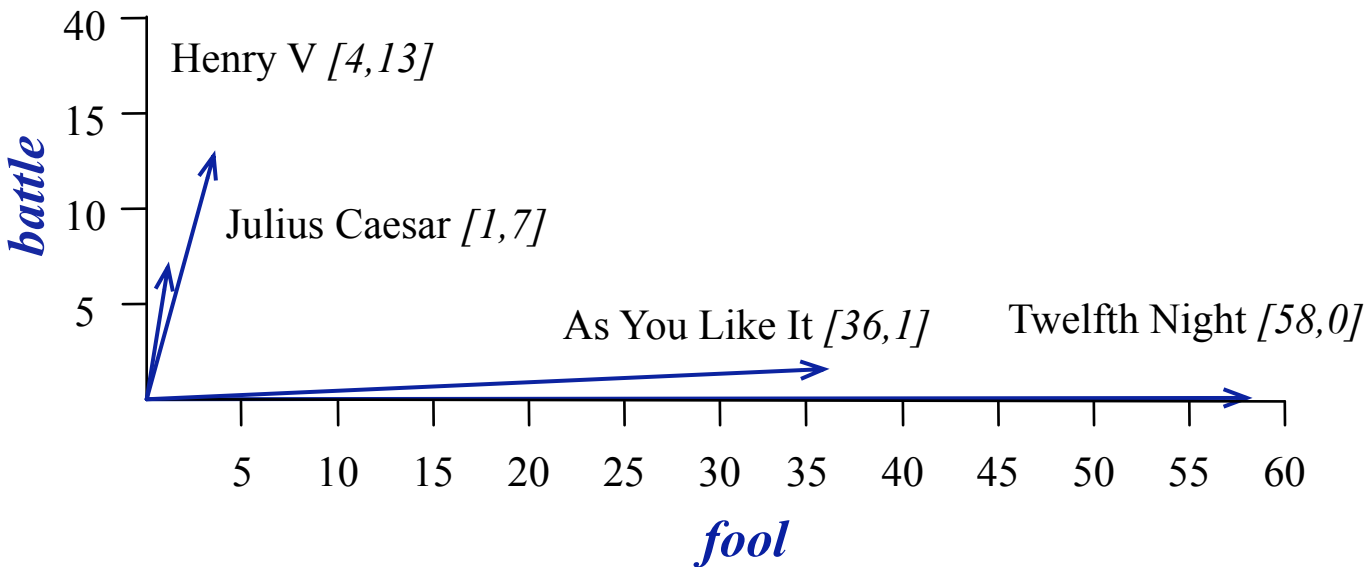
Words and Vectors

Term-document matrix

Each document is represented by a vector of words

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Visualizing document vectors



Vectors are the basis of information retrieval

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Vectors are similar for the two comedies
Different than the history

Comedies have more *fools* and *wit* and fewer *battles*.

Idea for word meaning: Words can be vectors too!!!

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

battle is "the kind of word that occurs in Julius Caesar and Henry V"

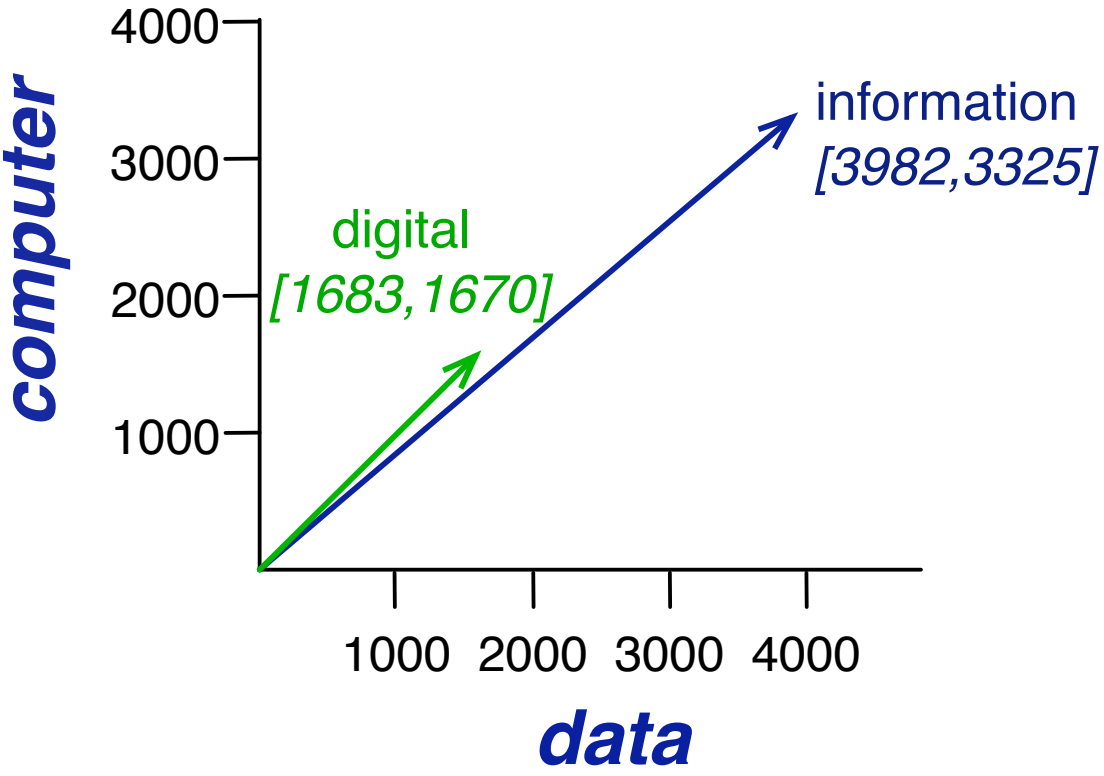
fool is "the kind of word that occurs in comedies, especially Twelfth Night"

More common: word-word matrix (or "term-context matrix")

Two **words** are similar in meaning if their context vectors are similar

is traditionally followed by	**cherry**	pie, a traditional dessert
often mixed, such as	**strawberry**	rhubarb pie. Apple pie
computer peripherals and personal	**digital**	assistants. These devices usually
a computer. This includes	**information**	available on the internet

	aardvark	...	**computer**	**data**	**result**	**pie**	**sugar**	...
cherry	0	...	2	8	9	442	25	...
strawberry	0	...	0	0	1	60	19	...
digital	0	...	1670	1683	85	5	4	...
information	0	...	3325	3982	378	5	13	...


4000
3000
2000
1000
computer
information
[3982,3325]
digital
[1683,1670]
1000 2000 3000 4000
data

Vector Semantics & Embeddings

Words and Vectors

Vector Semantics & Embeddings

Cosine for computing word similarity

Dot product and cosine

The dot product between two vectors is a scalar:

$$\text{dot product}(\mathbf{v},\mathbf{w}) = \mathbf{v}\cdot\mathbf{w} = \sum_{i=1}^{N} v_i w_i = v_1 w_1 + v_2 w_2 + \ldots + v_N w_N$$

The dot product tends to be high when the two vectors have large values in the same dimensions

Dot product can be a similarity metric between vectors

Problem with raw dot-product

Dot product favors long vectors

Dot product is higher if a vector is longer (has higher values in many dimension)

Vector length:

$$|\mathbf{v}| = \sqrt{\sum_{i=1}^{N} v_i^2}$$

Frequent words (of, the, you) have long vectors (since they occur many times with other words).

So dot product overly favors frequent words

Alternative: cosine for computing word similarity

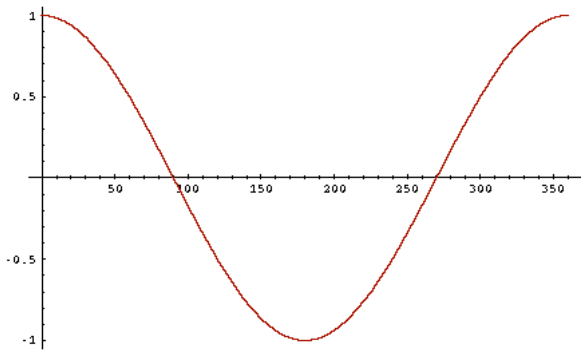
$$\text{cosine}(\vec{v},\vec{w}) = \frac{\vec{v}\cdot\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

Cosine as a similarity metric

-1: vectors point in opposite directions

+1: vectors point in same directions

0: vectors are orthogonal



But since raw frequency values are non-negative, the cosine for term-term matrix vectors ranges from 0–1

Cosine examples

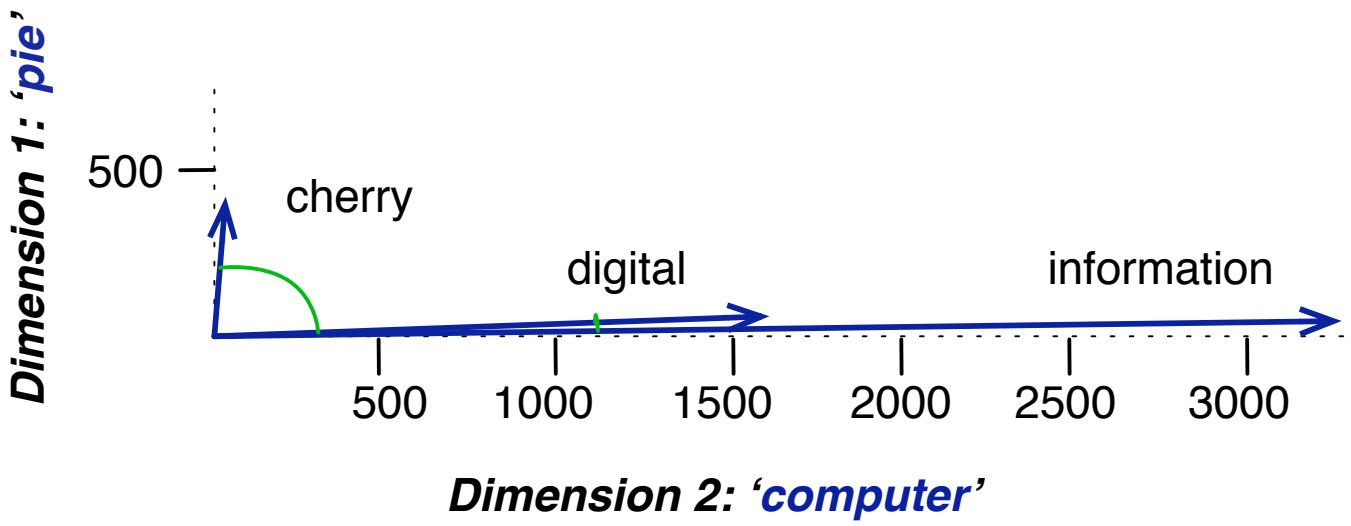
$$\cos(\vec{v},\vec{w}) = \frac{\vec{v}\bullet\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\vec{v}}{|\vec{v}|}\bullet\frac{\vec{w}}{|\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

	pie	data	computer
cherry	442	8	2
digital	5	1683	1670
information	5	3982	3325

$$\cos(\text{cherry},\text{information}) = \frac{442*5+8*3982+2*3325}{\sqrt{442^2+8^2+2^2}\sqrt{5^2+3982^2+3325^2}} = .017$$

$$\cos(\text{digital},\text{information}) = \frac{5*5+1683*3982+1670*3325}{\sqrt{5^2+1683^2+1670^2}\sqrt{5^2+3982^2+3325^2}} = .996$$

Visualizing cosines (well, angles)



Vector Semantics & Embeddings

Cosine for computing word similarity

Vector Semantics & Embeddings

TF-IDF

But raw frequency is a bad representation

- Frequency is clearly useful; if *sugar* appears a lot near *apricot*, that's useful information.
- But overly frequent words like *the*, *it,* or *they* are not very informative about the context
- Need a function that resolves this frequency paradox!

Two common solutions for word weighting

tf-idf: tf-idf value for word t in document d:

$$w_{t,d} = \text{tf}_{t,d} \times \text{idf}_t$$

Words like "the" or "good" have very low idf

PMI: (Pointwise mutual information)

- $\text{PMI}(w_1, w_2) = log \frac{p(w_1,w_2)}{p(w_1)p(w_2)}$

See if words like "good" appear more often with "great" than we would expect by chance

Term frequency (tf)

$tf_{t,d} = count(t,d)$

Instead of using raw count, we squash a bit:

$tf_{t,d} = \log_{10}(count(t,d)+1)$

Document frequency (df)

df_t *is* the number of documents *t* occurs in.

(note this is not collection frequency: total count across all documents)

"*Romeo*" is very distinctive for one Shakespeare play:

	Collection Frequency	Document Frequency
Romeo	113	1
action	113	31

Important: documents can be **anything**; we can call each paragraph a document

Inverse document frequency (idf)

$$\text{idf}_t = \log_{10}\left(\frac{N}{\text{df}_t}\right)$$

N is the total number of documents in the collection

Word	df	idf
Romeo	1	1.57
salad	2	1.27
Falstaff	4	0.967
forest	12	0.489
battle	21	0.246
wit	34	0.037
fool	36	0.012
good	37	0
sweet	37	0

Final tf-idf weighted value for a word

$$w_{t,d} = \text{tf}_{t,d} \times \text{idf}_t$$

Raw counts:

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Tf=idf:

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.074	0	0.22	0.28
good	0	0	0	0
fool	0.019	0.021	0.0036	0.0083
wit	0.049	0.044	0.018	0.022

Vector Semantics & Embeddings

TF-IDF

Vector Semantics & Embeddings

PPMI

Pointwise Mutual Information

Pointwise mutual information:
Do events x and y co-occur more than if they were independent?

$$\text{PMI}(X,Y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

PMI between two words: (Church & Hanks 1989)
Do words x and y co-occur more than if they were independent?

$$\text{PMI}(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}$$

Positive Pointwise Mutual Information

- PMI ranges from $-\infty$ to $+\infty$
- But the negative values are problematic
  - Things are co-occurring **less than** we expect by chance
  - Unreliable without enormous corpora
    - Imagine w1 and w2 whose probability is each 10^{-6}
    - Hard to be sure p(w1,w2) is significantly different than 10^{-12}
  - Plus it's not clear people are good at "unrelatedness"
- So we just replace negative PMI values by 0
- Positive PMI (**PPMI**) between word1 and word2:

$$\text{PPMI}(word_1, word_2) = \max\left(\log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}, 0\right)$$

Computing PPMI on a term-context matrix

Matrix F with W rows (words) and C columns (contexts)

f_{ij} is # of times w_i occurs in context c_j

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}} \qquad p_{i*} = \frac{\sum_{j=1}^{C} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}} \qquad p_{*j} = \frac{\sum_{i=1}^{W} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}}$$

	computer	data	result	pie	sugar	count(w)
cherry	2	8	9	442	25	486
strawberry	0	0	1	60	19	80
digital	1670	1683	85	5	4	3447
information	3325	3982	378	5	13	7703
count(context)	4997	5673	473	512	61	11716

$$pmi_{ij} = \log_2 \frac{p_{ij}}{p_{i*}p_{*j}} \qquad ppmi_{ij} = \begin{cases} pmi_{ij} & \text{if } pmi_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}}$$

	computer	data	result	pie	sugar	count(w)
cherry	2	8	9	442	25	486
strawberry	0	0	1	60	19	80
digital	1670	1683	85	5	4	3447
information	3325	3982	378	5	13	7703
count(context)	4997	5673	473	512	61	11716

p(w=information,c=data) = 3982/111716 = .3399

p(w=information) = 7703/11716 = .6575

p(c=data) = 5673/11716 = .4842

$$p(w_i) = \frac{\sum_{j=1}^{C} f_{ij}}{N} \qquad p(c_j) = \frac{\sum_{i=1}^{W} f_{ij}}{N}$$

	p(w,context)					p(w)
	computer	**data**	**result**	**pie**	**sugar**	**p(w)**
cherry	0.0002	0.0007	0.0008	0.0377	0.0021	0.0415
strawberry	0.0000	0.0000	0.0001	0.0051	0.0016	0.0068
digital	0.1425	0.1436	0.0073	0.0004	0.0003	0.2942
information	0.2838	0.3399	0.0323	0.0004	0.0011	0.6575
p(context)	0.4265	0.4842	0.0404	0.0437	0.0052	

$$pmi_{ij} = \log_2 \frac{p_{ij}}{p_{i*}p_{*j}}$$

	p(w,context)					p(w)
	computer	**data**	**result**	**pie**	**sugar**	**p(w)**
cherry	0.0002	0.0007	0.0008	0.0377	0.0021	0.0415
strawberry	0.0000	0.0000	0.0001	0.0051	0.0016	0.0068
digital	0.1425	0.1436	0.0073	0.0004	0.0003	0.2942
information	0.2838	0.3399	0.0323	0.0004	0.0011	0.6575
p(context)	0.4265	0.4842	0.0404	0.0437	0.0052	

pmi(information,data) = $\log_2$ (.3399 / (.6575*.4842)) = .0944

Resulting PPMI matrix (negatives replaced by 0)

	computer	**data**	**result**	**pie**	**sugar**
cherry	0	0	0	4.38	3.30
strawberry	0	0	0	4.10	5.51
digital	0.18	0.01	0	0	0
information	0.02	0.09	0.28	0	0

Weighting PMI

PMI is biased toward infrequent events

- Very rare words have very high PMI values

Two solutions:

- Give rare words slightly higher probabilities
- Use add-one smoothing (which has a similar effect)

Weighting PMI: Giving rare context words slightly higher probability

Raise the context probabilities to $\alpha = 0.75$:

$$\text{PPMI}_\alpha(w,c) = \max(\log_2 \frac{P(w,c)}{P(w)P_\alpha(c)}, 0)$$

$$P_\alpha(c) = \frac{count(c)^\alpha}{\sum_c count(c)^\alpha}$$

This helps because $P_\alpha(c) > P(c)$ for rare c

Consider two events, P(a) = .99 and P(b)=.01

$$P_\alpha(a) = \frac{.99^{.75}}{.99^{.75}+.01^{.75}} = .97 \quad P_\alpha(b) = \frac{.01^{.75}}{.01^{.75}+.01^{.75}} = .03$$

Vector Semantics & Embeddings

Dense vectors

Sparse versus dense vectors

tf-idf vectors are

- **long** (length |V|= 20,000 to 50,000)
- **sparse** (most elements are zero)

Alternative: learn vectors which are

- **short** (length 50-1000)
- **dense** (most elements are non-zero)

Sparse versus dense vectors

Why dense vectors?

- Short vectors may be easier to use as **features** in machine learning (fewer weights to tune)
- Dense vectors may **generalize** better than explicit counts
- They may do better at capturing synonymy:
  - *car* and *automobile* are synonyms; but are distinct dimensions
    - a word with *car* as a neighbor and a word with *automobile* as a neighbor should be similar, but aren't
- **In practice, they work better**

Common methods for getting short dense vectors

"Neural Language Model"-inspired models
- Word2vec (skipgram, CBOW), Glove

Singular Value Decomposition (SVD)
- A special case of this is called LSA – Latent Semantic Analysis

Alternative to these "static embeddings":
- Contextual Embeddings (ELMo, BERT)
- Compute distinct embeddings for a word in its context
- Separate embeddings for each token of a word
- We'll return to this in a later chapter

Vector Semantics & Embeddings

Dense vectors

Vector Semantics & Embeddings

Word2vec: The classifier

Embeddings you can download!

Word2vec (Mikolov et al)
https://code.google.com/archive/p/word2vec/

Glove (Pennington, Socher, Manning)
http://nlp.stanford.edu/projects/glove/

Word2vec

Popular embedding method

Very fast to train

Code available on the web

Idea: **predict** rather than **count**

Word2vec

Instead of **counting** how often each word *w* occurs near "*apricot*"

- Train a classifier on a binary **prediction** task:
  - Is *w* likely to show up near "*apricot*"?

We don't actually care about this task

- But we'll take the learned classifier weights as the word embeddings

Big idea: **self-supervision**:

- A word c that occurs near apricot in the corpus asks as the gold "correct answer" for supervised learning
- No need for human labels
- Bengio et al. (2003); Collobert et al. (2011)

Word2Vec: Skip-Gram Task

Word2vec provides a variety of options.
We'll do:

skip-gram with negative sampling (SGNS)

Approach: predict if candidate word *c* is a "neighbor"

1. Treat the target word *t* and a neighboring context word *c* as **positive examples**.
2. Randomly sample other words in the lexicon to get negative examples
3. Use logistic regression to train a classifier to distinguish those two cases
4. Use the learned weights as the embeddings

Skip-Gram Training Data

Assume a +/- 2 word window, given training sentence:

…lemon, a [tablespoon of apricot jam, a] pinch…

c1 c2 [target] c3 c4

Skip-Gram Classifier

(assuming a +/- 2 word window)

…lemon, a [tablespoon of apricot jam, a] pinch…
c1 c2 [target] c3 c4

Goal: train a classifier that is given a candidate (**w**ord, **c**ontext) pair
(apricot, tablespoon)
(apricot, aardvark)
…
And assigns each pair a probability:
$P(+|w, c)$

Similarity is computed from dot product

Remember: two vectors are similar if they have a high dot product

- Cosine is just a normalized dot product

So:

- Similarity(w,c) $\propto w \cdot c$

We'll need to normalize to get a probability

- (cosine isn't a probability either)

Turning dot products into probabilities

Sim(w,c) ≈ w · c

To turn this into a probability

We'll use the sigmoid from logistic regression:

$$P(+|w,c) = \sigma(c \cdot w) = \frac{1}{1+\exp(-c \cdot w)}$$

$$P(-|w,c) = 1 - P(+|w,c)$$

$$= \sigma(-c \cdot w) = \frac{1}{1+\exp(c \cdot w)}$$

How Skip-Gram Classifier computes $P(+|w, c)$

$$P(+|w,c) = \sigma(c \cdot w) = \frac{1}{1+\exp(-c \cdot w)}$$

This is for one context word, but we have lots of context words.
We'll assume independence and just multiply them:

$$P(+|w,c_{1:L}) = \prod_{i=1}^{L} \sigma(c_i \cdot w)$$

$$\log P(+|w,c_{1:L}) = \sum_{i=1}^{L} \log \sigma(c_i \cdot w)$$

Skip-gram classifier: summary

A probabilistic classifier that,

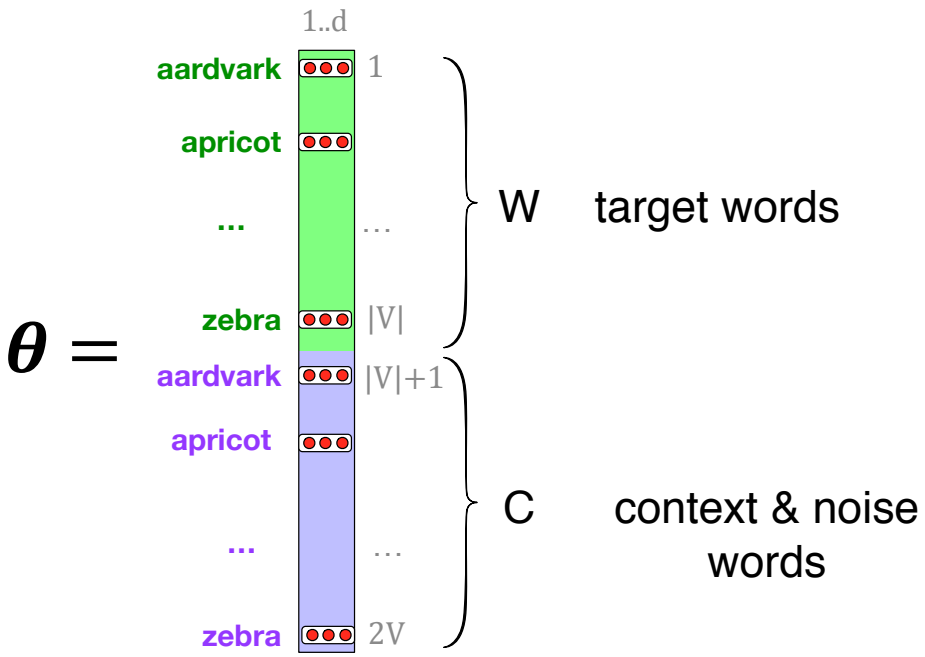
given a test target word *w*

its context window of *L* words $c_{1:L}$,

assigns a probability that w occurs in this window.

To compute this, we just need embeddings for all the words.

These embeddings we'll need: a set for w, a set for c



Vector Semantics & Embeddings

Word2vec: The classifier

Vector Semantics & Embeddings

Word2vec: Learning the embeddings

Skip-Gram Training data

…lemon, a [tablespoon of apricot jam, a] pinch…

c1 c2 [target] c3 c4

positive examples +

t	c
apricot	tablespoon
apricot	of
apricot	jam
apricot	a

Skip-Gram Training data

…lemon, a [tablespoon of apricot jam, a] pinch…

c1 c2 [target] c3 c4

positive examples +	
t	c
apricot	tablespoon
apricot	of
apricot	jam
apricot	a

For each positive example we'll grab k negative examples, sampling by frequency

Skip-Gram Training data

…lemon, a [tablespoon of apricot jam, a] pinch…

c1 c2 [target] c3 c4

positive examples +

t	c
apricot	tablespoon
apricot	of
apricot	jam
apricot	a

negative examples -

t	c	t	c
apricot	aardvark	apricot	seven
apricot	my	apricot	forever
apricot	where	apricot	dear
apricot	coaxial	apricot	if

Word2vec: how to learn vectors

Given the set of positive and negative training instances, and an initial set of embedding vectors

The goal of learning is to adjust those word vectors such that we:

- **Maximize** the similarity of the target word, context word pairs (w , c_{pos}) drawn from the positive data
- **Minimize** the similarity of the (w , c_{neg}) pairs drawn from the negative data.

Loss function for one *w with* c_{pos}, c_{neg1} ...c_{negk}

Maximize the dot product of the word with the actual context words, and minimize the dot products of the word with the *k* negative sampled non-neighbor words.

$$
\begin{aligned}
L_{CE} &= -\log\left[P(+|w,c_{pos})\prod_{i=1}^{k}P(-|w,c_{neg_i})\right] \\
&= -\left[\log P(+|w,c_{pos})+\sum_{i=1}^{k}\log P(-|w,c_{neg_i})\right] \\
&= -\left[\log P(+|w,c_{pos})+\sum_{i=1}^{k}\log\left(1-P(+|w,c_{neg_i})\right)\right] \\
&= -\left[\log \sigma(c_{pos}\cdot w)+\sum_{i=1}^{k}\log\sigma(-c_{neg_i}\cdot w)\right]
\end{aligned}
$$

Learning the classifier

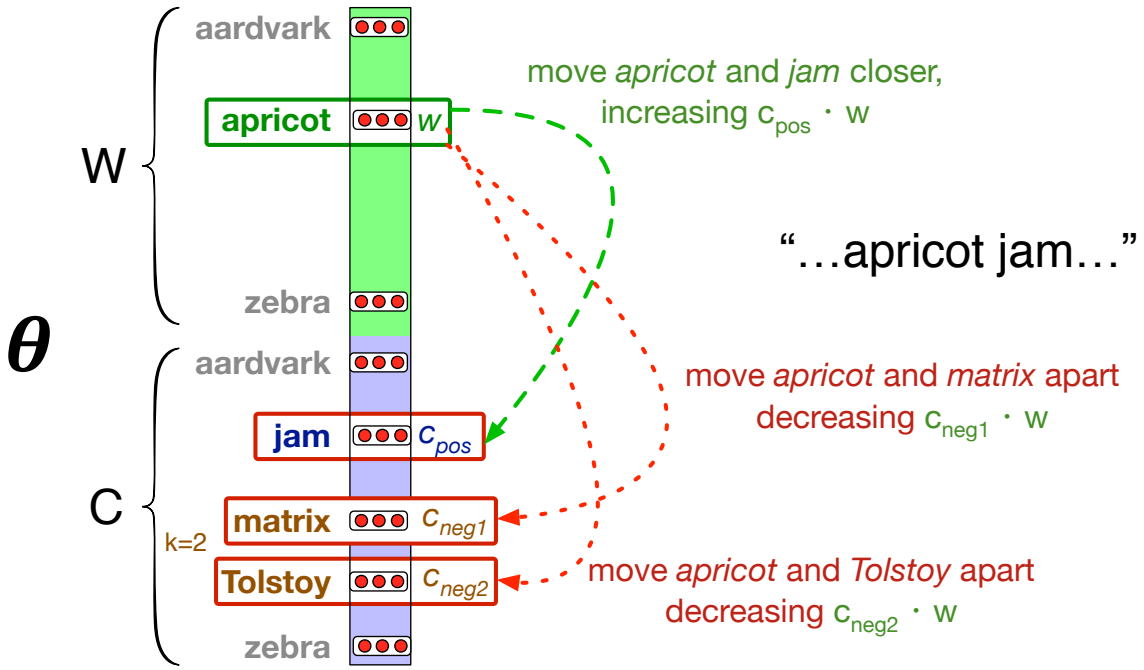
How to learn?

- Stochastic gradient descent!

We'll adjust the word weights to

- make the positive pairs more likely
- and the negative pairs less likely,
- over the entire training set.

Intuition of one step of gradient descent



The derivatives of the loss function

$$L_{CE} = -\left[\log \sigma(c_{pos} \cdot w) + \sum_{i=1}^{k} \log \sigma(-c_{neg_i} \cdot w)\right]$$

$$\frac{\partial L_{CE}}{\partial c_{pos}} = [\sigma(c_{pos} \cdot w) - 1]w$$

$$\frac{\partial L_{CE}}{\partial c_{neg}} = [\sigma(c_{neg} \cdot w)]w$$

$$\frac{\partial L_{CE}}{\partial w} = [\sigma(c_{pos} \cdot w) - 1]c_{pos} + \sum_{i=1}^{k} [\sigma(c_{\text{neg}_i} \cdot w)]c_{neg_i}$$

Update equation in SGD

Start with randomly initiatized C and W matrices, then incrementally do updates

$$c_{pos}^{t+1} = c_{pos}^{t} - \eta[\sigma(c_{pos}^{t} \cdot w) - 1]w$$

$$c_{neg}^{t+1} = c_{neg}^{t} - \eta[\sigma(c_{neg}^{t} \cdot w)]w$$

$$w^{t+1} = w^{t} - \eta[\sigma(c_{pos} \cdot w^{t}) - 1]c_{pos} + \sum_{i=1}^{k}[\sigma(c_{\mathrm{neg}_i} \cdot w^{t})]c_{neg_i}$$

Two sets of embeddings

SGNS learns two sets of embeddings

Target embeddings matrix W

Context embedding matrix C

It's common to just add them together, representing word *i* as the vector $w_i + c_i$

Summary: How to learn word2vec (skip-gram) embeddings

Start with V random d-dimentional vectors as initial embeddings

Train a classifier based on embedding similarity

- Take a corpus and take pairs of words that co-occur as positive examples
- Take pairs of words that don't co-occur as negative examples
- Train the classifier to distinguish these by slowly adjusting all the embeddings to improve the classifier performance
- Throw away the classifier code and keep the embeddings.

Vector Semantics & Embeddings

Word2vec: Learning the embeddings

Vector Semantics & Embeddings

Properties of Embeddings

The kinds of neighbors depend on window size

Large windows (C= +/- 5) : nearest words are related words in same semantic field

- *Hogwarts* nearest neighbors are Harry Potter world:
  - *Dumbledore, Half-blood, Malfoy*

Small windows (C= +/- 2) : nearest words are similar nouns words in same taxonomy

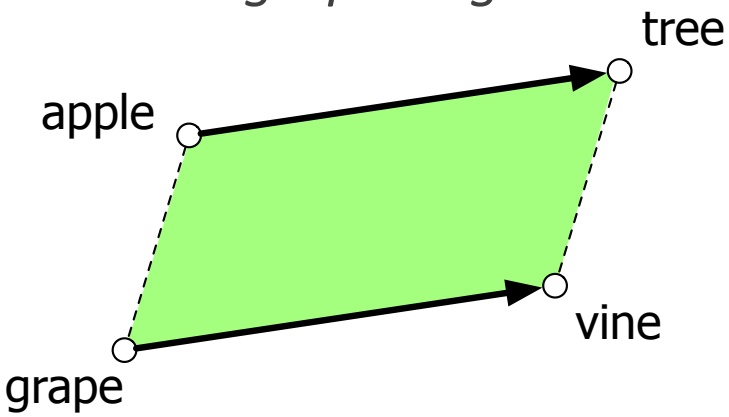
- *Hogwarts* nearest neighbors are other fictional schools
  - *Sunnydale, Evernight, Blandings*

Analogical relations

The classic parallelogram model of analogical reasoning
(Rumelhart and Abrahamson 1973)

To solve: *"apple is to tree as grape is to _____"*

Add $\overrightarrow{apple} - \overrightarrow{tree}$ to $\overrightarrow{grape}$ to get $\overrightarrow{vine}$



Analogical relations via parallelogram

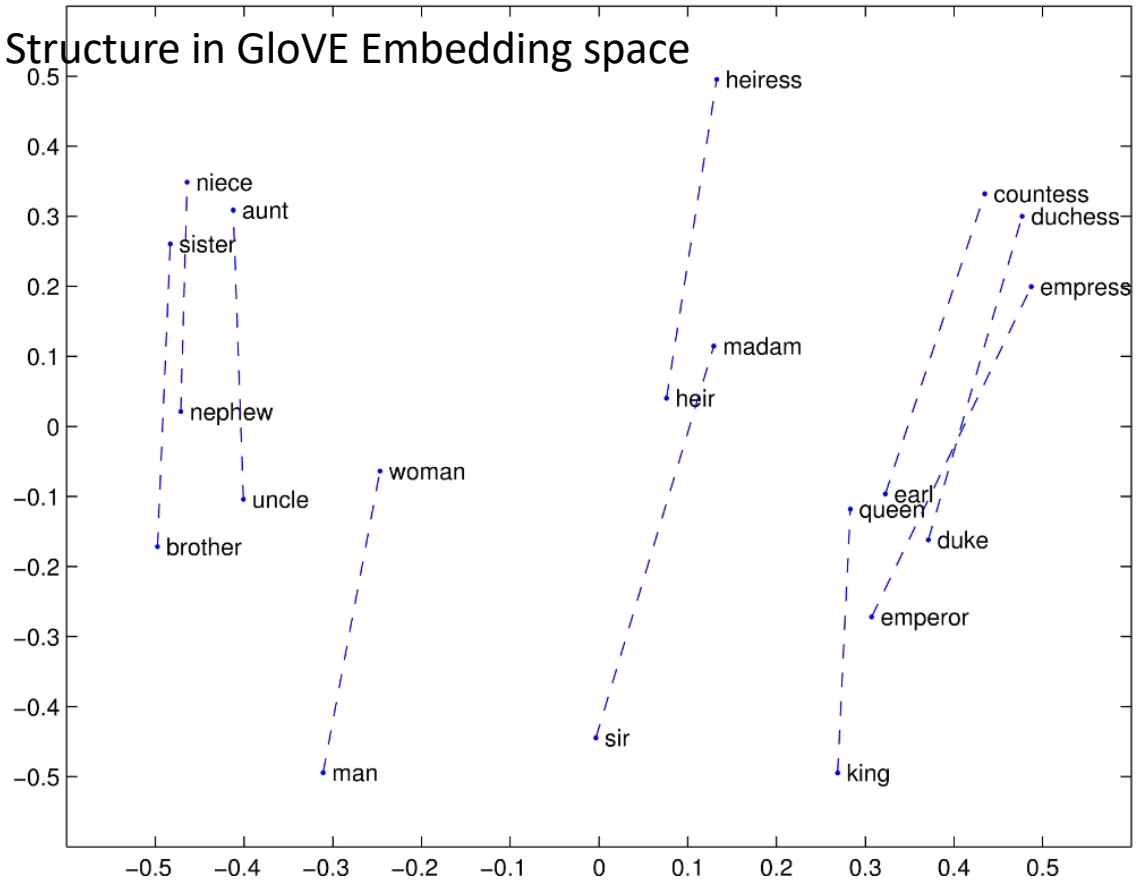
The parallelogram method can solve analogies with both sparse and dense embeddings (Turney and Littman 2005, Mikolov et al. 2013b)

$\overrightarrow{\text{king}} - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}}$ is close to $\overrightarrow{\text{queen}}$

$\overrightarrow{\text{Paris}} - \overrightarrow{\text{France}} + \overrightarrow{\text{Italy}}$ is close to $\overrightarrow{\text{Rome}}$

For a problem a:a*::b:b*, the parallelogram method is:

$$\hat{b}^* = \underset{x}{\text{argmax}}\ \text{distance}(x, a^* - a + b)$$


Structure in GloVE Embedding space
heiress
niece
aunt
sister
countess
duchess
empress
madam
heir
nephew
woman
uncle
earl
queen
duke
brother
emperor
sir
man
king
0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.5 −0.4 −0.3 −0.2 −0.1 0 0.1 0.2 0.3 0.4 0.5

Caveats with the parallelogram method

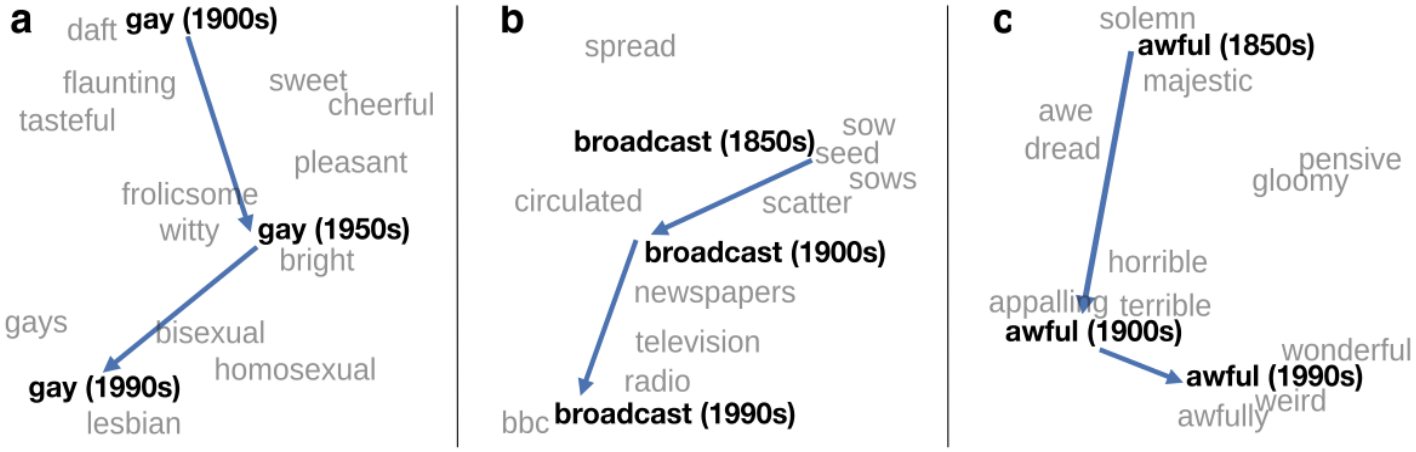
It only seems to work for frequent words, small distances and certain relations (relating countries to capitals, or parts of speech), but not others. (Linzen 2016, Gladkova et al. 2016, Ethayarajh et al. 2019a)

Understanding analogy is an open area of research (Peterson et al. 2020)

Embeddings as a window onto historical semantics

Train embeddings on different decades of historical text to see meanings shift

~30 million books, 1850-1990, Google Books data



William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. Proceedings of ACL.

Embeddings reflect cultural bias!

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." In *NeurIPS*, pp. 4349-4357. 2016.

Ask "Paris : France :: Tokyo : x"

- x = Japan

Ask "father : doctor :: mother : x"

- x = nurse

Ask "man : computer programmer :: woman : x"

- x = homemaker

Algorithms that use embeddings as part of e.g., hiring searches for programmers, might lead to bias in hiring

Historical embedding as a tool to study cultural biases

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences 115(16), E3635–E3644.

- Compute a **gender or ethnic bias** for each adjective: e.g., how much closer the adjective is to "woman" synonyms than "man" synonyms, or names of particular ethnicities
  - Embeddings for **competence** adjective (*smart, wise, brilliant, resourceful, thoughtful, logical)* are biased toward men, a bias slowly decreasing 1960-1990
  - Embeddings for **dehumanizing** adjectives (barbaric, monstrous, bizarre) were biased toward Asians in the 1930s, bias decreasing over the 20th century.
- These match the results of old surveys done in the 1930s

Vector Semantics & Embeddings

Properties of Embeddings

Chapter 3

Jurafsky's Text

Text of Chapter 6 of Jurafski, Dan, and James H. Martin. *Speech and Language Processing.* Third edition draft. https://web.stanford.edu/~jurafsky/slp3/

CHAPTER

6 Vector Semantics and Embeddings

荃者所以在鱼，得鱼而忘荃 Nets are for fish;
Once you get the fish, you can forget the net.
言者所以在意，得意而忘言 Words are for meaning;
Once you get the meaning, you can forget the words
庄子(Zhuangzi), Chapter 26

The asphalt that Los Angeles is famous for occurs mainly on its freeways. But in the middle of the city is another patch of asphalt, the La Brea tar pits, and this asphalt preserves millions of fossil bones from the last of the Ice Ages of the Pleistocene Epoch. One of these fossils is the *Smilodon*, or saber-toothed tiger, instantly recognizable by its long canines. Five million years ago or so, a completely different sabre-tooth tiger called *Thylacosmilus* lived in Argentina and other parts of South America. Thylacosmilus was a marsupial whereas Smilodon was a placental mammal, but Thylacosmilus had the same long upper canines and, like Smilodon, had a protective bone flange on the lower jaw. The similarity of these two mammals is one of many examples of parallel or convergent evolution, in which particular contexts or environments lead to the evolution of very similar structures in different species (Gould, 1980).

The role of context is also important in the similarity of a less biological kind of organism: the word. Words that occur in *similar contexts* tend to have *similar meanings*. This link between similarity in how words are distributed and similarity in what they mean is called the **distributional hypothesis**. The hypothesis was first formulated in the 1950s by linguists like Joos (1950), Harris (1954), and Firth (1957), who noticed that words which are synonyms (like *oculist* and *eye-doctor*) tended to occur in the same environment (e.g., near words like *eye* or *examined*) with the amount of meaning difference between two words "corresponding roughly to the amount of difference in their environments" (Harris, 1954, 157).

distributional hypothesis

In this chapter we introduce **vector semantics**, which instantiates this linguistic hypothesis by learning representations of the meaning of words, called **embeddings**, directly from their distributions in texts. These representations are used in every natural language processing application that makes use of meaning, and the **static embeddings** we introduce here underlie the more powerful dynamic or **contextualized embeddings** like **BERT** that we will see in Chapter 10.

vector semantics
embeddings

These word representations are also the first example in this book of **representation learning**, automatically learning useful representations of the input text. Finding such **self-supervised** ways to learn representations of the input, instead of creating representations by hand via **feature engineering**, is an important focus of NLP research (Bengio et al., 2013).

representation learning

6.1 Lexical Semantics

Let's begin by introducing some basic principles of word meaning. How should we represent the meaning of a word? In the n-gram models of Chapter 3, and in classical NLP applications, our only representation of a word is as a string of letters, or an index in a vocabulary list. This representation is not that different from a tradition in philosophy, perhaps you've seen it in introductory logic classes, in which the meaning of words is represented by just spelling the word with small capital letters; representing the meaning of "dog" as DOG, and "cat" as CAT.

Representing the meaning of a word by capitalizing it is a pretty unsatisfactory model. You might have seen a joke due originally to semanticist Barbara Partee (Carlson, 1977):

> Q: What's the meaning of life?
> A: LIFE'

Surely we can do better than this! After all, we'll want a model of word meaning to do all sorts of things for us. It should tell us that some words have similar meanings (*cat* is similar to *dog*), others are antonyms (*cold* is the opposite of *hot*), some have positive connotations (*happy*) while others have negative connotations (*sad*). It should represent the fact that the meanings of *buy*, *sell*, and *pay* offer differing perspectives on the same underlying purchasing event (If I buy something from you, you've probably sold it to me, and I likely paid you). More generally, a model of word meaning should allow us to draw inferences to address meaning-related tasks like question-answering or dialogue.

lexical semantics

In this section we summarize some of these desiderata, drawing on results in the linguistic study of word meaning, which is called **lexical semantics**; we'll return to and expand on this list in Chapter 18 and Chapter 10.

Lemmas and Senses Let's start by looking at how one word (we'll choose *mouse*) might be defined in a dictionary (simplified from the online dictionary WordNet):

```
mouse (N)
1.  any of numerous small rodents...
2.  a hand-operated device that controls a cursor...
```

lemma **citation form**

Here the form *mouse* is the **lemma**, also called the **citation form**. The form *mouse* would also be the lemma for the word *mice*; dictionaries don't have separate definitions for inflected forms like *mice*. Similarly *sing* is the lemma for *sing*, *sang*, *sung*. In many languages the infinitive form is used as the lemma for the verb, so Spanish *dormir* "to sleep" is the lemma for *duermes* "you sleep". The specific forms *sung* or *carpets* or *sing* or *duermes* are called **wordforms**.

wordform

As the example above shows, each lemma can have multiple meanings; the lemma *mouse* can refer to the rodent or the cursor control device. We call each of these aspects of the meaning of *mouse* a **word sense**. The fact that lemmas can be **polysemous** (have multiple senses) can make interpretation difficult (is someone who types "mouse info" into a search engine looking for a pet or a tool?). Chapter 18 will discuss the problem of polysemy, and introduce **word sense disambiguation**, the task of determining which sense of a word is being used in a particular context.

Synonymy One important component of word meaning is the relationship between word senses. For example when one word has a sense whose meaning is identical to a sense of another word, or nearly identical, we say the two senses of those two words are **synonyms**. Synonyms include such pairs as

synonym

couch/sofa vomit/throw up filbert/hazelnut car/automobile

A more formal definition of synonymy (between words rather than senses) is that two words are synonymous if they are substitutable for one another in any sentence without changing the *truth conditions* of the sentence, the situations in which the sentence would be true. We often say in this case that the two words have the same **propositional meaning**.

propositional meaning

While substitutions between some pairs of words like *car* / *automobile* or *water* / H_2O are truth preserving, the words are still not identical in meaning. Indeed, probably no two words are absolutely identical in meaning. One of the fundamental tenets of semantics, called the **principle of contrast** (Girard 1718, Bréal 1897, Clark 1987), states that a difference in linguistic form is always associated with some difference in meaning. For example, the word H_2O is used in scientific contexts and would be inappropriate in a hiking guide—*water* would be more appropriate— and this genre difference is part of the meaning of the word. In practice, the word *synonym* is therefore used to describe a relationship of approximate or rough synonymy.

principle of contrast

Word Similarity While words don't have many synonyms, most words do have lots of *similar* words. *Cat* is not a synonym of *dog*, but *cats* and *dogs* are certainly similar words. In moving from synonymy to similarity, it will be useful to shift from talking about relations between word senses (like synonymy) to relations between words (like similarity). Dealing with words avoids having to commit to a particular representation of word senses, which will turn out to simplify our task.

similarity

The notion of word **similarity** is very useful in larger semantic tasks. Knowing how similar two words are can help in computing how similar the meaning of two phrases or sentences are, a very important component of natural language understanding tasks like question answering, paraphrasing, and summarization. One way of getting values for word similarity is to ask humans to judge how similar one word is to another. A number of datasets have resulted from such experiments. For example the SimLex-999 dataset (Hill et al., 2015) gives values on a scale from 0 to 10, like the examples below, which range from near-synonyms (*vanish*, *disappear*) to pairs that scarcely seem to have anything in common (*hole*, *agreement*):

vanish	disappear	9.8
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

Word Relatedness The meaning of two words can be related in ways other than similarity. One such class of connections is called word **relatedness** (Budanitsky and Hirst, 2006), also traditionally called word **association** in psychology.

relatedness

association

Consider the meanings of the words *coffee* and *cup*. Coffee is not similar to cup; they share practically no features (coffee is a plant or a beverage, while a cup is a manufactured object with a particular shape). But coffee and cup are clearly related; they are associated by co-participating in an everyday event (the event of drinking coffee out of a cup). Similarly *scalpel* and *surgeon* are not similar but are related eventively (a surgeon tends to make use of a scalpel).

One common kind of relatedness between words is if they belong to the same **semantic field**. A semantic field is a set of words which cover a particular semantic domain and bear structured relations with each other. For example, words might be related by being in the semantic field of hospitals (*surgeon*, *scalpel*, *nurse*, *anesthetic*, *hospital*), restaurants (*waiter*, *menu*, *plate*, *food*, *chef*), or houses (*door*, *roof*,

semantic field

topic models

kitchen, *family*, *bed*). Semantic fields are also related to **topic models**, like **Latent Dirichlet Allocation**, **LDA**, which apply unsupervised learning on large sets of texts to induce sets of associated words from text. Semantic fields and topic models are very useful tools for discovering topical structure in documents.

In Chapter 18 we'll introduce more relations between senses like **hypernymy** or **IS-A**, **antonymy** (opposites) and **meronymy** (part-whole relations).

semantic frame

Semantic Frames and Roles Closely related to semantic fields is the idea of a **semantic frame**. A semantic frame is a set of words that denote perspectives or participants in a particular type of event. A commercial transaction, for example, is a kind of event in which one entity trades money to another entity in return for some good or service, after which the good changes hands or perhaps the service is performed. This event can be encoded lexically by using verbs like *buy* (the event from the perspective of the buyer), *sell* (from the perspective of the seller), *pay* (focusing on the monetary aspect), or nouns like *buyer*. Frames have semantic roles (like *buyer*, *seller*, *goods*, *money*), and words in a sentence can take on these roles.

Knowing that *buy* and *sell* have this relation makes it possible for a system to know that a sentence like *Sam bought the book from Ling* could be paraphrased as *Ling sold the book to Sam*, and that Sam has the role of the *buyer* in the frame and Ling the *seller*. Being able to recognize such paraphrases is important for question answering, and can help in shifting perspective for machine translation.

connotations

Connotation Finally, words have *affective meanings* or **connotations**. The word *connotation* has different meanings in different fields, but here we use it to mean the aspects of a word's meaning that are related to a writer or reader's emotions, sentiment, opinions, or evaluations. For example some words have positive connotations (*happy*) while others have negative connotations (*sad*). Even words whose meanings are similar in other ways can vary in connotation; consider the difference in connotations between *fake*, *knockoff*, *forgery*, on the one hand, and *copy*, *replica*, *reproduction* on the other, or *innocent* (positive connotation) and *naive* (negative connotation). Some words describe positive evaluation (*great*, *love*) and others negative evaluation (*terrible*, *hate*). Positive or negative evaluation language is called **sentiment**, as we saw in Chapter 4, and word sentiment plays a role in important tasks like sentiment analysis, stance detection, and applications of NLP to the language of politics and consumer reviews.

sentiment

Early work on affective meaning (Osgood et al., 1957) found that words varied along three important dimensions of affective meaning:

valence: the pleasantness of the stimulus

arousal: the intensity of emotion provoked by the stimulus

dominance: the degree of control exerted by the stimulus

Thus words like *happy* or *satisfied* are high on valence, while *unhappy* or *annoyed* are low on valence. *Excited* is high on arousal, while *calm* is low on arousal. *Controlling* is high on dominance, while *awed* or *influenced* are low on dominance. Each word is thus represented by three numbers, corresponding to its value on each of the three dimensions:

	Valence	Arousal	Dominance
courageous	8.05	5.5	7.38
music	7.67	5.57	6.5
heartbreak	2.45	5.65	3.58
cub	6.71	3.95	4.24

Osgood et al. (1957) noticed that in using these 3 numbers to represent the meaning of a word, the model was representing each word as a point in a three-dimensional space, a vector whose three dimensions corresponded to the word's rating on the three scales. This revolutionary idea that word meaning could be represented as a point in space (e.g., that part of the meaning of *heartbreak* can be represented as the point $[2.45, 5.65, 3.58]$) was the first expression of the vector semantics models that we introduce next.

6.2 Vector Semantics

vector semantics

Vectors semantics is the standard way to represent word meaning in NLP, helping us model many of the aspects of word meaning we saw in the previous section. The roots of the model lie in the 1950s when two big ideas converged: Osgood's (1957) idea mentioned above to use a point in three-dimensional space to represent the connotation of a word, and the proposal by linguists like Joos (1950), Harris (1954), and Firth (1957) to define the meaning of a word by its **distribution** in language use, meaning its neighboring words or grammatical environments. Their idea was that two words that occur in very similar distributions (whose neighboring words are similar) have similar meanings.

For example, suppose you didn't know the meaning of the word *ongchoi* (a recent borrowing from Cantonese) but you see it in the following contexts:

(6.1) Ongchoi is delicious sauteed with garlic.

(6.2) Ongchoi is superb over rice.

(6.3) ...ongchoi leaves with salty sauces...

And suppose that you had seen many of these context words in other contexts:

(6.4) ...spinach sauteed with garlic over rice...

(6.5) ...chard stems and leaves are delicious...

(6.6) ...collard greens and other salty leafy greens

The fact that *ongchoi* occurs with words like *rice* and *garlic* and *delicious* and *salty*, as do words like *spinach*, *chard*, and *collard greens* might suggest that ongchoi is a leafy green similar to these other leafy greens.[1] We can do the same thing computationally by just counting words in the context of *ongchoi*.

The idea of vector semantics is to represent a word as a point in a multidimensional semantic space that is derived (in ways we'll see) from the distributions of word neighbors. Vectors for representing words are called **embeddings** (although the term is sometimes more strictly applied only to dense vectors like word2vec (Section 6.8), rather than sparse tf-idf or PPMI vectors (Section 6.3-Section 6.6)). The word "embedding" derives from its mathematical sense as a mapping from one space or structure to another, although the meaning has shifted; see the end of the chapter.

embeddings

Fig. 6.1 shows a visualization of embeddings learned for sentiment analysis, showing the location of selected words projected down from 60-dimensional space into a two dimensional space. Notice the distinct regions containing positive words, negative words, and neutral function words.

[1] It's in fact *Ipomoea aquatica*, a relative of morning glory sometimes called *water spinach* in English.

Figure 6.1 A two-dimensional (t-SNE) projection of embeddings for some words and phrases, showing that words with similar meanings are nearby in space. The original 60-dimensional embeddings were trained for sentiment analysis. Simplified from Li et al. (2015) with colors added for explanation.

The fine-grained model of word similarity of vector semantics offers enormous power to NLP applications. NLP applications like the sentiment classifiers of Chapter 4 or Chapter 5 depend on the same words appearing in the training and test sets. But by representing words as embeddings, classifiers can assign sentiment as long as it sees some words with *similar meanings*. And as we'll see, vector semantic models can be learned automatically from text without supervision.

In this chapter we'll introduce the two most commonly used models. In the **tf-idf** model, an important baseline, the meaning of a word is defined by a simple function of the counts of nearby words. We will see that this method results in very long vectors that are **sparse**, i.e. mostly zeros (since most words simply never occur in the context of others). We'll introduce the **word2vec** model family for constructing short, **dense** vectors that have useful semantic properties. We'll also introduce the **cosine**, the standard way to use embeddings to compute *semantic similarity*, between two words, two sentences, or two documents, an important tool in practical applications like question answering, summarization, or automatic essay grading.

6.3 Words and Vectors

"The most important attributes of a vector in 3-space are {Location, Location, Location}"
Randall Munroe, `https://xkcd.com/2358/`

Vector or distributional models of meaning are generally based on a **co-occurrence matrix**, a way of representing how often words co-occur. We'll look at two popular matrices: the term-document matrix and the term-term matrix.

6.3.1 Vectors and documents

term-document matrix

In a **term-document matrix**, each row represents a word in the vocabulary and each column represents a document from some collection of documents. Fig. 6.2 shows a small selection from a term-document matrix showing the occurrence of four words in four plays by Shakespeare. Each cell in this matrix represents the number of times a particular word (defined by the row) occurs in a particular document (defined by the column). Thus *fool* appeared 58 times in *Twelfth Night*.

vector space model

The term-document matrix of Fig. 6.2 was first defined as part of the **vector space model** of information retrieval (Salton, 1971). In this model, a document is

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.2 The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document.

represented as a count vector, a column in Fig. 6.3.

vector

To review some basic linear algebra, a **vector** is, at heart, just a list or array of numbers. So *As You Like It* is represented as the list [1,114,36,20] (the first **column vector** in Fig. 6.3) and *Julius Caesar* is represented as the list [7,62,1,2] (the third column vector). A **vector space** is a collection of vectors, characterized by their **dimension**. In the example in Fig. 6.3, the document vectors are of dimension 4, just so they fit on the page; in real term-document matrices, the vectors representing each document would have dimensionality $|V|$, the vocabulary size.

vector space

dimension

The ordering of the numbers in a vector space indicates different meaningful dimensions on which documents vary. Thus the first dimension for both these vectors corresponds to the number of times the word *battle* occurs, and we can compare each dimension, noting for example that the vectors for *As You Like It* and *Twelfth Night* have similar values (1 and 0, respectively) for the first dimension.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.3 The term-document matrix for four words in four Shakespeare plays. The red boxes show that each document is represented as a column vector of length four.

We can think of the vector for a document as a point in $|V|$-dimensional space; thus the documents in Fig. 6.3 are points in 4-dimensional space. Since 4-dimensional spaces are hard to visualize, Fig. 6.4 shows a visualization in two dimensions; we've arbitrarily chosen the dimensions corresponding to the words *battle* and *fool*.

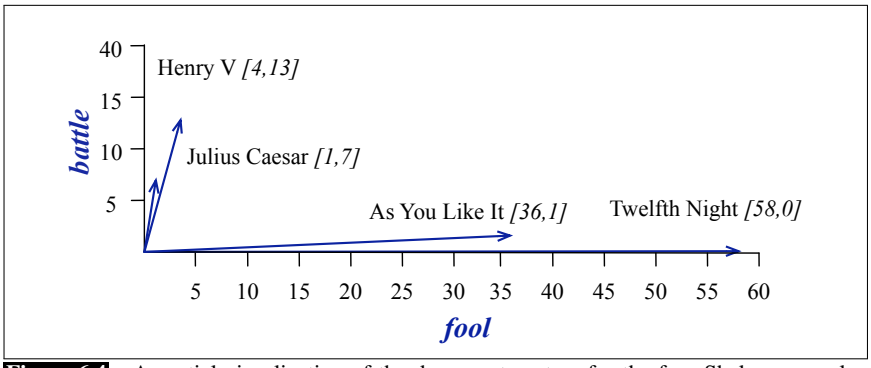


Figure 6.4 A spatial visualization of the document vectors for the four Shakespeare play documents, showing just two of the dimensions, corresponding to the words *battle* and *fool*. The comedies have high values for the *fool* dimension and low values for the *battle* dimension.

Term-document matrices were originally defined as a means of finding similar documents for the task of document **information retrieval**. Two documents that are

similar will tend to have similar words, and if two documents have similar words their column vectors will tend to be similar. The vectors for the comedies *As You Like It* [1,114,36,20] and *Twelfth Night* [0,80,58,15] look a lot more like each other (more fools and wit than battles) than they look like *Julius Caesar* [7,62,1,2] or *Henry V* [13,89,4,3]. This is clear with the raw numbers; in the first dimension (battle) the comedies have low numbers and the others have high numbers, and we can see it visually in Fig. 6.4; we'll see very shortly how to quantify this intuition more formally.

A real term-document matrix, of course, wouldn't just have 4 rows and columns, let alone 2. More generally, the term-document matrix has $|V|$ rows (one for each word type in the vocabulary) and D columns (one for each document in the collection); as we'll see, vocabulary sizes are generally in the tens of thousands, and the number of documents can be enormous (think about all the pages on the web).

information retrieval

Information retrieval (IR) is the task of finding the document d from the D documents in some collection that best matches a query q. For IR we'll therefore also represent a query by a vector, also of length $|V|$, and we'll need a way to compare two vectors to find how similar they are. (Doing IR will also require efficient ways to store and manipulate these vectors by making use of the convenient fact that these vectors are sparse, i.e., mostly zeros).

Later in the chapter we'll introduce some of the components of this vector comparison process: the tf-idf term weighting, and the cosine similarity metric.

6.3.2 Words as vectors: document dimensions

We've seen that documents can be represented as vectors in a vector space. But vector semantics can also be used to represent the meaning of *words*. We do this by associating each word with a word vector— a **row vector** rather than a column vector, hence with different dimensions, as shown in Fig. 6.5. The four dimensions of the vector for *fool*, [36,58,1,4], correspond to the four Shakespeare plays. Word counts in the same four dimensions are used to form the vectors for the other 3 words: *wit*, [20,15,2,3]; *battle*, [1,0,7,13]; and *good* [114,80,62,89].

row vector

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.5 The term-document matrix for four words in four Shakespeare plays. The red boxes show that each word is represented as a row vector of length four.

For documents, we saw that similar documents had similar vectors, because similar documents tend to have similar words. This same principle applies to words: similar words have similar vectors because they tend to occur in similar documents. The term-document matrix thus lets us represent the meaning of a word by the documents it tends to occur in.

6.3.3 Words as vectors: word dimensions

An alternative to using the term-document matrix to represent words as vectors of document counts, is to use the **term-term matrix**, also called the **word-word matrix** or the **term-context matrix**, in which the columns are labeled by words rather than documents. This matrix is thus of dimensionality $|V| \times |V|$ and each cell records

word-word matrix

the number of times the row (target) word and the column (context) word co-occur in some context in some training corpus. The context could be the document, in which case the cell represents the number of times the two words appear in the same document. It is most common, however, to use smaller contexts, generally a window around the word, for example of 4 words to the left and 4 words to the right, in which case the cell represents the number of times (in some training corpus) the column word occurs in such a ± 4 word window around the row word. For example here is one example each of some words in their windows:

is traditionally followed by	**cherry**	pie, a traditional dessert
often mixed, such as	**strawberry**	rhubarb pie. Apple pie
computer peripherals and personal	**digital**	assistants. These devices usually
a computer. This includes	**information**	available on the internet

If we then take every occurrence of each word (say **strawberry**) and count the context words around it, we get a word-word co-occurrence matrix. Fig. 6.6 shows a simplified subset of the word-word co-occurrence matrix for these four words computed from the Wikipedia corpus (Davies, 2015).

	aardvark	...	**computer**	**data**	**result**	**pie**	**sugar**	...
cherry	0	...	2	8	9	442	25	...
strawberry	0	...	0	0	1	60	19	...
digital	0	...	1670	1683	85	5	4	...
information	0	...	3325	3982	378	5	13	...

Figure 6.6 Co-occurrence vectors for four words in the Wikipedia corpus, showing six of the dimensions (hand-picked for pedagogical purposes). The vector for *digital* is outlined in red. Note that a real vector would have vastly more dimensions and thus be much sparser.

Note in Fig. 6.6 that the two words *cherry* and *strawberry* are more similar to each other (both *pie* and *sugar* tend to occur in their window) than they are to other words like *digital*; conversely, *digital* and *information* are more similar to each other than, say, to *strawberry*. Fig. 6.7 shows a spatial visualization.

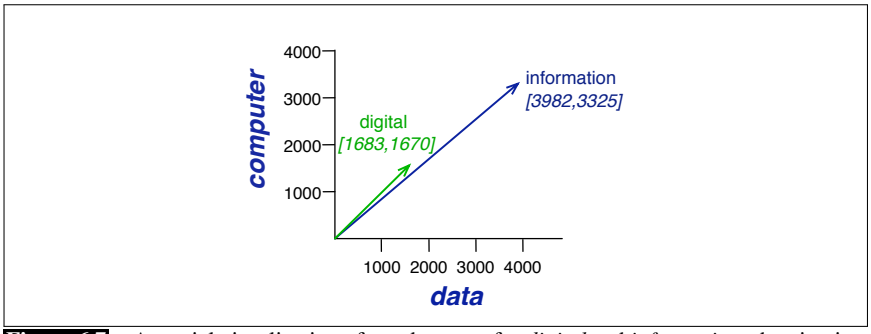


Figure 6.7 A spatial visualization of word vectors for *digital* and *information*, showing just two of the dimensions, corresponding to the words *data* and *computer*.

Note that $|V|$, the length of the vector, is generally the size of the vocabulary, often between 10,000 and 50,000 words (using the most frequent words in the training corpus; keeping words after about the most frequent 50,000 or so is generally not helpful). Since most of these numbers are zero these are **sparse** vector representations; there are efficient algorithms for storing and computing with sparse matrices.

Now that we have some intuitions, let's move on to examine the details of computing word similarity. Afterwards we'll discuss methods for weighting cells.

6.4 Cosine for measuring similarity

To measure similarity between two target words v and w, we need a metric that takes two vectors (of the same dimensionality, either both with words as dimensions, hence of length $|V|$, or both with documents as dimensions as documents, of length $|D|$) and gives a measure of their similarity. By far the most common similarity metric is the **cosine** of the angle between the vectors.

dot product
inner product

The cosine—like most measures for vector similarity used in NLP—is based on the **dot product** operator from linear algebra, also called the **inner product**:

$$\text{dot product}(\mathbf{v},\mathbf{w}) = \mathbf{v}\cdot\mathbf{w} = \sum_{i=1}^{N} v_i w_i = v_1 w_1 + v_2 w_2 + \ldots + v_N w_N \tag{6.7}$$

As we will see, most metrics for similarity between vectors are based on the dot product. The dot product acts as a similarity metric because it will tend to be high just when the two vectors have large values in the same dimensions. Alternatively, vectors that have zeros in different dimensions—orthogonal vectors—will have a dot product of 0, representing their strong dissimilarity.

vector length

This raw dot product, however, has a problem as a similarity metric: it favors **long** vectors. The **vector length** is defined as

$$|\mathbf{v}| = \sqrt{\sum_{i=1}^{N} v_i^2} \tag{6.8}$$

The dot product is higher if a vector is longer, with higher values in each dimension. More frequent words have longer vectors, since they tend to co-occur with more words and have higher co-occurrence values with each of them. The raw dot product thus will be higher for frequent words. But this is a problem; we'd like a similarity metric that tells us how similar two words are regardless of their frequency.

We modify the dot product to normalize for the vector length by dividing the dot product by the lengths of each of the two vectors. This **normalized dot product** turns out to be the same as the cosine of the angle between the two vectors, following from the definition of the dot product between two vectors **a** and **b**:

$$\begin{aligned} \mathbf{a}\cdot\mathbf{b} &= |\mathbf{a}||\mathbf{b}|\cos\theta \\ \frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}||\mathbf{b}|} &= \cos\theta \end{aligned} \tag{6.9}$$

cosine

The **cosine** similarity metric between two vectors **v** and **w** thus can be computed as:

$$\text{cosine}(\mathbf{v},\mathbf{w}) = \frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}} \tag{6.10}$$

unit vector

For some applications we pre-normalize each vector, by dividing it by its length, creating a **unit vector** of length 1. Thus we could compute a unit vector from **a** by dividing it by $|\mathbf{a}|$. For unit vectors, the dot product is the same as the cosine.

The cosine value ranges from 1 for vectors pointing in the same direction, through 0 for orthogonal vectors, to -1 for vectors pointing in opposite directions. But since raw frequency values are non-negative, the cosine for these vectors ranges from 0–1.

Let's see how the cosine computes which of the words *cherry* or *digital* is closer in meaning to *information*, just using raw counts from the following shortened table:

	pie	data	computer
cherry	442	8	2
digital	5	1683	1670
information	5	3982	3325

$$\cos(\text{cherry, information}) = \frac{442*5+8*3982+2*3325}{\sqrt{442^2+8^2+2^2}\sqrt{5^2+3982^2+3325^2}} = .017$$

$$\cos(\text{digital, information}) = \frac{5*5+1683*3982+1670*3325}{\sqrt{5^2+1683^2+1670^2}\sqrt{5^2+3982^2+3325^2}} = .996$$

The model decides that *information* is way closer to *digital* than it is to *cherry*, a result that seems sensible. Fig. 6.8 shows a visualization.

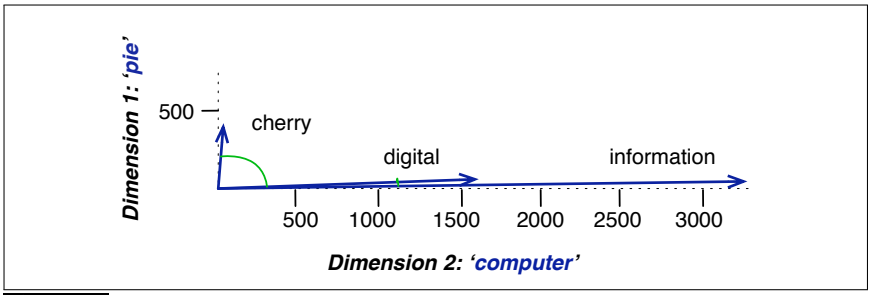


Figure 6.8 A (rough) graphical demonstration of cosine similarity, showing vectors for three words (*cherry*, *digital*, and *information*) in the two dimensional space defined by counts of the words *computer* and *pie* nearby. Note that the angle between *digital* and *information* is smaller than the angle between *cherry* and *information*. When two vectors are more similar, the cosine is larger but the angle is smaller; the cosine has its maximum (1) when the angle between two vectors is smallest (0°); the cosine of all other angles is less than 1.

6.5 TF-IDF: Weighing terms in the vector

The co-occurrence matrices above represent each cell by frequencies, either of words with documents (Fig. 6.5), or words with other words (Fig. 6.6). But raw frequency is not the best measure of association between words. Raw frequency is very skewed and not very discriminative. If we want to know what kinds of contexts are shared by *cherry* and *strawberry* but not by *digital* and *information*, we're not going to get good discrimination from words like *the*, *it*, or *they*, which occur frequently with all sorts of words and aren't informative about any particular word. We saw this also in Fig. 6.3 for the Shakespeare corpus; the dimension for the word *good* is not very discriminative between plays; *good* is simply a frequent word and has roughly equivalent high frequencies in each of the plays.

It's a bit of a paradox. Words that occur nearby frequently (maybe *pie* nearby *cherry*) are more important than words that only appear once or twice. Yet words

that are too frequent—ubiquitous, like *the* or *good*— are unimportant. How can we balance these two conflicting constraints?

There are two common solutions to this problem: in this section we'll describe the **tf-idf** algorithm, usually used when the dimensions are documents. In the next we introduce the **PPMI** algorithm (usually used when the dimensions are words).

The **tf-idf algorithm** (the '-' here is a hyphen, not a minus sign) is the product of two terms, each term capturing one of these two intuitions:

term frequency

The first is the **term frequency** (Luhn, 1957): the frequency of the word t in the document d. We can just use the raw count as the term frequency:

$$\text{tf}_{t,d} = \text{count}(t,d) \tag{6.11}$$

More commonly we squash the raw frequency a bit, by using the $\log_{10}$ of the frequency instead. The intuition is that a word appearing 100 times in a document doesn't make that word 100 times more likely to be relevant to the meaning of the document. Because we can't take the log of 0, we normally add 1 to the count:[2]

$$\text{tf}_{t,d} = \log_{10}(\text{count}(t,d)+1) \tag{6.12}$$

If we use log weighting, terms which occur 0 times in a document would have $\text{tf} = \log_{10}(1) = 0$, 10 times in a document $\text{tf} = \log_{10}(11) = 1.4$, 100 times $\text{tf} = \log_{10}(101) = 2.004$, 1000 times $\text{tf} = 3.00044$, and so on.

document frequency

The second factor in tf-idf is used to give a higher weight to words that occur only in a few documents. Terms that are limited to a few documents are useful for discriminating those documents from the rest of the collection; terms that occur frequently across the entire collection aren't as helpful. The **document frequency** df_t of a term t is the number of documents it occurs in. Document frequency is not the same as the **collection frequency** of a term, which is the total number of times the word appears in the whole collection in any document. Consider in the collection of Shakespeare's 37 plays the two words *Romeo* and *action*. The words have identical collection frequencies (they both occur 113 times in all the plays) but very different document frequencies, since Romeo only occurs in a single play. If our goal is to find documents about the romantic tribulations of Romeo, the word *Romeo* should be highly weighted, but not *action*:

	Collection Frequency	Document Frequency
Romeo	113	1
action	113	31

idf

We emphasize discriminative words like *Romeo* via the **inverse document frequency** or **idf** term weight (Sparck Jones, 1972). The idf is defined using the fraction N/df_t, where N is the total number of documents in the collection, and df_t is the number of documents in which term t occurs. The fewer documents in which a term occurs, the higher this weight. The lowest weight of 1 is assigned to terms that occur in all the documents. It's usually clear what counts as a document: in Shakespeare we would use a play; when processing a collection of encyclopedia articles like Wikipedia, the document is a Wikipedia page; in processing newspaper articles, the document is a single article. Occasionally your corpus might not have appropriate document divisions and you might need to break up the corpus into documents yourself for the purposes of computing idf.

[2] Or we can use this alternative: $\text{tf}_{t,d} = \begin{cases} 1+\log_{10}\text{count}(t,d) & \text{if } \text{count}(t,d) > 0 \\ 0 & \text{otherwise} \end{cases}$

Because of the large number of documents in many collections, this measure too is usually squashed with a log function. The resulting definition for inverse document frequency (idf) is thus

$$\text{idf}_t = \log_{10}\left(\frac{N}{\text{df}_t}\right) \tag{6.13}$$

Here are some idf values for some words in the Shakespeare corpus, ranging from extremely informative words which occur in only one play like *Romeo*, to those that occur in a few like *salad* or *Falstaff*, to those which are very common like *fool* or so common as to be completely non-discriminative since they occur in all 37 plays like *good* or *sweet*.[3]

Word	df	idf
Romeo	1	1.57
salad	2	1.27
Falstaff	4	0.967
forest	12	0.489
battle	21	0.246
wit	34	0.037
fool	36	0.012
good	37	0
sweet	37	0

tf-idf

The **tf-idf** weighted value $w_{t,d}$ for word t in document d thus combines term frequency $\text{tf}_{t,d}$ (defined either by Eq. 6.11 or by Eq. 6.12) with idf from Eq. 6.13:

$$w_{t,d} = \text{tf}_{t,d} \times \text{idf}_t \tag{6.14}$$

Fig. 6.9 applies tf-idf weighting to the Shakespeare term-document matrix in Fig. 6.2, using the tf equation Eq. 6.12. Note that the tf-idf values for the dimension corresponding to the word *good* have now all become 0; since this word appears in every document, the tf-idf algorithm leads it to be ignored. Similarly, the word *fool*, which appears in 36 out of the 37 plays, has a much lower weight.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.074	0	0.22	0.28
good	0	0	0	0
fool	0.019	0.021	0.0036	0.0083
wit	0.049	0.044	0.018	0.022

Figure 6.9 A tf-idf weighted term-document matrix for four words in four Shakespeare plays, using the counts in Fig. 6.2. For example the 0.049 value for *wit* in *As You Like It* is the product of $\text{tf} = \log_{10}(20+1) = 1.322$ and $\text{idf} = .037$. Note that the idf weighting has eliminated the importance of the ubiquitous word *good* and vastly reduced the impact of the almost-ubiquitous word *fool*.

The tf-idf weighting is the way for weighting co-occurrence matrices in information retrieval, but also plays a role in many other aspects of natural language processing. It's also a great baseline, the simple thing to try first. We'll look at other weightings like PPMI (Positive Pointwise Mutual Information) in Section 6.6.

[3] *Sweet* was one of Shakespeare's favorite adjectives, a fact probably related to the increased use of sugar in European recipes around the turn of the 16th century (Jurafsky, 2014, p. 175).

6.6 Pointwise Mutual Information (PMI)

An alternative weighting function to tf-idf, PPMI (positive pointwise mutual information), is used for term-term-matrices, when the vector dimensions correspond to words rather than documents. PPMI draws on the intuition that the best way to weigh the association between two words is to ask how much **more** the two words co-occur in our corpus than we would have a priori expected them to appear by chance.

pointwise mutual information

Pointwise mutual information (Fano, 1961)[4] is one of the most important concepts in NLP. It is a measure of how often two events x and y occur, compared with what we would expect if they were independent:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{6.16}$$

The pointwise mutual information between a target word w and a context word c (Church and Hanks 1989, Church and Hanks 1990) is then defined as:

$$\text{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)} \tag{6.17}$$

The numerator tells us how often we observed the two words together (assuming we compute probability by using the MLE). The denominator tells us how often we would **expect** the two words to co-occur assuming they each occurred independently; recall that the probability of two independent events both occurring is just the product of the probabilities of the two events. Thus, the ratio gives us an estimate of how much more the two words co-occur than we expect by chance. PMI is a useful tool whenever we need to find words that are strongly associated.

PMI values range from negative to positive infinity. But negative PMI values (which imply things are co-occurring *less often* than we would expect by chance) tend to be unreliable unless our corpora are enormous. To distinguish whether two words whose individual probability is each 10^{-6} occur together less often than chance, we would need to be certain that the probability of the two occurring together is significantly different than 10^{-12}, and this kind of granularity would require an enormous corpus. Furthermore it's not clear whether it's even possible to evaluate such scores of 'unrelatedness' with human judgments. For this reason it is more common to use Positive PMI (called **PPMI**) which replaces all negative PMI values with zero (Church and Hanks 1989, Dagan et al. 1993, Niwa and Nitta 1994)[5]:

PPMI

$$\text{PPMI}(w,c) = \max(\log_2 \frac{P(w,c)}{P(w)P(c)}, 0) \tag{6.18}$$

More formally, let's assume we have a co-occurrence matrix F with W rows (words) and C columns (contexts), where f_{ij} gives the number of times word w_i occurs in

[4] PMI is based on the **mutual information** between two random variables X and Y, defined as:

$$I(X,Y) = \sum_x \sum_y P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{6.15}$$

In a confusion of terminology, Fano used the phrase *mutual information* to refer to what we now call *pointwise mutual information* and the phrase *expectation of the mutual information* for what we now call *mutual information*

[5] Positive PMI also cleanly solves the problem of what to do with zero counts, using 0 to replace the $-\infty$ from $\log(0)$.

context c_j. This can be turned into a PPMI matrix where $ppmi_{ij}$ gives the PPMI value of word w_i with context c_j as follows:

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}} \quad p_{i*} = \frac{\sum_{j=1}^{C} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}} \quad p_{*j} = \frac{\sum_{i=1}^{W} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}} \tag{6.19}$$

$$\text{PPMI}_{ij} = \max(\log_2 \frac{p_{ij}}{p_{i*}p_{*j}}, 0) \tag{6.20}$$

Let's see some PPMI calculations. We'll use Fig. 6.10, which repeats Fig. 6.6 plus all the count marginals, and let's pretend for ease of calculation that these are the only words/contexts that matter.

	computer	data	result	pie	sugar	count(w)
cherry	2	8	9	442	25	486
strawberry	0	0	1	60	19	80
digital	1670	1683	85	5	4	3447
information	3325	3982	378	5	13	7703
count(context)	4997	5673	473	512	61	11716

Figure 6.10 Co-occurrence counts for four words in 5 contexts in the Wikipedia corpus, together with the marginals, pretending for the purpose of this calculation that no other words/contexts matter.

Thus for example we could compute PPMI(w=information,c=data), assuming we pretended that Fig. 6.6 encompassed all the relevant word contexts/dimensions, as follows:

$$\begin{aligned} P(\text{w=information,c=data}) &= \frac{3982}{11716} = .3399 \\ P(\text{w=information}) &= \frac{7703}{11716} = .6575 \\ P(\text{c=data}) &= \frac{5673}{11716} = .4842 \\ \text{ppmi(information,data)} &= \log 2(.3399/(.6575 * .4842)) = .0944 \end{aligned}$$

Fig. 6.11 shows the joint probabilities computed from the counts in Fig. 6.10, and Fig. 6.12 shows the PPMI values. Not surprisingly, *cherry* and *strawberry* are highly associated with both *pie* and *sugar*, and *data* is mildly associated with *information*.

	p(w,context)					p(w)
	computer	**data**	**result**	**pie**	**sugar**	**p(w)**
cherry	0.0002	0.0007	0.0008	0.0377	0.0021	0.0415
strawberry	0.0000	0.0000	0.0001	0.0051	0.0016	0.0068
digital	0.1425	0.1436	0.0073	0.0004	0.0003	0.2942
information	0.2838	0.3399	0.0323	0.0004	0.0011	0.6575
p(context)	0.4265	0.4842	0.0404	0.0437	0.0052	

Figure 6.11 Replacing the counts in Fig. 6.6 with joint probabilities, showing the marginals around the outside.

PMI has the problem of being biased toward infrequent events; very rare words tend to have very high PMI values. One way to reduce this bias toward low frequency

	computer	data	result	pie	sugar
cherry	0	0	0	4.38	3.30
strawberry	0	0	0	4.10	5.51
digital	0.18	0.01	0	0	0
information	0.02	0.09	0.28	0	0

Figure 6.12 The PPMI matrix showing the association between words and context words, computed from the counts in Fig. 6.11. Note that most of the 0 PPMI values are ones that had a negative PMI; for example PMI(*cherry,computer*) = -6.7, meaning that *cherry* and *computer* co-occur on Wikipedia less often than we would expect by chance, and with PPMI we replace negative values by zero.

events is to slightly change the computation for $P(c)$, using a different function $P_\alpha(c)$ that raises the probability of the context word to the power of α:

$$\text{PPMI}_\alpha(w,c) = \max(\log_2 \frac{P(w,c)}{P(w)P_\alpha(c)}, 0) \tag{6.21}$$

$$P_\alpha(c) = \frac{count(c)^\alpha}{\sum_c count(c)^\alpha} \tag{6.22}$$

Levy et al. (2015) found that a setting of $\alpha = 0.75$ improved performance of embeddings on a wide range of tasks (drawing on a similar weighting used for skip-grams described below in Eq. 6.32). This works because raising the count to $\alpha = 0.75$ increases the probability assigned to rare contexts, and hence lowers their PMI ($P_\alpha(c) > P(c)$ when c is rare).

Another possible solution is Laplace smoothing: Before computing PMI, a small constant k (values of 0.1-3 are common) is added to each of the counts, shrinking (discounting) all the non-zero values. The larger the k, the more the non-zero counts are discounted.

6.7 Applications of the tf-idf or PPMI vector models

In summary, the vector semantics model we've described so far represents a target word as a vector with dimensions corresponding either to to the documents in a large collection (the term-document matrix) or to the counts of words in some neighboring window (the term-term matrix). The values in each dimension are counts, weighted by tf-idf (for term-document matrices) or PPMI (for term-term matrices), and the vectors are sparse (since most values are zero).

The model computes the similarity between two words x and y by taking the cosine of their tf-idf or PPMI vectors; high cosine, high similarity. This entire model is sometimes referred to as the **tf-idf** model or the **PPMI** model, after the weighting function.

The tf-idf model of meaning is often used for document functions like deciding if two documents are similar. We represent a document by taking the vectors of all the words in the document, and computing the **centroid** of all those vectors. The centroid is the multidimensional version of the mean; the centroid of a set of vectors is a single vector that has the minimum sum of squared distances to each of the vectors in the set. Given k word vectors $w_1, w_2, ..., w_k$, the centroid **document vector** d is:

centroid

document vector

$$d = \frac{w_1 + w_2 + ... + w_k}{k} \tag{6.23}$$

Given two documents, we can then compute their document vectors d_1 and d_2, and estimate the similarity between the two documents by $\cos(d_1, d_2)$. Document similarity is also useful for all sorts of applications; information retrieval, plagiarism detection, news recommender systems, and even for digital humanities tasks like comparing different versions of a text to see which are similar to each other.

Either the PPMI model or the tf-idf model can be used to compute word similarity, for tasks like finding word paraphrases, tracking changes in word meaning, or automatically discovering meanings of words in different corpora. For example, we can find the 10 most similar words to any target word w by computing the cosines between w and each of the $V-1$ other words, sorting, and looking at the top 10.

6.8 Word2vec

In the previous sections we saw how to represent a word as a sparse, long vector with dimensions corresponding to words in the vocabulary or documents in a collection. We now introduce a more powerful word representation: **embeddings**, short dense vectors. Unlike the vectors we've seen so far, embeddings are **short**, with number of dimensions d ranging from 50-1000, rather than the much larger vocabulary size $|V|$ or number of documents D we've seen. These d dimensions don't have a clear interpretation. And the vectors are **dense**: instead of vector entries being sparse, mostly-zero counts or functions of counts, the values will be real-valued numbers that can be negative.

It turns out that dense vectors work better in every NLP task than sparse vectors. While we don't completely understand all the reasons for this, we have some intuitions. Representing words as 300-dimensional dense vectors requires our classifiers to learn far fewer weights than if we represented words as 50,000-dimensional vectors, and the smaller parameter space possibly helps with generalization and avoiding overfitting. Dense vectors may also do a better job of capturing synonymy. For example, in a sparse vector representation, dimensions for synonyms like *car* and *automobile* dimension are distinct and unrelated; sparse vectors may thus fail to capture the similarity between a word with *car* as a neighbor and a word with *automobile* as a neighbor.

skip-gram
SGNS
word2vec

In this section we introduce one method for computing embeddings: **skip-gram with negative sampling**, sometimes called **SGNS**. The skip-gram algorithm is one of two algorithms in a software package called **word2vec**, and so sometimes the algorithm is loosely referred to as word2vec (Mikolov et al. 2013, Mikolov et al. 2013a). The word2vec methods are fast, efficient to train, and easily available online with code and pretrained embeddings. Word2vec embeddings are **static embeddings**, meaning that the method learns one fixed embedding for each word in the vocabulary. In Chapter 10 we'll introduce methods for learning dynamic **contextual embeddings** like the popular **BERT** or **ELMO** representations, in which the vector for each word is different in different contexts.

static embeddings

The intuition of word2vec is that instead of counting how often each word w occurs near, say, *apricot*, we'll instead train a classifier on a binary prediction task: "Is word w likely to show up near *apricot*?" We don't actually care about this prediction task; instead we'll take the learned classifier *weights* as the word embeddings.

The revolutionary intuition here is that we can just use running text as implicitly supervised training data for such a classifier; a word c that occurs near the target word *apricot* acts as gold 'correct answer' to the question "Is word c likely to show

self-supervision

up near *apricot*?" This method, often called **self-supervision**, avoids the need for any sort of hand-labeled supervision signal. This idea was first proposed in the task of neural language modeling, when Bengio et al. (2003) and Collobert et al. (2011) showed that a neural language model (a neural network that learned to predict the next word from prior words) could just use the next word in running text as its supervision signal, and could be used to learn an embedding representation for each word as part of doing this prediction task.

We'll see how to do neural networks in the next chapter, but word2vec is a much simpler model than the neural network language model, in two ways. First, word2vec simplifies the task (making it binary classification instead of word prediction). Second, word2vec simplifies the architecture (training a logistic regression classifier instead of a multi-layer neural network with hidden layers that demand more sophisticated training algorithms). The intuition of skip-gram is:

1. Treat the target word and a neighboring context word as positive examples.
2. Randomly sample other words in the lexicon to get negative samples.
3. Use logistic regression to train a classifier to distinguish those two cases.
4. Use the learned weights as the embeddings.

6.8.1 The classifier

Let's start by thinking about the classification task, and then turn to how to train. Imagine a sentence like the following, with a target word *apricot*, and assume we're using a window of ±2 context words:

```
... lemon,  a [tablespoon of apricot jam,      a] pinch ...
              c1         c2    w     c3       c4
```

Our goal is to train a classifier such that, given a tuple (w,c) of a target word w paired with a candidate context word c (for example (*apricot*, *jam*), or perhaps (*apricot*, *aardvark*)) it will return the probability that c is a real context word (true for *jam*, false for *aardvark*):

$$P(+|w,c) \tag{6.24}$$

The probability that word c is not a real context word for w is just 1 minus Eq. 6.24:

$$P(-|w,c) = 1 - P(+|w,c) \tag{6.25}$$

How does the classifier compute the probability P? The intuition of the skip-gram model is to base this probability on embedding similarity: a word is likely to occur near the target if its embedding is similar to the target embedding. To compute similarity between these dense embeddings, we rely on the intuition that two vectors are similar if they have a high **dot product** (after all, cosine is just a normalized dot product). In other words:

$$Similarity(w,c) \approx c \cdot w \tag{6.26}$$

The dot product $c \cdot w$ is not a probability, it's just a number ranging from $-\infty$ to ∞ (since the elements in word2vec embeddings can be negative, the dot product can be negative). To turn the dot product into a probability, we'll use the **logistic** or **sigmoid** function $\sigma(x)$, the fundamental core of logistic regression:

$$\sigma(x) = \frac{1}{1+\exp(-x)} \tag{6.27}$$

We model the probability that word c is a real context word for target word w as:

$$P(+|w,c) = \sigma(c \cdot w) = \frac{1}{1+\exp(-c \cdot w)} \tag{6.28}$$

The sigmoid function returns a number between 0 and 1, but to make it a probability we'll also need the total probability of the two possible events (c is a context word, and c isn't a context word) to sum to 1. We thus estimate the probability that word c is not a real context word for w as:

$$\begin{aligned} P(-|w,c) &= 1 - P(+|w,c) \\ &= \sigma(-c \cdot w) = \frac{1}{1+\exp(c \cdot w)} \end{aligned} \tag{6.29}$$

Equation 6.28 gives us the probability for one word, but there are many context words in the window. Skip-gram makes the simplifying assumption that all context words are independent, allowing us to just multiply their probabilities:

$$P(+|w,c_{1:L}) = \prod_{i=1}^{L} \sigma(-c_i \cdot w) \tag{6.30}$$

$$\log P(+|w,c_{1:L}) = \sum_{i=1}^{L} \log \sigma(-c_i \cdot w) \tag{6.31}$$

In summary, skip-gram trains a probabilistic classifier that, given a test target word w and its context window of L words $c_{1:L}$, assigns a probability based on how similar this context window is to the target word. The probability is based on applying the logistic (sigmoid) function to the dot product of the embeddings of the target word with each context word. To compute this probability, we just need embeddings for each target word and context word in the vocabulary.

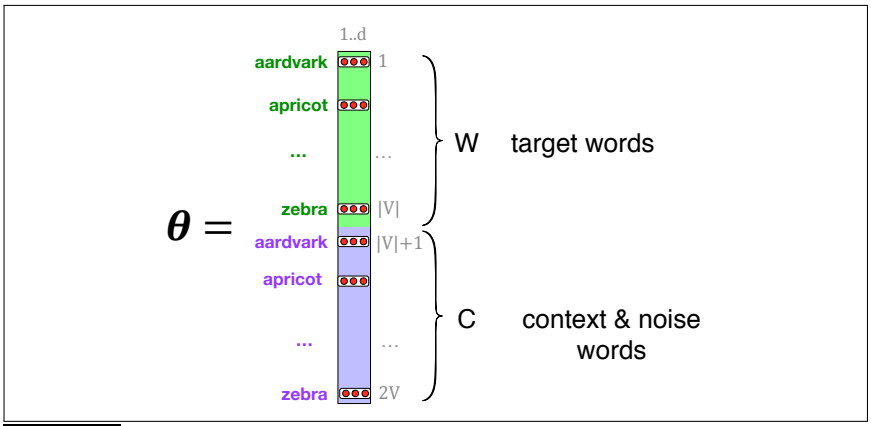


Figure 6.13 The embeddings learned by the skipgram model. The algorithm stores two embeddings for each word, the target embedding (sometimes called the input embedding) and the context embedding (sometimes called the output embedding). The parameter θ that the algorithm learns is thus a matrix of $2|V|$ vectors, each of dimension d, formed by concatenating two matrices, the target embeddings W and the context+noise embeddings C.

Fig. 6.13 shows the intuition of the parameters we'll need. Skip-gram actually stores two embeddings for each word, one for the word as a target, and one for the

word considered as context. Thus the parameters we need to learn are two matrices W and C, each containing an embedding for every one of the $|V|$ words in the vocabulary V.[6] Let's now turn to learning these embeddings (which is the real goal of training this classifier in the first place).

6.8.2 Learning skip-gram embeddings

Skip-gram learns embeddings by starting with random embedding vectors and then iteratively shifting the embedding of each word w to be more like the embeddings of words that occur nearby in texts, and less like the embeddings of words that don't occur nearby. Let's start by considering a single piece of training data:

```
... lemon,  a [tablespoon of apricot jam,      a] pinch ...
              c1         c2    w     c3        c4
```

This example has a target word w (apricot), and 4 context words in the $L = \pm 2$ window, resulting in 4 positive training instances (on the left below):

positive examples +

w	c_{pos}
apricot	tablespoon
apricot	of
apricot	jam
apricot	a

negative examples -

w	c_{neg}	w	c_{neg}
apricot	aardvark	apricot	seven
apricot	my	apricot	forever
apricot	where	apricot	dear
apricot	coaxial	apricot	if

For training a binary classifier we also need negative examples. In fact skip-gram with negative sampling (SGNS) uses more negative examples than positive examples (with the ratio between them set by a parameter k). So for each of these (w, c_{pos}) training instances we'll create k negative samples, each consisting of the target w plus a 'noise word' c_{neg}. A noise word is a random word from the lexicon, constrained not to be the target word w. The right above shows the setting where $k = 2$, so we'll have 2 negative examples in the negative training set $-$ for each positive example w, c_{pos}.

The noise words are chosen according to their weighted unigram frequency $p_\alpha(w)$, where α is a weight. If we were sampling according to unweighted frequency $p(w)$, it would mean that with unigram probability $p(\text{"}the\text{"})$ we would choose the word *the* as a noise word, with unigram probability $p(\text{"}aardvark\text{"})$ we would choose *aardvark*, and so on. But in practice it is common to set $\alpha = .75$, i.e. use the weighting $p^{\frac{3}{4}}(w)$:

$$P_\alpha(w) = \frac{count(w)^\alpha}{\sum_{w'} count(w')^\alpha} \tag{6.32}$$

Setting $\alpha = .75$ gives better performance because it gives rare noise words slightly higher probability: for rare words, $P_\alpha(w) > P(w)$. To illustrate this intuition, it might help to work out the probabilities for an example with two events, $P(a) = .99$ and $P(b) = .01$:

$$\begin{aligned} P_\alpha(a) &= \frac{.99^{.75}}{.99^{.75} + .01^{.75}} = .97 \\ P_\alpha(b) &= \frac{.01^{.75}}{.99^{.75} + .01^{.75}} = .03 \end{aligned} \tag{6.33}$$

[6] In principle the target matrix and the context matrix could use different vocabularies, but we'll simplify by assuming one shared vocabulary V.

Given the set of positive and negative training instances, and an initial set of embeddings, the goal of the learning algorithm is to adjust those embeddings to

- Maximize the similarity of the target word, context word pairs (w, c_{pos}) drawn from the positive examples
- Minimize the similarity of the (w, c_{neg}) pairs from the negative examples.

If we consider one word/context pair (w, c_{pos}) with its k noise words $c_{neg_1} \ldots c_{neg_k}$, we can express these two goals as the following loss function L to be minimized (hence the $-$); here the first term expresses that we want the classifier to assign the real context word c_{pos} a high probability of being a neighbor, and the second term expresses that we want to assign each of the noise words c_{neg_i} a high probability of being a non-neighbor, all multiplied because we assume independence:

$$\begin{aligned} L_{CE} &= -\log \left[P(+|w, c_{pos}) \prod_{i=1}^{k} P(-|w, c_{neg_i}) \right] \\ &= -\left[\log P(+|w, c_{pos}) + \sum_{i=1}^{k} \log P(-|w, c_{neg_i}) \right] \\ &= -\left[\log P(+|w, c_{pos}) + \sum_{i=1}^{k} \log \left(1 - P(+|w, c_{neg_i})\right) \right] \\ &= -\left[\log \sigma(c_{pos} \cdot w) + \sum_{i=1}^{k} \log \sigma(-c_{neg_i} \cdot w) \right] \end{aligned} \tag{6.34}$$

That is, we want to maximize the dot product of the word with the actual context words, and minimize the dot products of the word with the k negative sampled non-neighbor words.

We minimize this loss function using stochastic gradient descent. Fig. 6.14 shows the intuition of one step of learning.

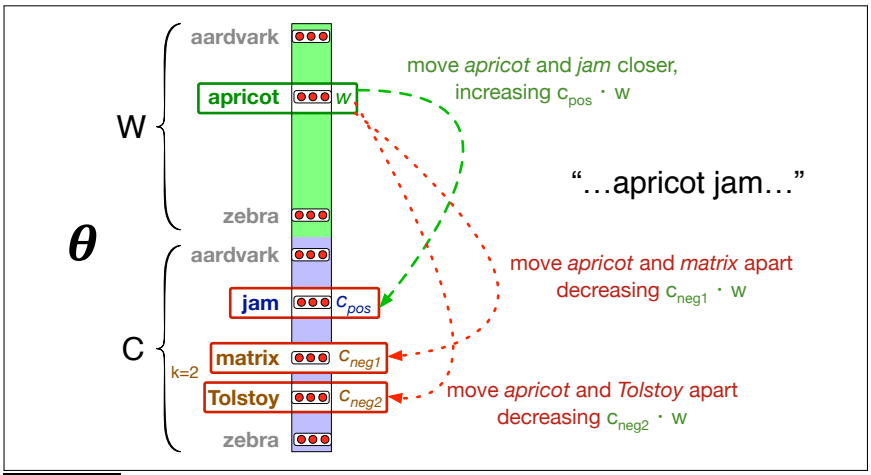


Figure 6.14 Intuition of one step of gradient descent. The skip-gram model tries to shift embeddings so the target embeddings (here for *apricot*) are closer to (have a higher dot product with) context embeddings for nearby words (here *jam*) and further from (lower dot product with) context embeddings for noise words that don't occur nearby (here *Tolstoy* and *matrix*).

To get the gradient, we need to take the derivative of Eq. 6.34 with respect to the different embeddings. It turns out the derivatives are the following (we leave the

proof as an exercise at the end of the chapter):

$$\frac{\partial L_{CE}}{\partial c_{pos}} = [\sigma(c_{pos} \cdot w) - 1]w \tag{6.35}$$

$$\frac{\partial L_{CE}}{\partial c_{neg}} = [\sigma(c_{neg} \cdot w)]w \tag{6.36}$$

$$\frac{\partial L_{CE}}{\partial w} = [\sigma(c_{pos} \cdot w) - 1]c_{pos} + \sum_{i=1}^{k}[\sigma(c_{\text{neg}_i} \cdot w)]c_{neg_i} \tag{6.37}$$

The update equations going from time step t to $t+1$ in stochastic gradient descent are thus:

$$c_{pos}^{t+1} = c_{pos}^{t} - \eta[\sigma(c_{pos}^{t} \cdot w) - 1]w \tag{6.38}$$

$$c_{neg}^{t+1} = c_{neg}^{t} - \eta[\sigma(c_{neg}^{t} \cdot w)]w \tag{6.39}$$

$$w^{t+1} = w^{t} - \eta[\sigma(c_{pos} \cdot w^{t}) - 1]c_{pos} + \sum_{i=1}^{k}[\sigma(c_{\text{neg}_i} \cdot w^{t})]c_{neg_i} \tag{6.40}$$

Just as in logistic regression, then, the learning algorithm starts with randomly initialized W and C matrices, and then walks through the training corpus using gradient descent to move W and C so as to maximize the objective in Eq. 6.34 by making the updates in (Eq. 6.39)-(Eq. 6.40).

target embedding
context embedding

Recall that the skip-gram model learns **two** separate embeddings for each word i: the **target embedding** w_i and the **context embedding** c_i, stored in two matrices, the **target matrix** W and the **context matrix** C. It's common to just add them together, representing word i with the vector $w_i + c_i$. Alternatively we can throw away the C matrix and just represent each word i by the vector w_i.

As with the simple count-based methods like tf-idf, the context window size L affects the performance of skip-gram embeddings, and experiments often tune the parameter L on a devset.

6.8.3 Other kinds of static embeddings

fasttext

There are many kinds of static embeddings. An extension of word2vec, **fasttext** (Bojanowski et al., 2017), deals with unknown words and sparsity in languages with rich morphology, by using subword models. Each word in fasttext is represented as itself plus a bag of constituent n-grams, with special boundary symbols < and > added to each word. For example, with $n = 3$ the word *where* would be represented by the sequence <where> plus the character n-grams:

```
<wh, whe, her, ere, re>
```

Then a skipgram embedding is learned for each constituent n-gram, and the word *where* is represented by the sum of all of the embeddings of its constituent n-grams. A fasttext open-source library, including pretrained embeddings for 157 languages, is available at https://fasttext.cc.

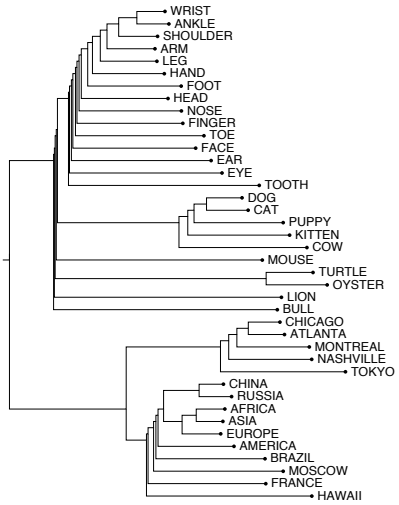
The most widely used static embedding model besides word2vec is GloVe (Pennington et al., 2014), short for Global Vectors, because the model is based on capturing global corpus statistics. GloVe is based on ratios of probabilities from the word-word co-occurrence matrix, combining the intuitions of count-based models like PPMI while also capturing the linear structures used by methods like word2vec.

It turns out that dense embeddings like word2vec actually have an elegant mathematical relationships with sparse embeddings like PPMI, in which word2vec can be seen as implicitly optimizing a shifted version of a PPMI matrix (Levy and Goldberg, 2014c).

6.9 Visualizing Embeddings

"I see well in many dimensions as long as the dimensions are around two."
The late economist Martin Shubik

Visualizing embeddings is an important goal in helping understand, apply, and improve these models of word meaning. But how can we visualize a (for example) 100-dimensional vector?



The simplest way to visualize the meaning of a word w embedded in a space is to list the most similar words to w by sorting the vectors for all words in the vocabulary by their cosine with the vector for w. For example the 7 closest words to *frog* using the GloVe embeddings are: *frogs*, *toad*, *litoria*, *leptodactylidae*, *rana*, *lizard*, and *eleutherodactylus* (Pennington et al., 2014).

Yet another visualization method is to use a clustering algorithm to show a hierarchical representation of which words are similar to others in the embedding space. The uncaptioned figure on the left uses hierarchical clustering of some embedding vectors for nouns as a visualization method (Rohde et al., 2006).

Probably the most common visualization method, however, is to project the 100 dimensions of a word down into 2 dimensions. Fig. 6.1 showed one such visualization, as does Fig. 6.16, using a projection method called t-SNE (van der Maaten and Hinton, 2008).

6.10 Semantic properties of embeddings

In this section we briefly summarize some of the semantic properties of embeddings that have been studied.

Different types of similarity or association: One parameter of vector semantic models that is relevant to both sparse tf-idf vectors and dense word2vec vectors is the size of the context window used to collect counts. This is generally between 1 and 10 words on each side of the target word (for a total context of 2-20 words).

The choice depends on the goals of the representation. Shorter context windows tend to lead to representations that are a bit more syntactic, since the information is coming from immediately nearby words. When the vectors are computed from short context windows, the most similar words to a target word w tend to be semantically similar words with the same parts of speech. When vectors are computed from long context windows, the highest cosine words to a target word w tend to be words that are topically related but not similar.

For example Levy and Goldberg (2014a) showed that using skip-gram with a window of ±2, the most similar words to the word *Hogwarts* (from the *Harry Potter* series) were names of other fictional schools: *Sunnydale* (from *Buffy the Vampire Slayer*) or *Evernight* (from a vampire series). With a window of ±5, the most similar words to *Hogwarts* were other words topically related to the *Harry Potter* series: *Dumbledore*, *Malfoy*, and *half-blood*.

first-order co-occurrence

It's also often useful to distinguish two kinds of similarity or association between words (Schütze and Pedersen, 1993). Two words have **first-order co-occurrence** (sometimes called **syntagmatic association**) if they are typically nearby each other. Thus *wrote* is a first-order associate of *book* or *poem*. Two words have **second-order co-occurrence** (sometimes called **paradigmatic association**) if they have similar neighbors. Thus *wrote* is a second-order associate of words like *said* or *remarked*.

second-order co-occurrence

Analogy/Relational Similarity: Another semantic property of embeddings is their ability to capture relational meanings. In an important early vector space model of cognition, Rumelhart and Abrahamson (1973) proposed the **parallelogram model** for solving simple analogy problems of the form *a is to b as a* is to what?*. In such problems, a system given a problem like *apple:tree::grape:?*, i.e., *apple is to tree as grape is to* ____, and must fill in the word *vine*. In the parallelogram model, illustrated in Fig. 6.15, the vector from the word *apple* to the word *tree* ($= \overrightarrow{\text{apple}} - \overrightarrow{\text{tree}}$) is added to the vector for *grape* ($\overrightarrow{\text{grape}}$); the nearest word to that point is returned.

parallelogram model

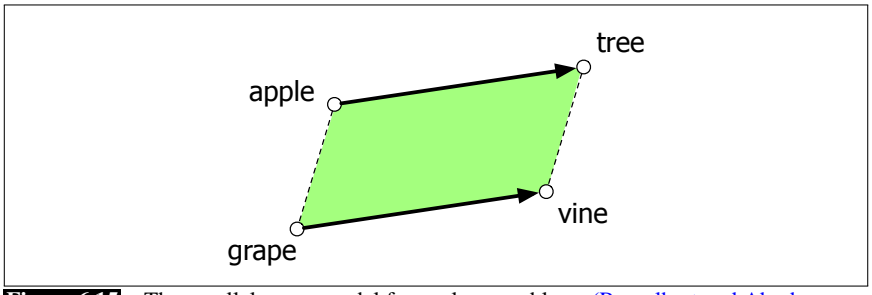


Figure 6.15 The parallelogram model for analogy problems (Rumelhart and Abrahamson, 1973): the location of $\overrightarrow{\text{vine}}$ can be found by subtracting $\overrightarrow{\text{tree}}$ from $\overrightarrow{\text{apple}}$ and adding $\overrightarrow{\text{grape}}$.

In early work with sparse embeddings, scholars showed that sparse vector models of meaning could solve such analogy problems (Turney and Littman, 2005), but the parallelogram method received more modern attention because of its success with word2vec or GloVe vectors (Mikolov et al. 2013b, Levy and Goldberg 2014b, Pennington et al. 2014). For example, the result of the expression $(\overrightarrow{\text{king}}) - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}}$ is a vector close to $\overrightarrow{\text{queen}}$. Similarly, $\overrightarrow{\text{Paris}} - \overrightarrow{\text{France}} + \overrightarrow{\text{Italy}}$) results in a vector that is close to $\overrightarrow{\text{Rome}}$. The embedding model thus seems to be extracting representations of relations like MALE-FEMALE, or CAPITAL-CITY-OF, or even COMPARATIVE/SUPERLATIVE, as shown in Fig. 6.16 from GloVe.

For a *a:b::a*:b** problem, meaning the algorithm is given *a, b,* and *a** and must find *b**, the parallelogram method is thus:

$$\hat{b}^* = \underset{x}{\text{argmax}}\ \text{distance}(x, a^* - a + b) \tag{6.41}$$

with the distance function defined either as cosine or as Euclidean distance.

There are some caveats. For example, the closest value returned by the parallelogram algorithm in word2vec or GloVe embedding spaces is usually not in fact b* but one of the 3 input words or their morphological variants (i.e., *cherry:red ::*

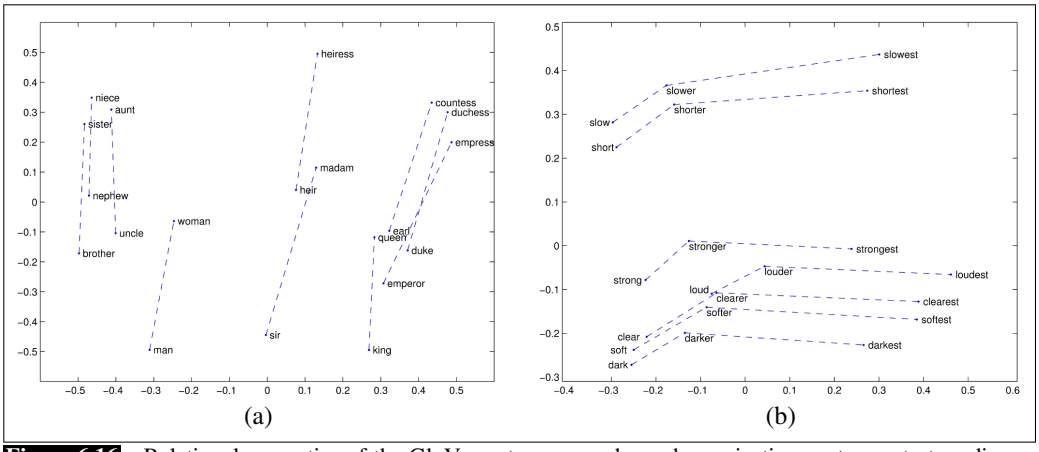


Figure 6.16 Relational properties of the GloVe vector space, shown by projecting vectors onto two dimensions. (a) $\overrightarrow{(\text{king})} - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}}$ is close to $\overrightarrow{\text{queen}}$. (b) offsets seem to capture comparative and superlative morphology (Pennington et al., 2014).

potato:x returns *potato* or *potatoes* instead of *brown*), so these must be explicitly excluded. Furthermore while embedding spaces perform well if the task involves frequent words, small distances, and certain relations (like relating countries with their capitals or verbs/nouns with their inflected forms), the parallelogram method with embeddings doesn't work as well for other relations (Linzen 2016, Gladkova et al. 2016, Ethayarajh et al. 2019a), and indeed Peterson et al. (2020) argue that the parallelogram method is in general too simple to model the human cognitive process of forming analogies of this kind.

6.10.1 Embeddings and Historical Semantics

Embeddings can also be a useful tool for studying how meaning changes over time, by computing multiple embedding spaces, each from texts written in a particular time period. For example Fig. 6.17 shows a visualization of changes in meaning in English words over the last two centuries, computed by building separate embedding spaces for each decade from historical corpora like Google N-grams (Lin et al., 2012) and the Corpus of Historical American English (Davies, 2012).

6.11 Bias and Embeddings

In addition to their ability to learn word meaning from text, embeddings, alas, also reproduce the implicit biases and stereotypes that were latent in the text. As the prior section just showed, embeddings can roughly model relational similarity: 'queen' as the closest word to 'king' - 'man' + 'woman' implies the analogy *man:woman::king:queen*. But these same embedding analogies also exhibit gender stereotypes. For example Bolukbasi et al. (2016) find that the closest occupation to 'man' - 'computer programmer' + 'woman' in word2vec embeddings trained on news text is 'homemaker', and that the embeddings similarly suggest the analogy 'father' is to 'doctor' as 'mother' is to 'nurse'. This could result in what Crawford (2017) and Blodgett et al. (2020) call an **allocational harm**, when a system allocates resources (jobs or credit) unfairly to different groups. For example algorithms

allocational harm

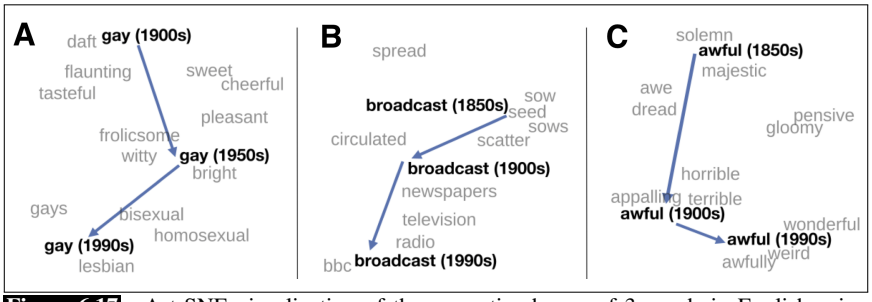


Figure 6.17 A t-SNE visualization of the semantic change of 3 words in English using word2vec vectors. The modern sense of each word, and the grey context words, are computed from the most recent (modern) time-point embedding space. Earlier points are computed from earlier historical embedding spaces. The visualizations show the changes in the word *gay* from meanings related to "cheerful" or "frolicsome" to referring to homosexuality, the development of the modern "transmission" sense of *broadcast* from its original sense of sowing seeds, and the pejoration of the word *awful* as it shifted from meaning "full of awe" to meaning "terrible or appalling" (Hamilton et al., 2016).

that use embeddings as part of a search for hiring potential programmers or doctors might thus incorrectly downweight documents with women's names.

bias amplification

It turns out that embeddings don't just reflect the statistics of their input, but also **amplify** bias; gendered terms become **more** gendered in embedding space than they were in the input text statistics (Zhao et al. 2017, Ethayarajh et al. 2019b, Jia et al. 2020), and biases are more exaggerated than in actual labor employment statistics (Garg et al., 2018).

Embeddings also encode the implicit associations that are a property of human reasoning. The Implicit Association Test (Greenwald et al., 1998) measures people's associations between concepts (like 'flowers' or 'insects') and attributes (like 'pleasantness' and 'unpleasantness') by measuring differences in the latency with which they label words in the various categories.[7] Using such methods, people in the United States have been shown to associate African-American names with unpleasant words (more than European-American names), male names more with mathematics and female names with the arts, and old people's names with unpleasant words (Greenwald et al. 1998, Nosek et al. 2002a, Nosek et al. 2002b). Caliskan et al. (2017) replicated all these findings of implicit associations using GloVe vectors and cosine similarity instead of human latencies. For example African-American names like 'Leroy' and 'Shaniqua' had a higher GloVe cosine with unpleasant words while European-American names ('Brad', 'Greg', 'Courtney') had a higher cosine with pleasant words. These problems with embeddings are an example of a **representational harm** (Crawford 2017, Blodgett et al. 2020), which is a harm caused by a system demeaning or even ignoring some social groups. Any embedding-aware algorithm that made use of word sentiment could thus exacerbate bias against African Americans.

representational harm

Recent research focuses on ways to try to remove these kinds of biases, for example by developing a transformation of the embedding space that removes gender stereotypes but preserves definitional gender (Bolukbasi et al. 2016, Zhao et al. 2017)

[7] Roughly speaking, if humans associate 'flowers' with 'pleasantness' and 'insects' with 'unpleasantness', when they are instructed to push a green button for 'flowers' (daisy, iris, lilac) and 'pleasant words' (love, laughter, pleasure) and a red button for 'insects' (flea, spider, mosquito) and 'unpleasant words' (abuse, hatred, ugly) they are faster than in an incongruous condition where they push a red button for 'flowers' and 'unpleasant words' and a green button for 'insects' and 'pleasant words'.

debiasing

or changing the training procedure (Zhao et al., 2018). However, although these sorts of **debiasing** may reduce bias in embeddings, they do not eliminate it (Gonen and Goldberg, 2019), and this remains an open problem.

Historical embeddings are also being used to measure biases in the past. Garg et al. (2018) used embeddings from historical texts to measure the association between embeddings for occupations and embeddings for names of various ethnicities or genders (for example the relative cosine similarity of women's names versus men's to occupation words like 'librarian' or 'carpenter') across the 20th century. They found that the cosines correlate with the empirical historical percentages of women or ethnic groups in those occupations. Historical embeddings also replicated old surveys of ethnic stereotypes; the tendency of experimental participants in 1933 to associate adjectives like 'industrious' or 'superstitious' with, e.g., Chinese ethnicity, correlates with the cosine between Chinese last names and those adjectives using embeddings trained on 1930s text. They also were able to document historical gender biases, such as the fact that embeddings for adjectives related to competence ('smart', 'wise', 'thoughtful', 'resourceful') had a higher cosine with male than female words, and showed that this bias has been slowly decreasing since 1960. We return in later chapters to this question about the role of bias in natural language processing.

6.12 Evaluating Vector Models

The most important evaluation metric for vector models is extrinsic evaluation on tasks, i.e., using vectors in an NLP task and seeing whether this improves performance over some other model.

Nonetheless it is useful to have intrinsic evaluations. The most common metric is to test their performance on **similarity**, computing the correlation between an algorithm's word similarity scores and word similarity ratings assigned by humans. **WordSim-353** (Finkelstein et al., 2002) is a commonly used set of ratings from 0 to 10 for 353 noun pairs; for example (*plane*, *car*) had an average score of 5.77. **SimLex-999** (Hill et al., 2015) is a more difficult dataset that quantifies similarity (*cup, mug*) rather than relatedness (*cup, coffee*), and including both concrete and abstract adjective, noun and verb pairs. The **TOEFL dataset** is a set of 80 questions, each consisting of a target word with 4 additional word choices; the task is to choose which is the correct synonym, as in the example: *Levied is closest in meaning to: imposed, believed, requested, correlated* (Landauer and Dumais, 1997). All of these datasets present words without context.

Slightly more realistic are intrinsic similarity tasks that include context. The Stanford Contextual Word Similarity (SCWS) dataset (Huang et al., 2012) and the Word-in-Context (WiC) dataset (Pilehvar and Camacho-Collados, 2019) offer richer evaluation scenarios. SCWS gives human judgments on 2,003 pairs of words in their sentential context, while WiC gives target words in two sentential contexts that are either in the same or different senses; see Section **??**. The *semantic textual similarity* task (Agirre et al. 2012, Agirre et al. 2015) evaluates the performance of sentence-level similarity algorithms, consisting of a set of pairs of sentences, each pair with human-labeled similarity scores.

Another task used for evaluation is the analogy task, discussed on page 24, where the system has to solve problems of the form *a is to b as a* is to b**, given *a, b,* and *a** and having to find *b** (Turney and Littman, 2005). A number of sets of tuples have

been created for this task, (Mikolov et al. 2013, Mikolov et al. 2013b, Gladkova et al. 2016), covering morphology (*city:cities::child:children*), lexicographic relations (*leg:table::spout::teapot*) and encyclopedia relations (*Beijing:China::Dublin:Ireland*), some drawing from the SemEval-2012 Task 2 dataset of 79 different relations (Jurgens et al., 2012).

All embedding algorithms suffer from inherent variability. For example because of randomness in the initialization and the random negative sampling, algorithms like word2vec may produce different results even from the same dataset, and individual documents in a collection may strongly impact the resulting embeddings (Hellrich and Hahn 2016, Antoniak and Mimno 2018). When embeddings are used to study word associations in particular corpora, therefore, it is best practice to train multiple embeddings with bootstrap sampling over documents and average the results (Antoniak and Mimno, 2018).

6.13 Summary

- In vector semantics, a word is modeled as a vector—a point in high-dimensional space, also called an **embedding**. In this chapter we focus on **static embeddings**, in each each word is mapped to a fixed embedding.
- Vector semantic models fall into two classes: **sparse** and **dense**. In sparse models each dimension corresponds to a word in the vocabulary V and cells are functions of **co-occurrence counts**. The **term-document** matrix has a row for each word (**term**) in the vocabulary and a column for each document. The **word-context** or **term-term** matrix has a row for each (target) word in the vocabulary and a column for each context term in the vocabulary. Two sparse weightings are common: the **tf-idf** weighting which weights each cell by its **term frequency** and **inverse document frequency**, and **PPMI** (pointwise positive mutual information) most common for for word-context matrices.
- Dense vector models have dimensionality 50–1000. **Word2vec** algorithms like **skip-gram** are a popular way to compute dense embeddings. Skip-gram trains a logistic regression classifier to compute the probability that two words are 'likely to occur nearby in text'. This probability is computed from the dot product between the embeddings for the two words.
- Skip-gram uses stochastic gradient descent to train the classifier, by learning embeddings that have a high dot product with embeddings of words that occur nearby and a low dot product with noise words.
- Other important embedding algorithms include **GloVe**, a method based on ratios of word co-occurrence probabilities.
- Whether using sparse or dense vectors, word and document similarities are computed by some function of the **dot product** between vectors. The cosine of two vectors—a normalized dot product—is the most popular such metric.

Bibliographical and Historical Notes

The idea of vector semantics arose out of research in the 1950s in three distinct fields: linguistics, psychology, and computer science, each of which contributed a

fundamental aspect of the model.

The idea that meaning is related to the distribution of words in context was widespread in linguistic theory of the 1950s, among distributionalists like Zellig Harris, Martin Joos, and J. R. Firth, and semioticians like Thomas Sebeok. As Joos (1950) put it,

> the linguist's "meaning" of a morpheme... is by definition the set of conditional probabilities of its occurrence in context with all other morphemes.

The idea that the meaning of a word might be modeled as a point in a multidimensional semantic space came from psychologists like Charles E. Osgood, who had been studying how people responded to the meaning of words by assigning values along scales like *happy/sad* or *hard/soft*. Osgood et al. (1957) proposed that the meaning of a word in general could be modeled as a point in a multidimensional Euclidean space, and that the similarity of meaning between two words could be modeled as the distance between these points in the space.

mechanical indexing

A final intellectual source in the 1950s and early 1960s was the field then called **mechanical indexing**, now known as **information retrieval**. In what became known as the **vector space model** for information retrieval (Salton 1971, Sparck Jones 1986), researchers demonstrated new ways to define the meaning of words in terms of vectors (Switzer, 1965), and refined methods for word similarity based on measures of statistical association between words like mutual information (Giuliano, 1965) and idf (Sparck Jones, 1972), and showed that the meaning of documents could be represented in the same vector spaces used for words.

Some of the philosophical underpinning of the distributional way of thinking came from the late writings of the philosopher Wittgenstein, who was skeptical of the possibility of building a completely formal theory of meaning definitions for each word, suggesting instead that "the meaning of a word is its use in the language" (Wittgenstein, 1953, PI 43). That is, instead of using some logical language to define each word, or drawing on denotations or truth values, Wittgenstein's idea is that we should define a word by how it is used by people in speaking and understanding in their day-to-day interactions, thus prefiguring the movement toward embodied and experiential models in linguistics and NLP (Glenberg and Robertson 2000, Lake and Murphy 2020, Bisk et al. 2020, Bender and Koller 2020).

More distantly related is the idea of defining words by a vector of discrete features, which has roots at least as far back as Descartes and Leibniz (Wierzbicka 1992, Wierzbicka 1996). By the middle of the 20th century, beginning with the work of Hjelmslev (Hjelmslev, 1969) (originally 1943) and fleshed out in early models of generative grammar (Katz and Fodor, 1963), the idea arose of representing meaning with **semantic features**, symbols that represent some sort of primitive meaning. For example words like *hen*, *rooster*, or *chick*, have something in common (they all describe chickens) and something different (their age and sex), representable as:

semantic feature

hen	`+female, +chicken, +adult`
rooster	`-female, +chicken, +adult`
chick	`+chicken, -adult`

The dimensions used by vector models of meaning to define words, however, are only abstractly related to this idea of a small fixed number of hand-built dimensions. Nonetheless, there has been some attempt to show that certain dimensions of embedding models do contribute some specific compositional aspect of meaning like these early semantic features.

The use of dense vectors to model word meaning, and indeed the term **embedding**, grew out of the **latent semantic indexing** (LSI) model (Deerwester et al.,

1988) recast as **LSA** (**latent semantic analysis**) (Deerwester et al., 1990). In LSA
SVD **singular value decomposition**—**SVD**— is applied to a term-document matrix (each cell weighted by log frequency and normalized by entropy), and then the first 300 dimensions are used as the LSA embedding. Singular Value Decomposition (SVD) is a method for finding the most important dimensions of a data set, those dimensions along which the data varies the most. LSA was then quickly widely applied: as a cognitive model Landauer and Dumais (1997), and for tasks like spell checking (Jones and Martin, 1997), language modeling (Bellegarda 1997, Coccaro and Jurafsky 1998, Bellegarda 2000) morphology induction (Schone and Jurafsky 2000, Schone and Jurafsky 2001b), multiword expressions (MWEs) (Schone and Jurafsky, 2001a), and essay grading (Rehder et al., 1998). Related models were simultaneously developed and applied to word sense disambiguation by Schütze (1992). LSA also led to the earliest use of embeddings to represent words in a probabilistic classifier, in the logistic regression document router of Schütze et al. (1995). The idea of SVD on the term-term matrix (rather than the term-document matrix) as a model of meaning for NLP was proposed soon after LSA by Schütze (1992). Schütze applied the low-rank (97-dimensional) embeddings produced by SVD to the task of word sense disambiguation, analyzed the resulting semantic space, and also suggested possible techniques like dropping high-order dimensions. See Schütze (1997).

A number of alternative matrix models followed on from the early SVD work, including Probabilistic Latent Semantic Indexing (PLSI) (Hofmann, 1999), Latent Dirichlet Allocation (LDA) (Blei et al., 2003), and Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999).

The LSA community seems to have first used the word "embedding" in Landauer et al. (1997), in a variant of its mathematical meaning as a mapping from one space or mathematical structure to another. In LSA, the word embedding seems to have described the mapping from the space of sparse count vectors to the latent space of SVD dense vectors. Although the word thus originally meant the mapping from one space to another, it has metonymically shifted to mean the resulting dense vector in the latent space. and it is in this sense that we currently use the word.

By the next decade, Bengio et al. (2003) and Bengio et al. (2006) showed that neural language models could also be used to develop embeddings as part of the task of word prediction. Collobert and Weston (2007), Collobert and Weston (2008), and Collobert et al. (2011) then demonstrated that embeddings could be used to represent word meanings for a number of NLP tasks. Turian et al. (2010) compared the value of different kinds of embeddings for different NLP tasks. Mikolov et al. (2011) showed that recurrent neural nets could be used as language models. The idea of simplifying the hidden layer of these neural net language models to create the skip-gram (and also CBOW) algorithms was proposed by Mikolov et al. (2013). The negative sampling training algorithm was proposed in Mikolov et al. (2013a). There are numerous surveys of static embeddings and their parameterizations (Bullinaria and Levy 2007, Bullinaria and Levy 2012, Lapesa and Evert 2014, Kiela and Clark 2014, Levy et al. 2015).

See Manning et al. (2008) for a deeper understanding of the role of vectors in information retrieval, including how to compare queries with documents, more details on tf-idf, and issues of scaling to very large datasets. See Kim (2019) for a clear and comprehensive tutorial on word2vec. Cruse (2004) is a useful introductory linguistic text on lexical semantics.

Exercises

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L., and Wiebe, J. (2015). 2015 SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. *SemEval-15*.

Agirre, E., Diab, M., Cer, D., and Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. *SemEval-12*.

Antoniak, M. and Mimno, D. (2018). Evaluating the stability of embedding-based word similarities. *TACL 6*, 107–119.

Bellegarda, J. R. (1997). A latent semantic analysis framework for large-span language modeling. *EUROSPEECH*.

Bellegarda, J. R. (2000). Exploiting latent semantic information in statistical language modeling. *Proceedings of the IEEE 89*(8), 1279–1296.

Bender, E. M. and Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *ACL*.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*(8), 1798–1828.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research 3*(Feb), 1137–1155.

Bengio, Y., Schwenk, H., Senécal, J.-S., Morin, F., and Gauvain, J.-L. (2006). Neural probabilistic language models. *Innovations in Machine Learning*, 137–186. Springer.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., and Turian, J. (2020). Experience grounds language.. arXiv preprint arXiv:2004.10151.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *JMLR 3*(5), 993–1022.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *ACL*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *TACL 5*, 135–146.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *NeurIPS*.

Bréal, M. (1897). *Essai de Sémantique: Science des significations*. Hachette.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics 32*(1), 13–47.

Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods 39*(3), 510–526.

Bullinaria, J. A. and Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stoplists, stemming, and SVD. *Behavior research methods 44*(3), 890–907.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science 356*(6334), 183–186.

Carlson, G. N. (1977). *Reference to kinds in English*. Ph.D. thesis, University of Massachusetts, Amherst. Forward.

Church, K. W. and Hanks, P. (1989). Word association norms, mutual information, and lexicography. *ACL*.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics 16*(1), 22–29.

Clark, E. (1987). The principle of contrast: A constraint on language acquisition. MacWhinney, B. (Ed.), *Mechanisms of language acquisition*, 1–33. LEA.

Coccaro, N. and Jurafsky, D. (1998). Towards better integration of semantic predictors in statistical language modeling. *ICSLP*.

Collobert, R. and Weston, J. (2007). Fast semantic extraction using a novel neural network architecture. *ACL*.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. *ICML*.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *JMLR 12*, 2493–2537.

Crawford, K. (2017). The trouble with bias. Keynote at NeurIPS.

Cruse, D. A. (2004). *Meaning in Language: an Introduction to Semantics and Pragmatics*. Oxford University Press. Second edition.

Dagan, I., Marcus, S., and Markovitch, S. (1993). Contextual word similarity and estimation from sparse data. *ACL*.

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora 7*(2), 121–157.

Davies, M. (2015). The Wikipedia Corpus: 4.6 million articles, 1.9 billion words. Adapted from Wikipedia. https://www.english-corpora.org/wiki/.

Deerwester, S. C., Dumais, S. T., Furnas, G. W., Harshman, R. A., Landauer, T. K., Lochbaum, K. E., and Streeter, L. (1988). Computer information retrieval using latent semantic structure: US Patent 4,839,853..

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantics analysis. *JASIS 41*(6), 391–407.

Ethayarajh, K., Duvenaud, D., and Hirst, G. (2019a). Towards understanding linear word analogies. *ACL*.

Ethayarajh, K., Duvenaud, D., and Hirst, G. (2019b). Understanding undesirable word embedding associations. *ACL*.

Fano, R. M. (1961). *Transmission of Information: A Statistical Theory of Communications*. MIT Press.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems 20*(1), 116—-131.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. *Studies in Linguistic Analysis*. Philological Society. Reprinted in Palmer, F. (ed.) 1968. Selected Papers of J. R. Firth. Longman, Harlow.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences 115*(16), E3635–E3644.

Girard, G. (1718). *La justesse de la langue françoise: ou les différentes significations des mots qui passent pour synonimes*. Laurent d'Houry, Paris.

Giuliano, V. E. (1965). The interpretation of word associations. Stevens, M. E., Giuliano, V. E., and Heilprin, L. B. (Eds.), *Statistical Association Methods For Mechanized Documentation. Symposium Proceedings. Washington, D.C., USA, March 17, 1964*. https://nvlpubs.nist.gov/nistpubs/Legacy/MP/nbsmiscellaneouspub269.pdf.

Gladkova, A., Drozd, A., and Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. *NAACL Student Research Workshop*.

Glenberg, A. M. and Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of memory and language 43*(3), 379–401.

Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *NAACL HLT*.

Gould, S. J. (1980). *The Panda's Thumb*. Penguin Group.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology 74*(6), 1464–1480.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *ACL*.

Harris, Z. S. (1954). Distributional structure. *Word 10*, 146–162. Reprinted in J. Fodor and J. Katz, *The Structure of Language*, Prentice Hall, 1964 and in Z. S. Harris, *Papers in Structural and Transformational Linguistics*, Reidel, 1970, 775–794.

Hellrich, J. and Hahn, U. (2016). Bad company—Neighborhoods in neural embedding spaces considered harmful. *COLING*.

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics 41*(4), 665–695.

Hjelmslev, L. (1969). *Prologomena to a Theory of Language*. University of Wisconsin Press. Translated by Francis J. Whitfield; original Danish edition 1943.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *SIGIR-99*.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. *ACL*.

Jia, S., Meng, T., Zhao, J., and Chang, K.-W. (2020). Mitigating gender bias amplification in distribution by posterior regularization. *ACL*.

Jones, M. P. and Martin, J. H. (1997). Contextual spelling correction using latent semantic analysis. *ANLP*.

Joos, M. (1950). Description of language design. *JASA 22*, 701–708.

Jurafsky, D. (2014). *The Language of Food*. W. W. Norton, New York.

Jurgens, D., Mohammad, S. M., Turney, P., and Holyoak, K. (2012). SemEval-2012 task 2: Measuring degrees of relational similarity. **SEM 2012*.

Katz, J. J. and Fodor, J. A. (1963). The structure of a semantic theory. *Language 39*, 170–210.

Kiela, D. and Clark, S. (2014). A systematic study of semantic vector space model parameters. *EACL 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*.

Kim, E. (2019). Optimize computational efficiency of skip-gram with negative sampling. https://aegis4048.github.io/optimize_computational_efficiency_of_skip-gram_with_negative_sampling.

Lake, B. M. and Murphy, G. L. (2020). Word meaning in minds and machines..

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*, 211–240.

Landauer, T. K., Laham, D., Rehder, B., and Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. *COGSCI*.

Lapesa, G. and Evert, S. (2014). A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *TACL 2*, 531–545.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature 401*(6755), 788–791.

Levy, O. and Goldberg, Y. (2014a). Dependency-based word embeddings. *ACL*.

Levy, O. and Goldberg, Y. (2014b). Linguistic regularities in sparse and explicit word representations. *CoNLL*.

Levy, O. and Goldberg, Y. (2014c). Neural word embedding as implicit matrix factorization. *NeurIPS*.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *TACL 3*, 211–225.

Li, J., Chen, X., Hovy, E. H., and Jurafsky, D. (2015). Visualizing and understanding neural models in NLP. *NAACL HLT*.

Lin, Y., Michel, J.-B., Lieberman Aiden, E., Orwant, J., Brockman, W., and Petrov, S. (2012). Syntactic annotations for the google books ngram corpus. *ACL*.

Linzen, T. (2016). Issues in evaluating semantic spaces using word analogies. *1st Workshop on Evaluating Vector-Space Representations for NLP*.

Luhn, H. P. (1957). A statistical approach to the mechanized encoding and searching of literary information. *IBM Journal of Research and Development 1*(4), 309–317.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). Efficient estimation of word representations in vector space. *ICLR 2013*.

Mikolov, T., Kombrink, S., Burget, L., Černockỳ, J. H., and Khudanpur, S. (2011). Extensions of recurrent neural network language model. *ICASSP*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. *NeurIPS*.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. *NAACL HLT*.

Niwa, Y. and Nitta, Y. (1994). Co-occurrence vectors from corpora vs. distance vectors from dictionaries. *ACL*.

Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002a). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice 6*(1), 101.

Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002b). Math=male, me=female, therefore math≠ me. *Journal of personality and social psychology 83*(1), 44.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957). *The Measurement of Meaning*. University of Illinois Press.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. *EMNLP*.

Peterson, J. C., Chen, D., and Griffiths, T. L. (2020). Parallelograms revisited: Exploring the limitations of vector space models for simple analogies. *Cognition 205*.

Pilehvar, M. T. and Camacho-Collados, J. (2019). WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. *NAACL HLT*.

Rehder, B., Schreiner, M. E., Wolfe, M. B. W., Laham, D., Landauer, T. K., and Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes 25*(2-3), 337–354.

Rohde, D. L. T., Gonnerman, L. M., and Plaut, D. C. (2006). An improved model of semantic similarity based on lexical co-occurrence. *CACM 8*, 627–633.

Rumelhart, D. E. and Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology 5*(1), 1–28.

Salton, G. (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall.

Schone, P. and Jurafsky, D. (2000). Knowlege-free induction of morphology using latent semantic analysis. *CoNLL*.

Schone, P. and Jurafsky, D. (2001a). Is knowledge-free induction of multiword unit dictionary headwords a solved problem?. *EMNLP*.

Schone, P. and Jurafsky, D. (2001b). Knowledge-free induction of inflectional morphologies. *NAACL*.

Schütze, H. (1992). Dimensions of meaning. *Proceedings of Supercomputing '92*. IEEE Press.

Schütze, H. (1997). *Ambiguity Resolution in Language Learning – Computational and Cognitive Models*. CSLI, Stanford, CA.

Schütze, H., Hull, D. A., and Pedersen, J. (1995). A comparison of classifiers and document representations for the routing problem. *SIGIR-95*.

Schütze, H. and Pedersen, J. (1993). A vector model for syntagmatic and paradigmatic relatedness. *9th Annual Conference of the UW Centre for the New OED and Text Research*.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation 28*(1), 11–21.

Sparck Jones, K. (1986). *Synonymy and Semantic Classification*. Edinburgh University Press, Edinburgh. Republication of 1964 PhD Thesis.

Switzer, P. (1965). Vector images in document retrieval. Stevens, M. E., Giuliano, V. E., and Heilprin, L. B. (Eds.), *Statistical Association Methods For Mechanized Documentation. Symposium Proceedings. Washington, D.C., USA, March 17, 1964*. https://nvlpubs.nist.gov/nistpubs/Legacy/MP/nbsmiscellaneouspub269.pdf.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. *ACL*.

Turney, P. D. and Littman, M. L. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning 60*(1-3), 251–278.

van der Maaten, L. and Hinton, G. E. (2008). Visualizing high-dimensional data using t-sne. *JMLR 9*, 2579–2605.

Wierzbicka, A. (1992). *Semantics, Culture, and Cognition: University Human Concepts in Culture-Specific Configurations*. Oxford University Press.

Wierzbicka, A. (1996). *Semantics: Primes and Universals*. Oxford University Press.

Wittgenstein, L. (1953). *Philosophical Investigations. (Translated by Anscombe, G.E.M.)*. Blackwell.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *EMNLP*.

Zhao, J., Zhou, Y., Li, Z., Wang, W., and Chang, K.-W. (2018). Learning gender-neutral word embeddings. *EMNLP*.

Chapter 4

Mikolov et al.

Mikolov, Tomas and Chen, Kai and Corrado, Greg and Dean, Jeffrey. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781. https://arxiv.org/abs/1301.3781

Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

Greg Corrado
Google Inc., Mountain View, CA
gcorrado@google.com

Jeffrey Dean
Google Inc., Mountain View, CA
jeff@google.com

Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

1 Introduction

Many current NLP systems and techniques treat words as atomic units - there is no notion of similarity between words, as these are represented as indices in a vocabulary. This choice has several good reasons - simplicity, robustness and the observation that simple models trained on huge amounts of data outperform complex systems trained on less data. An example is the popular N-gram model used for statistical language modeling - today, it is possible to train N-grams on virtually all available data (trillions of words [3]).

However, the simple techniques are at their limits in many tasks. For example, the amount of relevant in-domain data for automatic speech recognition is limited - the performance is usually dominated by the size of high quality transcribed speech data (often just millions of words). In machine translation, the existing corpora for many languages contain only a few billions of words or less. Thus, there are situations where simple scaling up of the basic techniques will not result in any significant progress, and we have to focus on more advanced techniques.

With progress of machine learning techniques in recent years, it has become possible to train more complex models on much larger data set, and they typically outperform the simple models. Probably the most successful concept is to use distributed representations of words [10]. For example, neural network based language models significantly outperform N-gram models [1, 27, 17].

1.1 Goals of the Paper

The main goal of this paper is to introduce techniques that can be used for learning high-quality word vectors from huge data sets with billions of words, and with millions of words in the vocabulary. As far as we know, none of the previously proposed architectures has been successfully trained on more

than a few hundred of millions of words, with a modest dimensionality of the word vectors between 50 - 100.

We use recently proposed techniques for measuring the quality of the resulting vector representations, with the expectation that not only will similar words tend to be close to each other, but that words can have **multiple degrees of similarity** [20]. This has been observed earlier in the context of inflectional languages - for example, nouns can have multiple word endings, and if we search for similar words in a subspace of the original vector space, it is possible to find words that have similar endings [13, 14].

Somewhat surprisingly, it was found that similarity of word representations goes beyond simple syntactic regularities. Using a word offset technique where simple algebraic operations are performed on the word vectors, it was shown for example that *vector("King") - vector("Man") + vector("Woman")* results in a vector that is closest to the vector representation of the word *Queen* [20].

In this paper, we try to maximize accuracy of these vector operations by developing new model architectures that preserve the linear regularities among words. We design a new comprehensive test set for measuring both syntactic and semantic regularities[1], and show that many such regularities can be learned with high accuracy. Moreover, we discuss how training time and accuracy depends on the dimensionality of the word vectors and on the amount of the training data.

1.2 Previous Work

Representation of words as continuous vectors has a long history [10, 26, 8]. A very popular model architecture for estimating neural network language model (NNLM) was proposed in [1], where a feedforward neural network with a linear projection layer and a non-linear hidden layer was used to learn jointly the word vector representation and a statistical language model. This work has been followed by many others.

Another interesting architecture of NNLM was presented in [13, 14], where the word vectors are first learned using neural network with a single hidden layer. The word vectors are then used to train the NNLM. Thus, the word vectors are learned even without constructing the full NNLM. In this work, we directly extend this architecture, and focus just on the first step where the word vectors are learned using a simple model.

It was later shown that the word vectors can be used to significantly improve and simplify many NLP applications [4, 5, 29]. Estimation of the word vectors itself was performed using different model architectures and trained on various corpora [4, 29, 23, 19, 9], and some of the resulting word vectors were made available for future research and comparison[2]. However, as far as we know, these architectures were significantly more computationally expensive for training than the one proposed in [13], with the exception of certain version of log-bilinear model where diagonal weight matrices are used [23].

2 Model Architectures

Many different types of models were proposed for estimating continuous representations of words, including the well-known Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). In this paper, we focus on distributed representations of words learned by neural networks, as it was previously shown that they perform significantly better than LSA for preserving linear regularities among words [20, 31]; LDA moreover becomes computationally very expensive on large data sets.

Similar to [18], to compare different model architectures we define first the computational complexity of a model as the number of parameters that need to be accessed to fully train the model. Next, we will try to maximize the accuracy, while minimizing the computational complexity.

[1] The test set is available at `www.fit.vutbr.cz/~imikolov/rnnlm/word-test.v1.txt`

[2] `http://ronan.collobert.com/senna/`
`http://metaoptimize.com/projects/wordreprs/`
`http://www.fit.vutbr.cz/~imikolov/rnnlm/`
`http://ai.stanford.edu/~ehhuang/`

For all the following models, the training complexity is proportional to

$$O = E \times T \times Q, \tag{1}$$

where E is number of the training epochs, T is the number of the words in the training set and Q is defined further for each model architecture. Common choice is $E = 3 - 50$ and T up to one billion. All models are trained using stochastic gradient descent and backpropagation [26].

2.1 Feedforward Neural Net Language Model (NNLM)

The probabilistic feedforward neural network language model has been proposed in [1]. It consists of input, projection, hidden and output layers. At the input layer, N previous words are encoded using 1-of-V coding, where V is size of the vocabulary. The input layer is then projected to a projection layer P that has dimensionality $N \times D$, using a shared projection matrix. As only N inputs are active at any given time, composition of the projection layer is a relatively cheap operation.

The NNLM architecture becomes complex for computation between the projection and the hidden layer, as values in the projection layer are dense. For a common choice of $N = 10$, the size of the projection layer (P) might be 500 to 2000, while the hidden layer size H is typically 500 to 1000 units. Moreover, the hidden layer is used to compute probability distribution over all the words in the vocabulary, resulting in an output layer with dimensionality V. Thus, the computational complexity per each training example is

$$Q = N \times D + N \times D \times H + H \times V, \tag{2}$$

where the dominating term is $H \times V$. However, several practical solutions were proposed for avoiding it; either using hierarchical versions of the softmax [25, 23, 18], or avoiding normalized models completely by using models that are not normalized during training [4, 9]. With binary tree representations of the vocabulary, the number of output units that need to be evaluated can go down to around $log_2(V)$. Thus, most of the complexity is caused by the term $N \times D \times H$.

In our models, we use hierarchical softmax where the vocabulary is represented as a Huffman binary tree. This follows previous observations that the frequency of words works well for obtaining classes in neural net language models [16]. Huffman trees assign short binary codes to frequent words, and this further reduces the number of output units that need to be evaluated: while balanced binary tree would require $log_2(V)$ outputs to be evaluated, the Huffman tree based hierarchical softmax requires only about $log_2(Unigram\_perplexity(V))$. For example when the vocabulary size is one million words, this results in about two times speedup in evaluation. While this is not crucial speedup for neural network LMs as the computational bottleneck is in the $N \times D \times H$ term, we will later propose architectures that do not have hidden layers and thus depend heavily on the efficiency of the softmax normalization.

2.2 Recurrent Neural Net Language Model (RNNLM)

Recurrent neural network based language model has been proposed to overcome certain limitations of the feedforward NNLM, such as the need to specify the context length (the order of the model N), and because theoretically RNNs can efficiently represent more complex patterns than the shallow neural networks [15, 2]. The RNN model does not have a projection layer; only input, hidden and output layer. What is special for this type of model is the recurrent matrix that connects hidden layer to itself, using time-delayed connections. This allows the recurrent model to form some kind of short term memory, as information from the past can be represented by the hidden layer state that gets updated based on the current input and the state of the hidden layer in the previous time step.

The complexity per training example of the RNN model is

$$Q = H \times H + H \times V, \tag{3}$$

where the word representations D have the same dimensionality as the hidden layer H. Again, the term $H \times V$ can be efficiently reduced to $H \times log_2(V)$ by using hierarchical softmax. Most of the complexity then comes from $H \times H$.

2.3 Parallel Training of Neural Networks

To train models on huge data sets, we have implemented several models on top of a large-scale distributed framework called DistBelief [6], including the feedforward NNLM and the new models proposed in this paper. The framework allows us to run multiple replicas of the same model in parallel, and each replica synchronizes its gradient updates through a centralized server that keeps all the parameters. For this parallel training, we use mini-batch asynchronous gradient descent with an adaptive learning rate procedure called Adagrad [7]. Under this framework, it is common to use one hundred or more model replicas, each using many CPU cores at different machines in a data center.

3 New Log-linear Models

In this section, we propose two new model architectures for learning distributed representations of words that try to minimize computational complexity. The main observation from the previous section was that most of the complexity is caused by the non-linear hidden layer in the model. While this is what makes neural networks so attractive, we decided to explore simpler models that might not be able to represent the data as precisely as neural networks, but can possibly be trained on much more data efficiently.

The new architectures directly follow those proposed in our earlier work [13, 14], where it was found that neural network language model can be successfully trained in two steps: first, continuous word vectors are learned using simple model, and then the N-gram NNLM is trained on top of these distributed representations of words. While there has been later substantial amount of work that focuses on learning word vectors, we consider the approach proposed in [13] to be the simplest one. Note that related models have been proposed also much earlier [26, 8].

3.1 Continuous Bag-of-Words Model

The first proposed architecture is similar to the feedforward NNLM, where the non-linear hidden layer is removed and the projection layer is shared for all words (not just the projection matrix); thus, all words get projected into the same position (their vectors are averaged). We call this architecture a bag-of-words model as the order of words in the history does not influence the projection. Furthermore, we also use words from the future; we have obtained the best performance on the task introduced in the next section by building a log-linear classifier with four future and four history words at the input, where the training criterion is to correctly classify the current (middle) word. Training complexity is then

$$Q = N \times D + D \times log_2(V). \tag{4}$$

We denote this model further as CBOW, as unlike standard bag-of-words model, it uses continuous distributed representation of the context. The model architecture is shown at Figure 1. Note that the weight matrix between the input and the projection layer is shared for all word positions in the same way as in the NNLM.

3.2 Continuous Skip-gram Model

The second architecture is similar to CBOW, but instead of predicting the current word based on the context, it tries to maximize classification of a word based on another word in the same sentence. More precisely, we use each current word as an input to a log-linear classifier with continuous projection layer, and predict words within a certain range before and after the current word. We found that increasing the range improves quality of the resulting word vectors, but it also increases the computational complexity. Since the more distant words are usually less related to the current word than those close to it, we give less weight to the distant words by sampling less from those words in our training examples.

The training complexity of this architecture is proportional to

$$Q = C \times (D + D \times log_2(V)), \tag{5}$$

where C is the maximum distance of the words. Thus, if we choose $C = 5$, for each training word we will select randomly a number R in range $< 1; C >$, and then use R words from history and

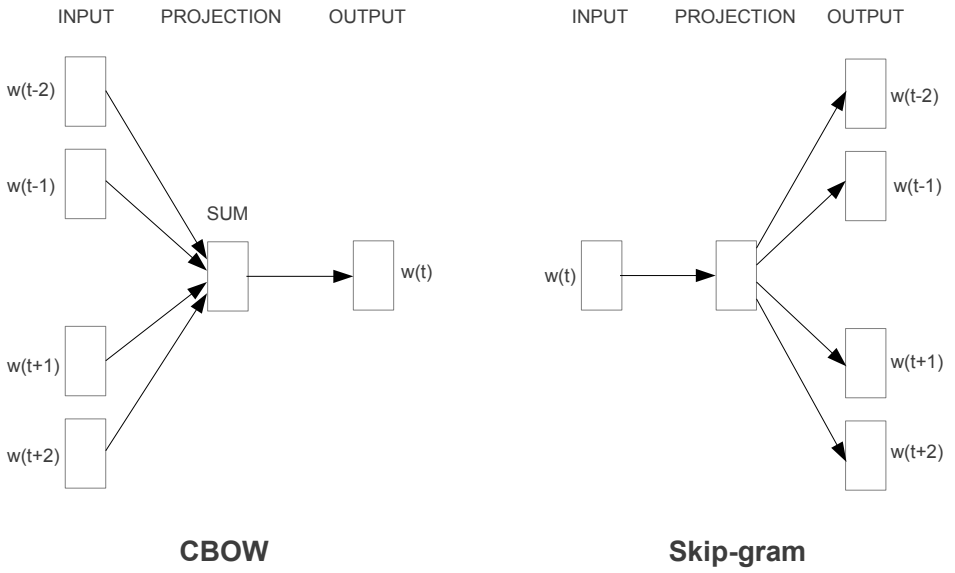


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

R words from the future of the current word as correct labels. This will require us to do $R \times 2$ word classifications, with the current word as input, and each of the $R + R$ words as output. In the following experiments, we use $C = 10$.

4 Results

To compare the quality of different versions of word vectors, previous papers typically use a table showing example words and their most similar words, and understand them intuitively. Although it is easy to show that word *France* is similar to *Italy* and perhaps some other countries, it is much more challenging when subjecting those vectors in a more complex similarity task, as follows. We follow previous observation that there can be many different types of similarities between words, for example, word *big* is similar to *bigger* in the same sense that *small* is similar to *smaller*. Example of another type of relationship can be word pairs *big* - *biggest* and *small* - *smallest* [20]. We further denote two pairs of words with the same relationship as a question, as we can ask: "What is the word that is similar to *small* in the same sense as *biggest* is similar to *big*?"

Somewhat surprisingly, these questions can be answered by performing simple algebraic operations with the vector representation of words. To find a word that is similar to *small* in the same sense as *biggest* is similar to *big*, we can simply compute vector $X = vector(\text{"}biggest\text{"}) - vector(\text{"}big\text{"}) + vector(\text{"}small\text{"})$. Then, we search in the vector space for the word closest to X measured by cosine distance, and use it as the answer to the question (we discard the input question words during this search). When the word vectors are well trained, it is possible to find the correct answer (word $smallest$) using this method.

Finally, we found that when we train high dimensional word vectors on a large amount of data, the resulting vectors can be used to answer very subtle semantic relationships between words, such as a city and the country it belongs to, e.g. France is to Paris as Germany is to Berlin. Word vectors with such semantic relationships could be used to improve many existing NLP applications, such as machine translation, information retrieval and question answering systems, and may enable other future applications yet to be invented.

Table 1: *Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship test set.*

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

4.1 Task Description

To measure quality of the word vectors, we define a comprehensive test set that contains five types of semantic questions, and nine types of syntactic questions. Two examples from each category are shown in Table 1. Overall, there are 8869 semantic and 10675 syntactic questions. The questions in each category were created in two steps: first, a list of similar word pairs was created manually. Then, a large list of questions is formed by connecting two word pairs. For example, we made a list of 68 large American cities and the states they belong to, and formed about 2.5K questions by picking two word pairs at random. We have included in our test set only single token words, thus multi-word entities are not present (such as *New York*).

We evaluate the overall accuracy for all question types, and for each question type separately (semantic, syntactic). Question is assumed to be correctly answered only if the closest word to the vector computed using the above method is exactly the same as the correct word in the question; synonyms are thus counted as mistakes. This also means that reaching 100% accuracy is likely to be impossible, as the current models do not have any input information about word morphology. However, we believe that usefulness of the word vectors for certain applications should be positively correlated with this accuracy metric. Further progress can be achieved by incorporating information about structure of words, especially for the syntactic questions.

4.2 Maximization of Accuracy

We have used a Google News corpus for training the word vectors. This corpus contains about 6B tokens. We have restricted the vocabulary size to 1 million most frequent words. Clearly, we are facing time constrained optimization problem, as it can be expected that both using more data and higher dimensional word vectors will improve the accuracy. To estimate the best choice of model architecture for obtaining as good as possible results quickly, we have first evaluated models trained on subsets of the training data, with vocabulary restricted to the most frequent 30k words. The results using the CBOW architecture with different choice of word vector dimensionality and increasing amount of the training data are shown in Table 2.

It can be seen that after some point, adding more dimensions or adding more training data provides diminishing improvements. So, we have to increase both vector dimensionality and the amount of the training data together. While this observation might seem trivial, it must be noted that it is currently popular to train word vectors on relatively large amounts of data, but with insufficient size

Table 2: *Accuracy on subset of the Semantic-Syntactic Word Relationship test set, using word vectors from the CBOW architecture with limited vocabulary. Only questions containing words from the most frequent 30k words are used.*

Dimensionality / Training words	24M	49M	98M	196M	391M	783M
50	13.4	15.7	18.6	19.1	22.5	23.2
100	19.4	23.1	27.8	28.7	33.4	32.2
300	23.2	29.2	35.3	38.6	43.7	45.9
600	24.0	30.1	36.5	40.8	46.6	50.4

Table 3: *Comparison of architectures using models trained on the same data, with 640-dimensional word vectors. The accuracies are reported on our Semantic-Syntactic Word Relationship test set, and on the syntactic relationship test set of [20]*

Model Architecture	Semantic-Syntactic Word Relationship test set		MSR Word Relatedness Test Set [20]
	Semantic Accuracy [%]	Syntactic Accuracy [%]	
RNNLM	9	36	35
NNLM	23	53	47
CBOW	24	64	61
Skip-gram	55	59	56

(such as 50 - 100). Given Equation 4, increasing amount of training data twice results in about the same increase of computational complexity as increasing vector size twice.

For the experiments reported in Tables 2 and 4, we used three training epochs with stochastic gradient descent and backpropagation. We chose starting learning rate 0.025 and decreased it linearly, so that it approaches zero at the end of the last training epoch.

4.3 Comparison of Model Architectures

First we compare different model architectures for deriving the word vectors using the same training data and using the same dimensionality of 640 of the word vectors. In the further experiments, we use full set of questions in the new Semantic-Syntactic Word Relationship test set, i.e. unrestricted to the 30k vocabulary. We also include results on a test set introduced in [20] that focuses on syntactic similarity between words[3].

The training data consists of several LDC corpora and is described in detail in [18] (320M words, 82K vocabulary). We used these data to provide a comparison to a previously trained recurrent neural network language model that took about 8 weeks to train on a single CPU. We trained a feedforward NNLM with the same number of 640 hidden units using the DistBelief parallel training [6], using a history of 8 previous words (thus, the NNLM has more parameters than the RNNLM, as the projection layer has size 640×8).

In Table 3, it can be seen that the word vectors from the RNN (as used in [20]) perform well mostly on the syntactic questions. The NNLM vectors perform significantly better than the RNN - this is not surprising, as the word vectors in the RNNLM are directly connected to a non-linear hidden layer. The CBOW architecture works better than the NNLM on the syntactic tasks, and about the same on the semantic one. Finally, the Skip-gram architecture works slightly worse on the syntactic task than the CBOW model (but still better than the NNLM), and much better on the semantic part of the test than all the other models.

Next, we evaluated our models trained using one CPU only and compared the results against publicly available word vectors. The comparison is given in Table 4. The CBOW model was trained on subset

[3]We thank Geoff Zweig for providing us the test set.

Table 4: *Comparison of publicly available word vectors on the Semantic-Syntactic Word Relationship test set, and word vectors from our models. Full vocabularies are used.*

Model	Vector Dimensionality	Training words	Accuracy [%]		
			Semantic	Syntactic	Total
Collobert-Weston NNLM	50	660M	9.3	12.3	11.0
Turian NNLM	50	37M	1.4	2.6	2.1
Turian NNLM	200	37M	1.4	2.2	1.8
Mnih NNLM	50	37M	1.8	9.1	5.8
Mnih NNLM	100	37M	3.3	13.2	8.8
Mikolov RNNLM	80	320M	4.9	18.4	12.7
Mikolov RNNLM	640	320M	8.6	36.5	24.6
Huang NNLM	50	990M	13.3	11.6	12.3
Our NNLM	20	6B	12.9	26.4	20.3
Our NNLM	50	6B	27.9	55.8	43.2
Our NNLM	100	6B	34.2	**64.5**	50.8
CBOW	300	783M	15.5	53.1	36.1
Skip-gram	300	783M	**50.0**	55.9	**53.3**

Table 5: *Comparison of models trained for three epochs on the same data and models trained for one epoch. Accuracy is reported on the full Semantic-Syntactic data set.*

Model	Vector Dimensionality	Training words	Accuracy [%]			Training time [days]
			Semantic	Syntactic	Total	
3 epoch CBOW	300	783M	15.5	53.1	36.1	1
3 epoch Skip-gram	300	783M	50.0	55.9	53.3	3
1 epoch CBOW	300	783M	13.8	49.9	33.6	0.3
1 epoch CBOW	300	1.6B	16.1	52.6	36.1	0.6
1 epoch CBOW	600	783M	15.4	53.3	36.2	0.7
1 epoch Skip-gram	300	783M	45.6	52.2	49.2	1
1 epoch Skip-gram	300	1.6B	52.2	55.1	53.8	2
1 epoch Skip-gram	600	783M	56.7	54.5	55.5	2.5

of the Google News data in about a day, while training time for the Skip-gram model was about three days.

For experiments reported further, we used just one training epoch (again, we decrease the learning rate linearly so that it approaches zero at the end of training). Training a model on twice as much data using one epoch gives comparable or better results than iterating over the same data for three epochs, as is shown in Table 5, and provides additional small speedup.

4.4 Large Scale Parallel Training of Models

As mentioned earlier, we have implemented various models in a distributed framework called DistBelief. Below we report the results of several models trained on the Google News 6B data set, with mini-batch asynchronous gradient descent and the adaptive learning rate procedure called Adagrad [7]. We used 50 to 100 model replicas during the training. The number of CPU cores is an

Table 6: *Comparison of models trained using the DistBelief distributed framework. Note that training of NNLM with 1000-dimensional vectors would take too long to complete.*

Model	Vector Dimensionality	Training words	Accuracy [%]			Training time [days x CPU cores]
			Semantic	Syntactic	Total	
NNLM	100	6B	34.2	64.5	50.8	14 x 180
CBOW	1000	6B	57.3	68.9	63.7	2 x 140
Skip-gram	1000	6B	66.1	65.1	65.6	2.5 x 125

Table 7: *Comparison and combination of models on the Microsoft Sentence Completion Challenge.*

Architecture	Accuracy [%]
4-gram [32]	39
Average LSA similarity [32]	49
Log-bilinear model [24]	54.8
RNNLMs [19]	55.4
Skip-gram	48.0
Skip-gram + RNNLMs	**58.9**

estimate since the data center machines are shared with other production tasks, and the usage can fluctuate quite a bit. Note that due to the overhead of the distributed framework, the CPU usage of the CBOW model and the Skip-gram model are much closer to each other than their single-machine implementations. The result are reported in Table 6.

4.5 Microsoft Research Sentence Completion Challenge

The Microsoft Sentence Completion Challenge has been recently introduced as a task for advancing language modeling and other NLP techniques [32]. This task consists of 1040 sentences, where one word is missing in each sentence and the goal is to select word that is the most coherent with the rest of the sentence, given a list of five reasonable choices. Performance of several techniques has been already reported on this set, including N-gram models, LSA-based model [32], log-bilinear model [24] and a combination of recurrent neural networks that currently holds the state of the art performance of 55.4% accuracy on this benchmark [19].

We have explored the performance of Skip-gram architecture on this task. First, we train the 640-dimensional model on 50M words provided in [32]. Then, we compute score of each sentence in the test set by using the unknown word at the input, and predict all surrounding words in a sentence. The final sentence score is then the sum of these individual predictions. Using the sentence scores, we choose the most likely sentence.

A short summary of some previous results together with the new results is presented in Table 7. While the Skip-gram model itself does not perform on this task better than LSA similarity, the scores from this model are complementary to scores obtained with RNNLMs, and a weighted combination leads to a new state of the art result 58.9% accuracy (59.2% on the development part of the set and 58.7% on the test part of the set).

5 Examples of the Learned Relationships

Table 8 shows words that follow various relationships. We follow the approach described above: the relationship is defined by subtracting two word vectors, and the result is added to another word. Thus for example, *Paris* - *France* + *Italy* = *Rome*. As it can be seen, accuracy is quite good, although there is clearly a lot of room for further improvements (note that using our accuracy metric that

Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

assumes exact match, the results in Table 8 would score only about 60%). We believe that word vectors trained on even larger data sets with larger dimensionality will perform significantly better, and will enable the development of new innovative applications. Another way to improve accuracy is to provide more than one example of the relationship. By using ten examples instead of one to form the relationship vector (we average the individual vectors together), we have observed improvement of accuracy of our best models by about 10% absolutely on the semantic-syntactic test.

It is also possible to apply the vector operations to solve different tasks. For example, we have observed good accuracy for selecting out-of-the-list words, by computing average vector for a list of words, and finding the most distant word vector. This is a popular type of problems in certain human intelligence tests. Clearly, there is still a lot of discoveries to be made using these techniques.

6 Conclusion

In this paper we studied the quality of vector representations of words derived by various models on a collection of syntactic and semantic language tasks. We observed that it is possible to train high quality word vectors using very simple model architectures, compared to the popular neural network models (both feedforward and recurrent). Because of the much lower computational complexity, it is possible to compute very accurate high dimensional word vectors from a much larger data set. Using the DistBelief distributed framework, it should be possible to train the CBOW and Skip-gram models even on corpora with one trillion words, for basically unlimited size of the vocabulary. That is several orders of magnitude larger than the best previously published results for similar models.

An interesting task where the word vectors have recently been shown to significantly outperform the previous state of the art is the SemEval-2012 Task 2 [11]. The publicly available RNN vectors were used together with other techniques to achieve over 50% increase in Spearman's rank correlation over the previous best result [31]. The neural network based word vectors were previously applied to many other NLP tasks, for example sentiment analysis [12] and paraphrase detection [28]. It can be expected that these applications can benefit from the model architectures described in this paper.

Our ongoing work shows that the word vectors can be successfully applied to automatic extension of facts in Knowledge Bases, and also for verification of correctness of existing facts. Results from machine translation experiments also look very promising. In the future, it would be also interesting to compare our techniques to Latent Relational Analysis [30] and others. We believe that our comprehensive test set will help the research community to improve the existing techniques for estimating the word vectors. We also expect that high quality word vectors will become an important building block for future NLP applications.

7 Follow-Up Work

After the initial version of this paper was written, we published single-machine multi-threaded C++ code for computing the word vectors, using both the continuous bag-of-words and skip-gram architectures[4]. The training speed is significantly higher than reported earlier in this paper, i.e. it is in the order of billions of words per hour for typical hyperparameter choices. We also published more than 1.4 million vectors that represent named entities, trained on more than 100 billion words. Some of our follow-up work will be published in an upcoming NIPS 2013 paper [21].

References

[1] Y. Bengio, R. Ducharme, P. Vincent. A neural probabilistic language model. Journal of Machine Learning Research, 3:1137-1155, 2003.

[2] Y. Bengio, Y. LeCun. Scaling learning algorithms towards AI. In: Large-Scale Kernel Machines, MIT Press, 2007.

[3] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean. Large language models in machine translation. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning, 2007.

[4] R. Collobert and J. Weston. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In International Conference on Machine Learning, ICML, 2008.

[5] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research, 12:2493-2537, 2011.

[6] J. Dean, G.S. Corrado, R. Monga, K. Chen, M. Devin, Q.V. Le, M.Z. Mao, M.A. Ranzato, A. Senior, P. Tucker, K. Yang, A. Y. Ng., Large Scale Distributed Deep Networks, NIPS, 2012.

[7] J.C. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 2011.

[8] J. Elman. Finding Structure in Time. Cognitive Science, 14, 179-211, 1990.

[9] Eric H. Huang, R. Socher, C. D. Manning and Andrew Y. Ng. Improving Word Representations via Global Context and Multiple Word Prototypes. In: Proc. Association for Computational Linguistics, 2012.

[10] G.E. Hinton, J.L. McClelland, D.E. Rumelhart. Distributed representations. In: Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations, MIT Press, 1986.

[11] D.A. Jurgens, S.M. Mohammad, P.D. Turney, K.J. Holyoak. Semeval-2012 task 2: Measuring degrees of relational similarity. In: Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), 2012.

[12] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In Proceedings of ACL, 2011.

[13] T. Mikolov. Language Modeling for Speech Recognition in Czech, Masters thesis, Brno University of Technology, 2007.

[14] T. Mikolov, J. Kopecký, L. Burget, O. Glembek and J. Černocký. Neural network based language models for higly inflective languages, In: Proc. ICASSP 2009.

[15] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, S. Khudanpur. Recurrent neural network based language model, In: Proceedings of Interspeech, 2010.

[16] T. Mikolov, S. Kombrink, L. Burget, J. Černocký, S. Khudanpur. Extensions of recurrent neural network language model, In: Proceedings of ICASSP 2011.

[17] T. Mikolov, A. Deoras, S. Kombrink, L. Burget, J. Černocký. Empirical Evaluation and Combination of Advanced Language Modeling Techniques, In: Proceedings of Interspeech, 2011.

[4]The code is available at `https://code.google.com/p/word2vec/`

[18] T. Mikolov, A. Deoras, D. Povey, L. Burget, J. Černocký. Strategies for Training Large Scale Neural Network Language Models, In: Proc. Automatic Speech Recognition and Understanding, 2011.

[19] T. Mikolov. Statistical Language Models based on Neural Networks. PhD thesis, Brno University of Technology, 2012.

[20] T. Mikolov, W.T. Yih, G. Zweig. Linguistic Regularities in Continuous Space Word Representations. NAACL HLT 2013.

[21] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. Accepted to NIPS 2013.

[22] A. Mnih, G. Hinton. Three new graphical models for statistical language modelling. ICML, 2007.

[23] A. Mnih, G. Hinton. A Scalable Hierarchical Distributed Language Model. Advances in Neural Information Processing Systems 21, MIT Press, 2009.

[24] A. Mnih, Y.W. Teh. A fast and simple algorithm for training neural probabilistic language models. ICML, 2012.

[25] F. Morin, Y. Bengio. Hierarchical Probabilistic Neural Network Language Model. AISTATS, 2005.

[26] D. E. Rumelhart, G. E. Hinton, R. J. Williams. Learning internal representations by back-propagating errors. Nature, 323:533.536, 1986.

[27] H. Schwenk. Continuous space language models. Computer Speech and Language, vol. 21, 2007.

[28] R. Socher, E.H. Huang, J. Pennington, A.Y. Ng, and C.D. Manning. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In NIPS, 2011.

[29] J. Turian, L. Ratinov, Y. Bengio. Word Representations: A Simple and General Method for Semi-Supervised Learning. In: Proc. Association for Computational Linguistics, 2010.

[30] P. D. Turney. Measuring Semantic Similarity by Latent Relational Analysis. In: Proc. International Joint Conference on Artificial Intelligence, 2005.

[31] A. Zhila, W.T. Yih, C. Meek, G. Zweig, T. Mikolov. Combining Heterogeneous Models for Measuring Relational Similarity. NAACL HLT 2013.

[32] G. Zweig, C.J.C. Burges. The Microsoft Research Sentence Completion Challenge, Microsoft Research Technical Report MSR-TR-2011-129, 2011.

Chapter 5

Kozlowski et al.

Kozlowski, Austin C., Matt Taddy, and Evans, James A. (2019) "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings". American Sociological Review 84:5.

AMERICAN SOCIOLOGICAL ASSOCIATION

American Sociological Review
2019, Vol. 84(5) 905–949

DOI: 10.1177/0003122419877135
journals.sagepub.com/home/asr

SAGE

The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings

Austin C. Kozlowski,[a] Matt Taddy,[b] and James A. Evans[a,c]

Abstract

We argue word embedding models are a useful tool for the study of culture using a historical analysis of shared understandings of social class as an empirical case. Word embeddings represent semantic relations between words as relationships between vectors in a high-dimensional space, specifying a relational model of meaning consistent with contemporary theories of culture. Dimensions induced by word differences (*rich* – *poor*) in these spaces correspond to dimensions of cultural meaning, and the projection of words onto these dimensions reflects widely shared associations, which we validate with surveys. Analyzing text from millions of books published over 100 years, we show that the markers of class continuously shifted amidst the economic transformations of the twentieth century, yet the basic cultural dimensions of class remained remarkably stable. The notable exception is education, which became tightly linked to affluence independent of its association with cultivated taste.

Keywords

word embeddings, *word2vec*, culture, computational sociology, methodology, text analysis, content analysis

People classify objects along myriad axes of meaning to interpret the social world. In addition to core social categories such as gender and race, a diverse array of cultural dimensions—fair/unfair, beautiful/ugly, new/old—play key roles in patterning interactions and structuring institutions. Although dominant theories of culture posit such a multidimensional matrix of meanings, empirical investigations commonly limit their attention to one or two facets due to the analytic and methodological difficulties associated with incorporating higher dimensionality.

Social class, a central sociological construct, is itself a complex and multidimensional attribution. Stratification scholars commonly treat class as a composite of several distinct factors, including affluence, education, and occupation, as well as status and cultivated taste. Decades of social science research has produced extensive knowledge of how these various socioeconomic dimensions are materially and causally interrelated (Chan and

[a]University of Chicago
[b]Amazon
[c]Santa Fe Institute

Corresponding Author:
James A. Evans, Department of Sociology, University of Chicago, 1126 E. 59th Street, Chicago IL, 60637
Email: jevans@uchicago.edu

Goldthorpe 2007; DiMaggio 1982; Hout 2012). Yet the multiple dimensions of class are not only attributes used by analysts to articulate an individual's economic standing; they also serve as axes of cultural distinction actors deploy in daily life. People, groups, and everyday objects carry cultural associations of affluence, education, cultivation, and status, which together comprise profiles of classed meaning (Bourdieu 1984; Warner, Meeker, and Eells 1949). These profiles are evoked when individuals make decisions regarding what to purchase, where to spend the evening, how to present themselves, and who to befriend. Stratification scholars have developed strong conceptions of the material relations between the multiple dimensions of class, but understanding how the *meanings* of these dimensions relate to one another and co-evolve over time remains underspecified.

In this article, we apply an emerging computational approach—neural-network word embedding models—to analyze the cultural dimensions of social class and their evolution over the twentieth century. Word embedding algorithms input large collections of digitized text and output a high-dimensional vector-space model[1] in which each unique word is represented as a vector in the space (Mikolov, Yih, and Zweig 2013; Pennington, Socher, and Manning 2014). This means each word appearing in the analyzed documents is ascribed a set of coordinates that fix its location in a geometric space in relation to every other word. Words are positioned in this space based on their surrounding "context" words in the text, such that words sharing many contexts are positioned near one another, and words that inhabit different linguistic contexts are located farther apart. Previous work with word embeddings in computational linguistics shows that words frequently sharing contexts, and thus located nearby in the vector space, tend to share similar meanings.

We provide new evidence that the dimensions of word embedding vector space models closely correspond to meaningful "cultural dimensions," such as *rich-poor, moral-immoral*, and *masculine-feminine*. We show that a word vector's position on these dimensions reflects the word's respective cultural associations. For example, projecting occupation names on an "affluence dimension," we find that traditionally well-compensated occupations, such as banker and lawyer, are positioned at one end of the dimension, and poorly paid occupations, such as nanny and carpenter, lie at the other. This occurs because with each discursive context that "banker" shares with wealthy words like "affluent," "moneyed," and "rich," it is nudged toward the rich pole of the affluence dimension, and each time "nanny" shares a context with terms like "needy," "destitute," and "poor," it is nudged toward the poor pole.

After empirically validating word embeddings' ability to capture widely shared cultural associations, we apply this method to the question of how collective understandings of social class evolved in the United States over the course of the twentieth century. To gain new leverage on this question, we train word embedding models on text from millions of books published over the entire twentieth century digitized in the Google Ngram corpus. We then identify dimensions in these models corresponding to five cultural dimensions of class described by classical and contemporary sociological theory as well as two other cultural dimensions frequently invoked in association with class: affluence, employment, status, education, cultivation, morality, and gender.[2]

Comparing texts from each decade of the twentieth century, we discover that the cultural dimensions of class comprise a complex yet remarkably stable semantic structure. We find that affluence and status serve as cultural mediators between a cluster of education, cultivation, and morality on one hand and associations of employment and ownership on the other. This persistent and intransitive structure requires high dimensionality to represent without distortion. Furthermore, we find that the cultural markers signifying positions within this robust structure are in continual flux, with terms distinguishing high and low class shifting over the decades,

following steady patterns of cultural circulation and turnover.

MULTIDIMENSIONALITY OF CLASS

Social class, the systematic and hierarchical distinction between persons and groups in social standing, has long been recognized to operate along multiple distinct dimensions. Affluence is often treated as a core aspect of class, with income commonly serving as a proxy for socioeconomic status. This is not an arbitrary selection; money is quickly and easily convertible into many forms of capital, power, and influence, making it a particularly salient element of class (Simmel [1900] 2004).

Nevertheless, scholars long have argued that the economics of class cannot be reduced to affluence alone. Analysts in the Marxist tradition foreground socio-structural position and relation to capital as the basis of social class instead (Gramsci 1992; Marx [1867] 2004; Wright 1979). From this perspective, it is not the accumulation of wealth, but rather one's position as an owner or worker, that determines a shared interest with respect to politics, culture, and social life (Marx and Engels 1970). In addition to occupational position and wealth, social scientists frequently include education as a third element of socioeconomic status. Education became particularly central to the study of class after World War II, when the expansion of mass schooling and the demands of a changing labor market turned education into a critical axis of social division (Fischer and Hout 2006).

Theorists have also noted that a full conception of social class requires accounting for its symbolic manifestations. In an early articulation of this distinction, Weber (1978) contrasted economic class with status (*Stand*), which operates via social honor and prestige. Because status refers to actors' ability to make a credible claim of esteem rather than their power in a market, it need not always coincide with affluence (Chan and Goldthorpe 2007). Recent research confirms the empirical relevance of this theoretical distinction, finding that status shapes associational networks independently of economic factors, and individuals commonly distinguish prestige from earnings in their subjective evaluations of occupational social standing (Chan and Goldthorpe 2004; Freeland and Hoey 2018).

Another line of research establishes how cultivated tastes serve as a crucial marker of class distinct from individual or collective status. Veblen ([1899] 1912) and Elias (1978) articulated this connection between cultivation and class early in the twentieth century, and Bourdieu (1984) recentered this association at century's end with the concept of cultural capital. Numerous studies show how actors parlay cultural capital into economic gains (DiMaggio and Mohr 1985), but Bourdieu's (1984) original conception draws a more complex connection between cultivated taste and affluence, with cultural elites such as artists and intellectuals comprising their own high-status social groups that stand in opposition to the economic elite.

The cultural associations of class are entwined with many diverse dimensions of social classification. For example, a growing literature on valuation and moralized markets outlines how socioeconomic attributions are shaped by moral classifications (Fourcade and Healy 2007; Zelizer 1979). This scholarship details how moral distinctions become mapped onto socioeconomic positions (Svallfors 2006) and how moral sentiments shape economic valuation (Fourcade 2011). In this vein, Lamont (1992, 2000) illustrates how middle- and working-class Americans deploy moral and socioeconomic distinctions in tandem when forming judgments about their neighbors, their friends, and themselves. Classed associations similarly interact with understandings of gender. Feminist scholars have shown how gender permeates class in the labor process (Hochschild 2012; Salzinger 2003), consumption patterns (Cohen 2003; Illouz 1997; Mears 2010), and the macro system of economic stratification (Cha and Weeden 2014; Gilman 1999; Ridgeway 2011). Arising from historical processes that

differentially distribute power and prestige by gender, classed meanings are frequently also gendered meanings (Veblen [1899] 1912).

Together, contemporary and classical work paint class as a complex construct with many facets at once connected yet analytically and culturally distinct. The precise ways these cultural dimensions of class relate to one another, however, and how these interrelations have evolved over time, remain open empirical questions.

Social Class in the Twentieth Century

The twentieth century was a period of dramatic class transformation in the United States and beyond. Large organizations came to dominate the Western world's industrial and economic landscape, mass education heightened the importance of formal credentials for occupational attainment, and the gender composition of the workforce shifted radically as women entered historically male jobs and the incidence of divorce spiked (Collins 1979; Fischer and Hout 2006). Nevertheless, it is unclear whether the system of class-based meanings used by lay actors underwent parallel transformations. Despite voluminous scholarship focused on how shared understandings of class operate on the micro-level in particular times and places (e.g., Bourgois 2003; Khan 2010; Willis 1977), macro-historical analyses of the dimensions of meaning undergirding class remain rare.

Commentators offer competing narratives about the cultural trajectory of class in the twentieth century. Some characterize the twentieth century as the eclipse of social-structural positions by identities and lifestyles. According to this line of inquiry, noneconomic identifiers, such as gender, race, education, and consumption patterns, form the new backbone of political organization and group solidarity (Clark 2018; Hunter 1992; Pakulski and Waters 1996). This literature coincides with the popularization of cultural capital in anglophone sociology, which stresses the rising importance of symbolic attributions in determining class (DiMaggio and Mohr 1985).

Other scholars argue against the "death of class" narrative, claiming that occupation and position in class structure continue to play key roles in determining wealth and shaping collective identity (Weeden and Grusky 2005; Wright 2000). Yet most research on the durable importance of occupational position and control of capital focuses on their relations to observable life chances and is not directly concerned with the cultural matrix of class. It remains unclear whether sociology's increasing attention to identity and lifestyle in transforming social class reflects concurrent trends in how class is understood in public discourse.

A third possibility is that symbolic factors like cultivation and status have always been central to how class is collectively understood. For instance, Accominotti, Kahn, and Storer's (2018) analysis of New York Philharmonic attendance recounts how cultivated taste developed into a currency of cultural capital among a middle-class intelligentsia in the nineteenth century. Moreover, classical accounts of status and cultivation suggest these symbolic components have been structuring class since at least the end of the Industrial Revolution (Elias 1978; Veblen [1899] 1912; Weber 1978).

These considerations suggest the possibility that collective understandings of class are founded on a durable system of meanings resilient to large-scale economic transformations. Empirical investigation into whether class's cultural components remained stable over the twentieth century has been stymied by methodological difficulties associated with macro-cultural analysis. Following a line of successful inquiry (Bearman and Stovel 2000; Franzosi 2004; Mohr, Wagner-Pacifici, and Breiger 2015), we propose formal text analysis as a promising avenue for recovering widely-shared understandings of class from historical populations no longer available for direct observation.

FORMAL TEXT ANALYSIS IN THE STUDY OF CULTURE

Cultural scholars from sociology, anthropology, and socio-linguistics have commonly

theorized that a group's language reflects its cultural system (Lévi-Strauss 1963; Whorf 1956). Following this insight, text has served as a key source of data for scholars investigating cultural categories and meaning structures. Text is particularly well-suited to historical-cultural analysis, as it is often the most semantically-rich record a group leaves behind. In sociology, analysis of text has historically been dominated by qualitative approaches, the two most common being interpretivist close-reading and systematic qualitative coding.

Interpretive text analysis, in which the researcher draws insights from a holistic deep reading, has produced great advances in sociological understandings of culture, but it suffers from clear limitations in reproducibility (Ricoeur 1981). Qualitative coding, in which the researcher selects a number of themes and systematically tracks their deployment in text (Glaser and Strauss 1967), can be more reproducible than a singular close reading, but it suffers from low inter-coder reliability when themes are complex or subtle. Because these dominant techniques are not easily replicable and rely on the analyst's intuition and finesse, the study of culture in sociology has largely remained a "virtuoso affair" (DiMaggio 1997). Furthermore, both interpretive text analysis and qualitative coding are limited by the pace of human reading, so neither are well suited for the analysis of very large corpora or entire socio-cultural domains.

Limitations of qualitative textual analysis have motivated scholars of culture in the social sciences and humanities to develop an array of formal and quantitative methods of text analysis (Evans and Aceves 2016). Two such methods that have gained popularity in recent years are semantic network analysis and topic modeling. Semantic networks are typically constructed by treating words as nodes in a network and textual co-occurrences as links (Carley 1994; Hoffman et al. 2017; Kaufer and Carley 1993; Lee and Martin 2015). Examining structural characteristics of a semantic network, such as central words or words that bridge semantic or cultural holes, can provide insight into the relationship between individual words and the overall conceptual structure undergirding a text (Corman et al. 2002; Pachucki and Breiger 2010; Vilhena et al. 2014).

Alternatively, topic modeling is a more recent approach that uses a well-formed probability model to enable inductive discovery of "topics" structuring a corpus, each learned as a sparse distribution over words that tend to co-occur in text (Blei, Ng, and Jordan 2003; Mohr and Bogdanov 2013). Topic modeling can detect polysemy by tracing words that exist in multiple topics, and heteroglossia, the multiple voices of a single text, by inducing the mixture of distinct topics across documents (Blei 2012; DiMaggio, Nag, and Blei 2013).

Both methods can generate important insights into the cultural system that produced a text, but there remain many sociologically important questions for which these methods are poorly suited. When corpora grow sufficiently large, standard semantic network analysis metrics fail to distinguish between concepts that are close or distant by considering topological information alone.[3] Topic modeling sorts words into a predetermined number of clusters, or topics, based on co-occurrence in text, and such discrete clusters do not capture continuous relationships between words.

As such, both networks and topic models are ill-suited for representing the multifarious associations and cultural valances that characterize all words in a corpus. Questions regarding how masculine or feminine, good or bad, high- or low-class a given object is within a cultural system remain difficult to answer using existing formal methods for text analysis. Furthermore, investigation into the relations between cultural dimensions, such as how closely a culture's rich/poor distinction relates to its masculine/feminine dimension, is beyond the scope of prior approaches.

WORD EMBEDDING MODELS AND COMPLEX SEMANTIC RELATIONSHIPS

Recent work in natural language processing has made great strides by representing relationships between words in a corpus not as

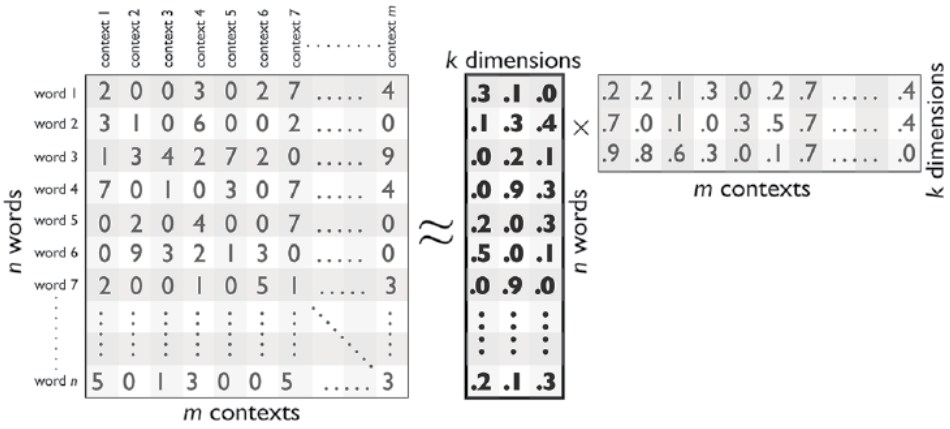


Figure 1. Schematic Illustration of the Descriptive Problem Neural Word Embeddings Solve—How to Represent All Words from a Corpus within a *k*-Dimensional Space That Best Preserves Distances between Words in Their Local Contexts

networks or topical clusters but as vectors in a dense, continuous, high-dimensional space (Joulin et al. 2016; Mikolov, Yih, et al. 2013; Pennington et al. 2014). These vector space models, known collectively as *word embeddings*, have attracted widespread interest among computer scientists and computational linguists due to their ability to capture and represent complex semantic relations.

In a word embedding model, each word is represented as a vector in shared vector space. Words sharing similar contexts within the text will be positioned nearby in the space, whereas words that appear only in distinct and disconnected contexts will be positioned farther apart. Figure 1 schematically illustrates the structure of the descriptive problem that word embeddings attempt to solve: how to represent all words from a corpus within the *k*-dimensional space that best preserves distances between *n* words across *m* semantic contexts. The solution, which we illustrate in subsequent figures, is an *n*-by-*k* matrix of values, where $k \ll m$, bolded here where $k = 3$.

An early approach to word embeddings, Latent Semantic Analysis (LSA), used singular-value decomposition (SVD) to factorize this word-context matrix when contexts were large—entire documents containing hundreds, thousands, or tens of thousands of words. The first singular value explained the most variation in the original *n*-by-*m* word-context matrix, the second component the second most, and so on, such that *k* was typically trimmed when the marginal *k*th singular value explained arbitrarily little variation in the matrix.

From efficiency considerations, SVD placed strict upper limits on the number of documents and lower limits on the size of semantic contexts they could factorize. Neural word embeddings use heuristic optimization of a neural network with at least one "hidden-layer" of *k* internal, dependent variables. This enables factorization of much larger word-context matrices constructed from vast numbers of documents containing many distinct words (large *n*) but very local word contexts (large *m*).[4]

In these models, $k \ll m$, but substantial natural language corpora require $k \geq 300$ to minimize the error of word-context matrix reconstruction (Mikolov, Yih, et al. 2013). Note that because the optimal distance between two vectors is a function of shared context rather than strict co-occurrence, words need not co-occur for their vectors to be positioned close together. If "doctor" and "lawyer" both appear near the word "work" or "office," then the vectors for "doctor" and "lawyer" would be located near each other in the embedding, even if they never appear together in text.

Distance between words in an embedding space is typically assessed using "cosine similarity," the cosine of the angle between two word vectors. This is preferred to the

Euclidean (straight-line) distance due to properties of high-dimensional spaces that violate intuitions formed in two or three dimensions. For example, as the dimensionality of a hypersphere grows, its volume shrinks relative to its surface area as more of that volume resides near the surface.[5] We normalize all word vectors (Levy, Goldberg, and Dagan 2015) such that they lie on the surface of a hypersphere of the same dimensionality as the space.

Word2vec, the most widely used word embedding algorithm and the primary approach we apply in the following analyses, uses a shallow, two-layered neural network architecture that optimizes the prediction of words based on shared context with other words.[6] Because words are located together in the embedding model if they appear in similar local contexts in the corpus, abutting words in the vector space tend to share similar meanings.

A word's nearest neighbors are often either its synonyms or syntactic variants. A word's broader neighborhood in the embedding space is typically populated by a host of terms with related meanings. Therefore, a great deal of semantic and cultural information is available simply by examining the word vectors that surround a word of interest. Kulkarni and colleagues (2015) have used word embedding models in this way to trace shifts in the meaning of the word "gay" over the course of the twentieth century, from a location in the vector space beside "cheerful" and "frolicsome" to one near "lesbian" and "bisexual." Hamilton, Leskovec, and Jurafsky (2016) similarly used word embedding models to investigate how a word's rate of semantic change, measured as change in the word's overall position in space, depends on its frequency and polysemy, finding that words occurring with high frequency change meaning more slowly and polysemous words change more rapidly.

Past work with word embedding models also shows that semantically meaningful relations can be found between words not directly proximate in the space. *Word2vec* initially attracted a great deal of attention by virtue of its intriguing ability to solve analogy problems by applying simple linear algebra to word vectors (Mikolov, Chen, et al. 2013). For example, the analogy "man is to woman as king is to _____" can be solved with a model trained on a large body of text by performing the arithmetic operation with the word vectors $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman}$, with the resulting vector most proximate to the word vector for $\overrightarrow{queen}$. *Word2vec* can achieve success rates as high as 74 percent (Ji et al. 2016) on a challenging analogy test comprising 20,000 questions involving semantic comparisons ranging from currency-country (*kwanza* is to Angola as *rial* is to Iran) and male-female (man is to woman as waiter is to waitress) to syntactic comparisons involving opposites, plural nouns, comparatives, superlatives, and verb conjugations (e.g., past tense, present participle) (Mikolov, Chen, et al. 2013). We provide a more detailed technical discussion of word embedding models in Appendix Part A.[7]

CULTURAL DIMENSIONS OF WORD EMBEDDINGS

In this article, we present a novel method for applying word embedding models to the sociological analysis of culture. We show that derived dimensions of word embedding vector spaces correspond closely to "cultural dimensions," such as affluence, gender, and status, which individuals use in everyday life to classify agents and objects in the world. By discovering and examining these culturally meaningful dimensions in a word embedding, analysts can reveal individual words' associations on those dimensions and determine how these dimensions are positioned relative to one another in that space.

For instance, an analyst can use a word embedding model to determine whether "opera" is considered more affluent than "jazz" by projecting the word vectors corresponding to "opera" and "jazz" onto the dimension of the space corresponding to affluence. Similarly, the researcher can determine if "jazz" is more masculine or feminine than "opera" by projecting these words onto

the dimension corresponding to gender in the same space. This dimensional approach emphasizes that semantic meaning is contained not only in the distance between two word vectors but also in the *direction* of that distance.

The technique we present for discovery of cultural dimensions in a word embedding vector space builds on logic for solving analogies with word embeddings. One interpretation for why $\overrightarrow{king} + \overrightarrow{woman} - \overrightarrow{man} \approx \overrightarrow{queen}$ in word embedding models is because $(\overrightarrow{woman} - \overrightarrow{man})$ closely corresponds to a "gender dimension." Adding $(\overrightarrow{woman} - \overrightarrow{man})$ to $\overrightarrow{king}$ has the effect of starting at $\overrightarrow{king}$ and taking one step on the gender dimension in the direction of femininity. Similarly, adding $(\overrightarrow{affluence} - \overrightarrow{poverty})$ to a word has the effect of taking one step in the direction of affluence. Following this intuition, we find with an embedding trained on contemporary Google News text that $\overrightarrow{hockey} + \overrightarrow{affluence} - \overrightarrow{poverty} \approx \overrightarrow{lacrosse}$. Conversely, $(\overrightarrow{poverty} - \overrightarrow{affluence})$ corresponds to one step in the direction of poverty on the same dimension.

An approximation of the affluence dimension is captured not only by $(\overrightarrow{affluence} - \overrightarrow{poverty})$, but also by any other pairs of words whose semantic difference corresponds to that cultural dimension of interest, such as $\overrightarrow{rich} - \overrightarrow{poor}$, $\overrightarrow{priceless} - \overrightarrow{worthless}$, or $\overrightarrow{prosperous} - \overrightarrow{bankrupt}$. Because we expect these similar word pairs to approximate the same cultural dimension of affluence, we calculate a single, robust affluence dimension by simply taking the arithmetic mean of a set of such pairs.[8] Other cultural dimensions, such as gender or race, can be similarly constructed with sets of antonym pairs such as $\overrightarrow{masculine} - \overrightarrow{feminine}$ or $\overrightarrow{black} - \overrightarrow{white}$, respectively.[9]

The process we propose for identifying cultural associations with word embeddings is diagrammed in Figure 2. To identify the cultural valence of a word, we calculate the orthogonal projection of the word vector onto the cultural dimension of interest. Because vectors are normalized, the projection of a word vector onto a "cultural dimension" vector is equivalent to the cosine of the angle between the two vectors. For instance, to determine the affluence association for the word "tennis," we project $\overrightarrow{tennis}$ onto the class dimension of $(\overrightarrow{affluence} - \overrightarrow{poverty}) + (\overrightarrow{rich} - \overrightarrow{poor}) + (\overrightarrow{priceless} - \overrightarrow{worthless}) + \ldots$. In this case, a more positive projection would indicate an association with affluence, and more negative values an association with poverty.[10] By comparing the projections of multiple words on a single cultural dimension, we can compare their connotations within the given spectrum of meaning.

Panel A of Figure 2 shows the construction of an affluence dimension by averaging the differences of several related antonym pairs. Panel B depicts how, by projecting the names of several sports onto the affluence dimension, we find that "boxing" and "camping" project onto the poor side of the dimension, "baseball" and "basketball" are nearly orthogonal to affluence, indicating no strong class association, and "golf," "tennis," and "volleyball" all project rich. Panel C shows how this process can be repeated for another dimension, in this example gender, and how words may be simultaneously positioned along multiple cultural dimensions. The angle between these dimensions can be calculated to capture the similarity between axes of cultural meaning, and it can be evaluated at multiple time points to trace shifts in categorical relations. Induced dimensions like affluence or gender will be approximately orthogonal if those dimensions are semantically and contextually unrelated.[11] When the angle between dimensions deviates from 90 degrees, it suggests a meaningful relationship between them, as we will demonstrate.

Our technique for identifying cultural dimensions is closely related to recent work using word embedding models to detect bias[12] in texts. Caliskan, Bryson, and Narayanan (2017) show that a word's position relative to gendered or racialized labels in a word embedding model is strongly associated with that word's associations measured by Implicit Association Tests (IAT) capturing unconscious bias (Greenwald, McGhee, and Schwartz 1998). They use this evidence to argue that word embedding models reveal

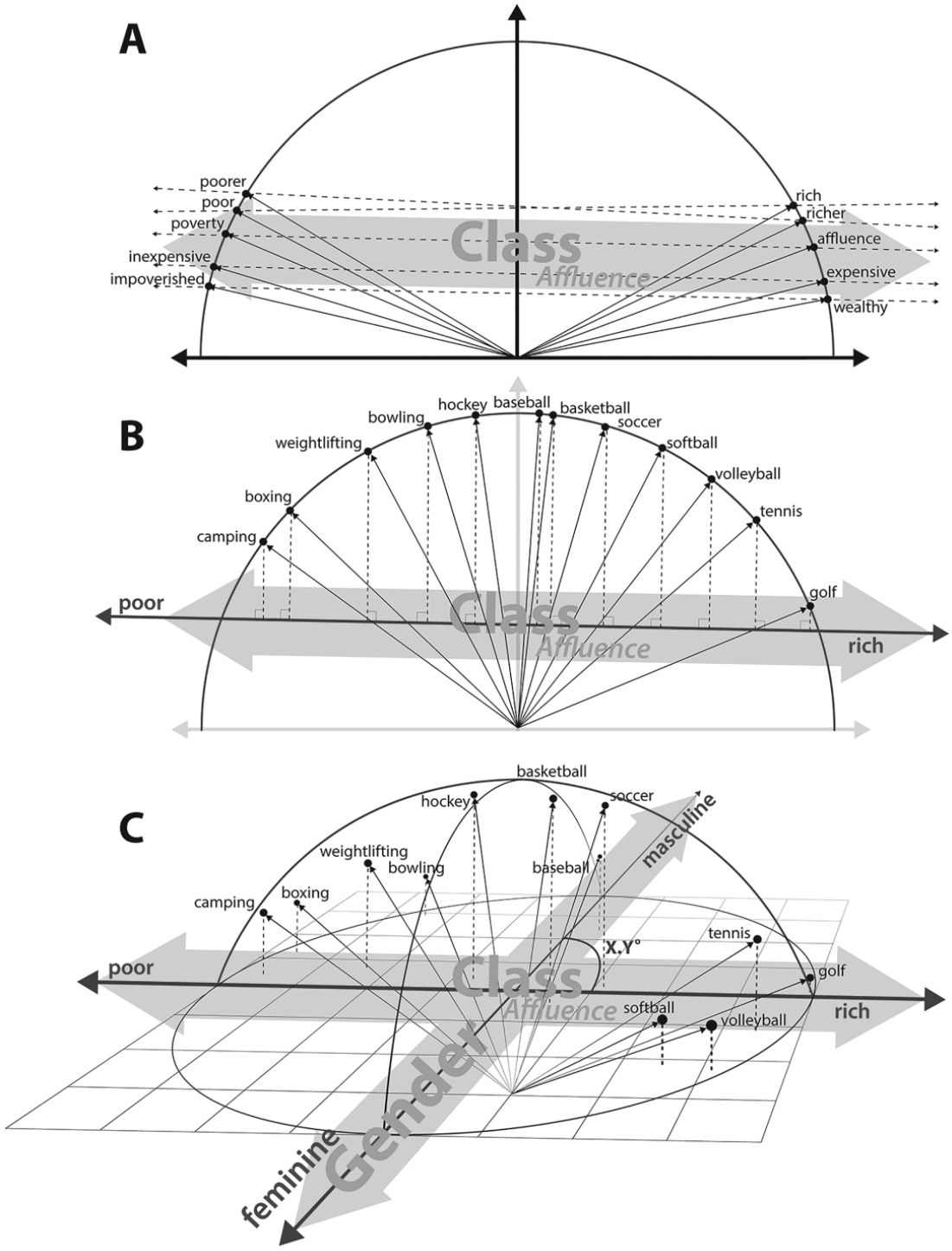


Figure 2. Conceptual Diagram of (A) the Construction of a Cultural Dimension; (B) the Projection of Words onto That Dimension; and (C) the Simultaneous Projection of Words onto Multiple Dimensions

negative racial and gender stereotypes implicit in texts. Bolukbasi and colleagues (2016) deploy a related approach to neutralize such biased associations in text.

Our work builds on these studies in several ways. First, we show that word position within the embedding model correlates not only with the unconscious associations but also with widely shared, conscious associations measured by surveys. Second, we argue that the method presented here detects not only hidden biases but a vast array of cultural valances. Many associations we find here are indeed biased: "criminal" is consistently

found to be more "poor" than "rich," and "scientist" more "masculine" than "feminine." Word embeddings include harmful stereotypes, however, only because they accurately reflect cultural systems that are themselves rife with such stereotypes. Thus, it is rarely in cultural analysts' interest to "debias" a word embedding model as Bolukbasi and colleagues (2016) propose. Rather, it is by interrogating these biases, as well as the neutral cultural associations present in the models, that analysts can cultivate an understanding of the multifaceted word meanings and cultural categories deployed in text.

Garg and colleagues (2018) begin to move in this direction by using word embeddings to study change in gender and ethnic stereotypes over time. By examining change in stereotypes, they recognize them as more than simply distortions of the semantic system, but rather meaningful characteristics that reflect the culture in which texts were produced. Their analysis, however, like Bolukbasi and colleagues (2016) and Caliskan and colleagues (2017), remains couched in the analysis of bias rather than cultural categories in general. Our approach builds on these studies by interpreting the dimensions of embedding models as representative of meaningful cultural categories rather than simply biases, distortions, or deficits in the semantic system. We then use these dimensions as tools to illuminate complex cultural relations associated with class in a given social context, across contexts, and over time. More broadly, this article is the first to specifically demonstrate the utility of word embedding models for sociological and cultural inquiry.

WORD EMBEDDINGS AND CULTURAL THEORY

Word embedding models at once align and contend with dominant theories of culture in a number of significant ways. First, word embedding models are fundamentally *relational* in how they represent meaning. "Posh" only has meaning in that it is positioned near "wealth" but closer to "style," near "fashion" but closer to "rich," and distant from "plain" and "cheap." At the same time, "wealth," "style," "fashion," "plain," and "cheap" themselves achieve meaning through their position relative to "posh" and other words in the space.

This purely relational approach to modeling meaning parallels a diverse body of cultural theorizing, including the structuralist models of meaning developed by Saussure (1916), which posit that individual signifiers are arbitrary and acquire meaning only through placement in a complex system of signification. The fundamental insight that meaning is not immanent within words and phrases but rather coheres within a broader cultural system is inherent in any word embedding analysis. This theoretical congruence makes word embeddings an effective tool for advancing empirical research within relational frameworks popular among contemporary theorists of culture (DiMaggio 2011; Emirbayer 1997; Mische 2011).

Meaning and Dimensionality

Dominant theories of culture often conceptualize meaning in terms of semantic dimensions. Considering objects' multiple, cross-cutting valences along dimensions such as good/bad, rich/poor, and masculine/feminine not only resonates with structuralist thought (Douglas 1966; Lévi-Strauss 1963), but it is central to contemporary intersectionality, affect control, and field theories. Inducing labeled cultural dimensions from word embeddings thus makes it possible to operationalize and engage with these prominent theoretical traditions using large-scale text. Distances between terms can also be fruitfully analyzed without imposing labeled semantic dimensions onto the space. Word embeddings may therefore be applied to "non-dimensional" theories of meaning, such as those based on cognitive prototypes or family resemblances (Rosch and Mervis 1975; Tversky and Gati 1978).

The ability of word embedding models to simultaneously locate objects on multiple cultural dimensions, including race, gender, class, and many others, makes them a powerful tool for studies of intersectionality. The fundamental insight of the intersectionality literature is

that cultural categories, particularly those of identity, cannot be isolated and understood independently (Crenshaw 1991; McCall 2005). Rather, analysts must always consider ways in which the meanings of cultural categories change as they overlap and intersect one another. Interrogation of the intersection of cultural categories becomes empirically tractable through word embedding models.

For example, comparing words that project high on both affluence and masculinity to those that project high on affluence and femininity will reveal how markers of class differ across gender lines within the cultural world in which the texts were produced. The theory that identity is defined by numerous cross-cutting and overlapping categories is itself predicated on a "high-dimensional" model of culture similar to that modeled by Euclidean word embeddings. Indeed, the empirical success of word embedding models to represent cultural dimensions promotes a radical view of intersectional identity, modeled not as a low-dimensional matrix, but rather a high-dimensional array composed of hundreds or thousands of interacting cultural associations.

Our use of word embeddings also shares much in common with Osgood's semantic differential method, which similarly rates words along cultural dimensions. In the semantic differential method, respondents are asked in an interview to place words on culturally meaningful spectra: for example, "Is 'dictator' closer to 'smooth' or 'rough?'" (Osgood, Suci, and Tannenbaum 1957). A key finding from this method is that much of the variance across all dimensions tested can be explained by just three core factors: evaluation (good versus bad), potency (powerful versus weak), and activity (lively versus torpid). Osgood's insight matured within sociology into Heise's (1979, 1987) affect control theory, which posits that individuals interpret events and plot courses of action by accounting for culturally based affective meanings, operationalized as the position of words on evaluation, potency, and activity (EPA) dimensions (see also Schröder, Hoey, and Rogers 2016).

Osgood and colleagues' (1957) work has at times been used to argue for the low dimensionality of meaning systems, but this interpretation overlooks key findings of the semantic differential research program. When Osgood and colleagues (1957) had respondents rate words on a set of semantic dimensions purposely selected to be unrelated in meaning, they found the EPA dimensions captured a relatively small portion of the total variance. Motivated by such results, Osgood (1969) concluded that the semantic differential only effectively captures the "affective" components of objects' meanings while systematically missing more denotative elements. The recent discovery that word embedding models require upward of 200 dimensions to successfully recover complex semantic relationships suggests that although three dimensions may be able to coarsely bin concepts and predict approximate human responses, higher dimensionality enables fine-grained classification along a rich set of distinctions particularly useful for sociologists, who are often concerned with subtle nuances of meaning between specific dimensions, such as gender, status, and education.

Word embedding models also operationalize and extend key elements of Bourdieu's field theory. Bourdieusian cultural fields offer a model of how individuals, objects, and positions in social structure are located relative to one another in structurally homologous "social spaces," with relations between entities described in terms of "distances" (Bourdieu 1989). Bourdieu (1984) frequently represented these social spaces geometrically using the method of correspondence analysis (Greenacre 2017), rendering distances between entities and meaningful dimensions of the field visible by placing them in a two-dimensional plane. By overlaying the space of economic relations with the homologous space of cultural relations, Bourdieu underscores how social class operates at once materially and symbolically.

The vector-space models produced by word embeddings similarly position objects relative to one another in a shared space based on cultural similarity. By leveraging the wealth of information contained in a large corpus, however, word embeddings are able

to position words in a semantically-rich, high-dimensional space that need not be reduced to low dimensionality for interpretation. Indeed, the low-dimensional projection of correspondence analysis operationalizes a theory of cultural capital that is itself low-dimensional: social actors struggle to obtain and maintain dominant positions within a cultural field through a single currency of cultural capital and a single dimension of status-distinguishing tastes and preferences (Bourdieu 1984).[13] In this vein, Lamont (1992) criticizes Bourdieu's approach for overemphasizing distinctions based on aesthetic cultivation such as common/rare while neglecting moral distinctions such as honest/dishonest or fair/unfair.

By preserving higher dimensionality in a cultural space, word embeddings can facilitate the development and testing of high-dimensional theories for how actors acquire and exploit varied cultural capitals along multiple dimensions of distinction. Moreover, identifying cultural dimensions using antonym pairs does not require interpreting orthogonal dimensions like correspondence analysis, but instead allows analysts to examine relations between correlated but distinct semantic dimensions. The high dimensionality of word embeddings thus leaves room for complex interrelations between multiple axes of cultural distinction and opens the relationship between these axes as grounds for empirical investigation.

DATA AND METHODS

Our investigation relies on multiple data sources,[14] first for validation of our method and second for examination of historical trends in the cultural dimensions of class. To determine the ecological validity of our general approach, we compare results from word embedding models to human-rated cultural associations assessed by surveys, both contemporary and historical. Having established the validity of our method, we train word embedding models on Google Ngrams text from books published over the span of the twentieth century, and we use these models to interrogate broadly shared understandings of social class.

Surveys of Cultural Association

To establish a basis of comparison between human-reported associations and associations represented in word embedding models, we fielded a survey of cultural associations to 398 respondents on Amazon Mechanical Turk. The survey was fielded in 2016 and 2017 and was open only to Mechanical Turk users located in the United States. Although our sample cannot be said to be representative of the general U.S. population, responses to basic demographic questions indicate wide diversity in age, gender, and racial composition (Levay, Freese, and Druckman 2016). To improve representativeness, we apply post-stratification weights to the sample, weighting on race (white, black, or other), education (bachelor's degree or less), and sex (male or female). The results presented here include post-stratification weighting, but unweighted models produce substantively similar findings. This survey and the weighting procedures are detailed in Appendix Part B.

In the survey, respondents were asked to rate 59 different items on scales representing association along class, race, and gender lines. All questions followed the format, "On a scale from 0 to 100, with 0 representing *very working class* and 100 representing *very upper class*, how would you rate a *steak*?" For measuring race and gender associations, the survey posed similarly worded questions, replacing "working class" and "upper class" with "white" and "African American," or "feminine" and "masculine," respectively. A full list of items asked on the survey is available in Appendix Table B1. Words were selected in seven topical domains: occupations, foods, clothing, vehicles, music genres, sports, and first names. A diverse array of topical domains were chosen to test the capacity of word embedding models to detect cultural associations across very different subjects. Specific terms were selected within each topical domain to ensure high variance across dimensions.[15] We calculate the weighted mean of responses for each item, and we use these means as our estimates of a general cultural association. The end product

is thus a rating between 0 and 100 on a class dimension, a race dimension, and a gender dimension for each of the 59 words listed in Table B1. Measurement of broadly shared cultural associations with a Mechanical Turk survey is likely to suffer from bias and measurement error, but these weaknesses should only attenuate the correspondence between the surveyed associations and those recovered from word embedding models. Therefore, the associations presented here between survey and word embedding models can be interpreted as conservative estimates.

For historical validation, we draw on a similar dataset collected in the 1950s by semantic differential researchers. To produce a standard set of word scores for social psychologists to use across studies, Jenkins, Russell, and Suci (1958) had 30 college students rate 360 common terms on 20 semantic dimensions, such as *hard-soft* and *good-bad*, and published a table reporting the average rating for every word on each semantic dimension. We use these average scores as measures of self-reported cultural associations from the 1950s, enabling us to at once test a broader range of semantic dimensions and validate word embeddings for historical analysis. We exclude 11 terms from the analysis either because they are two-word phrases (e.g., "neurotic man") or they did not appear frequently enough in the Google Ngrams text to be rendered in the vector space (e.g., "briny"), resulting in a total of 349 words used in the analysis, each scored on 20 semantic dimensions.

Word Embedding Data

We analyze several word embedding models trained on multiple textual archives. The majority of our analyses utilize embedding models trained on publicly-available Google Ngram texts. The Google Ngram corpus, the product of a massive project in text digitization across thousands of the world's libraries, distills text from 6 percent of all books ever published (Lin et al. 2012; Michel et al. 2011). Any sequence of five words that occurs more than 40 times over the entirety of the scanned texts appears in the collection of 5-grams, along with the number of times it occurred each year. Because word embeddings require local context to determine the meaning of words, we limit our analysis to the collection of 5-grams, and we exclude data on the occurrence of 4-grams, 3-grams, 2-grams, and single words.[16] All characters were converted to lowercase in preprocessing to increase the frequency of rare words. Although the Google Ngrams corpus does not represent one single, identifiable voice, it includes a vast number of documents spanning a variety of genres, including novels, government documents, academic texts, and technical reports, making it sensitive to subtle associations that appear diffusely in general discourse. Google Ngrams are poorly suited for identifying subcultural or contextually-specific meanings, but they are able to successfully capture pervasive and widely-shared meanings that characterize terms across contexts.

The Google Ngram corpus is a uniquely powerful source of textual data, but it suffers from various weaknesses. Google Ngrams have been subject to criticism because the composition of the corpus in a given year may not be representative of total literary output (Pechenick, Danforth, and Dodds 2015). We also recognize that authors whose books and periodicals appear in Google Ngrams are by no means a culturally representative sample of the U.S. general public. Instead, we must limit our generalizations to a relatively elite, "literary public"; a group whose cultural framework of class is consequential given its wide dissemination but possibly different from more marginalized populations underrepresented in the corpus. Word embedding models require very large collections of text to reproduce accurate semantic relationships, and Google Ngrams provide the largest and most extensive sampling of historical English texts. Furthermore, our contemporary and historical validations suggest Google Ngrams over the twentieth century are able to produce cultural associations that mirror human reports on numerous diverse semantic dimensions. We therefore proceed with Google Ngrams as our primary source of historical text and reflect on limitations of our analyses in the discussion.

We train word embedding models on Google Ngrams texts for both the historical analysis of class and contemporary validations. The Google Ngrams corpus contains metadata specifying the year of publication for each string of text, making it possible to trace semantic changes over time. We divide the corpus by decade, training separate models on texts from 1900 to 1909, 1910 to 1919, and so on through 1990 to 1999, resulting in 10 independently constructed word embedding models. By comparing these models side-by-side, we are able to trace macro-cultural trends over this 100-year period. Only words that appear at least 25 times are rendered in the model for a given decade, thus excluding words mentioned too rarely to be accurately placed.

For contemporary validation, we train an embedding model on Google Ngrams of publications dating from 2000 through 2012. We use this range of years because Google Ngrams do not include publications more recent than 2012, and this duration is similar to those used in our historical analyses. For additional validation, we compare the performance of the Google Ngrams embedding to two widely used, pre-trained embeddings: one trained on contemporary Google News text with *word-2vec* and one trained on a broad scraping of website text from the Common Crawl with *GloVe*. These alternative embeddings are discussed in greater detail in Appendix Part A.

For validation with the 1950s semantic differential survey data, we use the same embedding model trained on 1950 to 1959 Google Ngrams that we use in our historical analysis. We train all word embeddings with *word2vec* skipgram architecture with 300 dimensions, following standards that prior research found to be effective in solving analogy tasks (Mikolov, Chen, et al. 2013). We also test the validity of our approach across different corpora and word embedding algorithms, including large samples of twenty-first-century news and webpages, which we detail in Appendix Part A.

We identify a diverse set of cultural dimensions in our embedding models for validation and for historical analysis. For contemporary validation, we construct cultural dimensions corresponding to three core sociological axes of classification: affluence, gender, and race (black/white). For historical validation, we construct 20 cultural dimensions corresponding to those measured by Jenkins and colleagues (1958). Finally, for our historical analysis of collective understandings of class, we construct cultural dimensions corresponding to those identified in the literature as being constitutive of, or deeply intertwined with, social class. For these analyses, we again construct dimensions for affluence and gender, and we add dimensions of education, employment (owner/worker), status, cultivation, and morality.

Measuring Cultural Dimensions

To identify cultural dimensions in word embedding models, we average numerous pairs of antonym words. Cultural dimensions are calculated by simply taking the mean of all word pair differences that approximate a given dimension, $\frac{\sum_{p}^{|P|} \overrightarrow{p_1} - \overrightarrow{p_2}}{|P|}$, where p are all antonym word pairs in relevant set P, and $\overrightarrow{p_1}$ and $\overrightarrow{p_2}$ are the first and second word vectors of each pair.[17] The projection of a normalized word vector onto a cultural dimension is calculated with cosine similarity, as is the angle between cultural dimensions.

We bound our estimates with 90 percent confidence intervals constructed through a nonparametric subsampling approach. This method involves splitting the corpus into 20 non-overlapping subsamples, independently constructing embedding models on these 20 subcorpora, and calculating the desired estimates on all 20 embedding models. The variance between these estimates is then used to quantify how sensitive the estimates are to particular usages in the text. If a word is used infrequently and appears in several very different contexts, it will produce a wider error bound than a word used frequently in consistent contexts. Technical details regarding our calculation of these confidence intervals is available in Appendix Part C.

To assemble effective lists of antonym terms, we used five thesauri: three contemporary (*Bartlett's Roget's Thesaurus* 1996; *Oxford Thesaurus* 1992; *Webster's Collegiate Thesaurus* 1976) and two historical (Roget 1912; Smith 1903). Drawing words from historical thesauri ensures our list of terms is robust for the early and more recent decades of the twentieth century. Indeed, certain terms only appear in the early decades of the century (e.g., "luxuriant" and "penurious") and others only appear at the end (e.g., "privileged" and "underprivileged"). Antonym pairs that do not appear in a given decade's embedding are excluded from calculation of the average cultural dimension. As a result, the terms that comprise a cultural dimension shift as the terms used in discourse to designate the cultural dimension themselves shift.

Some cultural dimensions are characterized by a much larger set of words in the English language than others, leading to substantial differences in the number of antonym pairs included for each. Furthermore, selection of antonym pairs requires some discretion on the part of the analyst, because thesauri often contain a wide range of loosely synonymous terms inappropriate for the given analysis. We present supplemental analyses suggesting that cultural dimensions constructed from fewer antonym pairs may be less robust, but results do not differ substantially between those constructed from 10 pairs and those trained on 40. We further find that the exact ways words are paired (e.g., $\overrightarrow{rich} - \overrightarrow{poor}$ instead of $\overrightarrow{rich} - \overrightarrow{impoverished}$) has a minimal effect on the effectiveness of the dimension in predicting human-rated associations. The full sets of antonym pairs we use for all cultural dimensions analyzed in this study are listed in Appendix Part D, and robustness checks are presented in Part E. Corpus sizes are listed in Appendix Part F.

We contextualize our cultural analysis of class by comparing associations held in the general public to those expressed in sociological literature. To produce clear grounds for formal comparison, we compute word embedding models trained on a corpus of all sociology articles published in the twentieth century in the JSTOR collection. The class-based associations we find in this corpus generally accord with widely recognized disciplinary trends (see Appendix Part H).

RESULTS

Validation of Cultural Dimensions

We validate the ability of word embedding models to reflect widely shared cultural associations by calculating the Pearson's correlation between a word's mean rating on a given survey scale and the word's projection on the corresponding cultural dimension in an embedding model. Correlations are calculated using the 59 terms listed in Appendix Table B1. We compare the validation results from the Google Ngrams embedding to two widely-used, pre-trained embedding models to illuminate the strengths and weaknesses of Google Ngrams compared with other corpora. Results are presented in Table 1.

The first column of Table 1 presents correlations between survey responses and word vector projections for class. We see that association for the Google Ngrams embedding is .53, and correlations with the two alternative embeddings are .57 and .58 (details in Appendix Part A). The second column displays the correlation between gendered associations in survey response and projection on the embedding's gender dimension. For gender associations, the Google Ngrams embedding correlates with surveyed ratings at .76, and alternative embeddings correlate at .88 and .90. These correlations attest to how well a gender dimension elicited from the word embedding model corresponds to contemporary individuals' understandings of masculinity and femininity. The third column shows correlations between word embedding projections and survey ratings for racial associations. The Google Ngrams corpus does relatively poorly in this test, correlating at only .27 with survey response. Other embeddings range widely from .42 to .75.

There are many possible explanations for the Google Ngrams' relatively poor performance in picking up racial associations. The

Table 1. Pearson Correlations between Survey Estimates and Word Embedding Estimates for Gender, Class, and Race Associations

	Class (Affluence)	Gender	Race
Google Ngrams *word2vec* Embedding†	.53	.76	.27
Google News *word2vec* Embedding	.58	.88	.75
Common Crawl *GloVe* Embedding	.57	.90	.44

Note: $N = 59$, except †$N = 58$ where one word measured in the survey did not occur frequently enough in the text to appear in the word embedding.

subject matter of news articles and general internet postings may be imbued with more racial associations than the Ngrams corpus, which contains significant non-fiction, including technical reports and scientific publications without narrative content that could invoke ambient, contemporary racial associations within that embedding model's projections. Additionally, as noted earlier, the Google Ngrams text were reduced to lowercase in preprocessing, which decreased the available number of antonym word pairs for constructing the race dimension from seven to five, possibly resulting in decreased accuracy of the dimension. This poses pronounced difficulties for analyses of race, given that the semantic dimension *black-white* will likely capture a host of associations related to color but unrelated to race. Because of these difficulties in recovering racial associations from the Ngram corpus, we refrain from analyses of race in our subsequent analyses of class associations over time.

Figure 3 plots the correspondence between word embedding models and our survey of cultural associations. The figure reveals how several music genres—jazz, rap, opera, punk, techno, hip hop, and bluegrass—are arrayed on the cultural dimensions of class and race by survey response and the word embedding trained on Google News, with the average survey rating of a word depicted in black and the projection in gray. Comparing survey ratings to word embedding projections, we see striking similarity in the relative positions of words. In both methods, opera holds the association of being both high class and white. Techno, punk, and bluegrass are similarly white but of distinctly lower class than opera. On the right end of the panel, jazz is associated with both African Americans and high class, whereas hip hop and rap tend toward the working class. Projecting words simultaneously into multiple dimensions, it is clear how word embeddings can be used to examine intersectionality by revealing how class markers vary across racial lines.

We next validate results from an embedding trained on 1950s Google Ngrams text on data from a semantic differential survey fielded in 1958 (Jenkins et al. 1958). This validation assesses the ability of Google Ngrams embeddings to capture historical associations and their capacity to reflect a wide variety of semantic dimensions beyond core sociological categories. We take the same set of 349 words and 20 cultural dimensions measured by Jenkins and colleagues and produce a corresponding embedding-derived dataset by projecting the respective word vectors onto corresponding cultural dimensions from the embedding model. The sets of antonym pairs used to construct these cultural dimensions in the embedding are listed in Appendix Table D2.

Figure 4 depicts Pearson correlations between word embedding projections and human ratings for 20 semantic dimensions. We find a statistically significant ($p < .01$), positive association between human-rated associations and embedding projection on all dimensions. Many correlations are impressively high; correlations on six dimensions exceed .60, including *kind-cruel, good-bad, beautiful-ugly*, and *true-false*. We see more modest correlations on other dimensions, but we also find that lower correlations generally correspond to lower variance in average human ratings on those dimensions. This

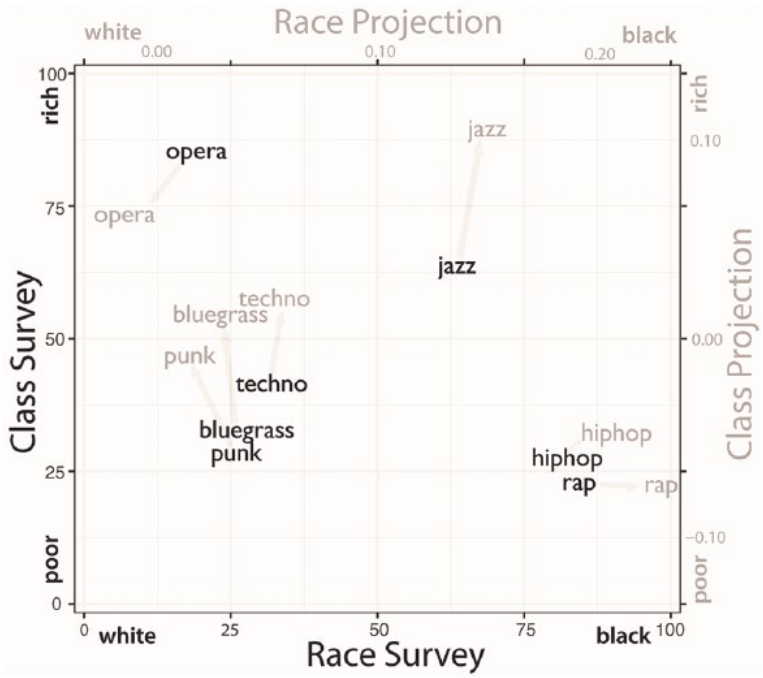


Figure 3. Projection of Music Genres onto Race and Class Dimensions of the Google News Word Embedding (Gray) and Average Survey Ratings for Race and Class Associations (Black)

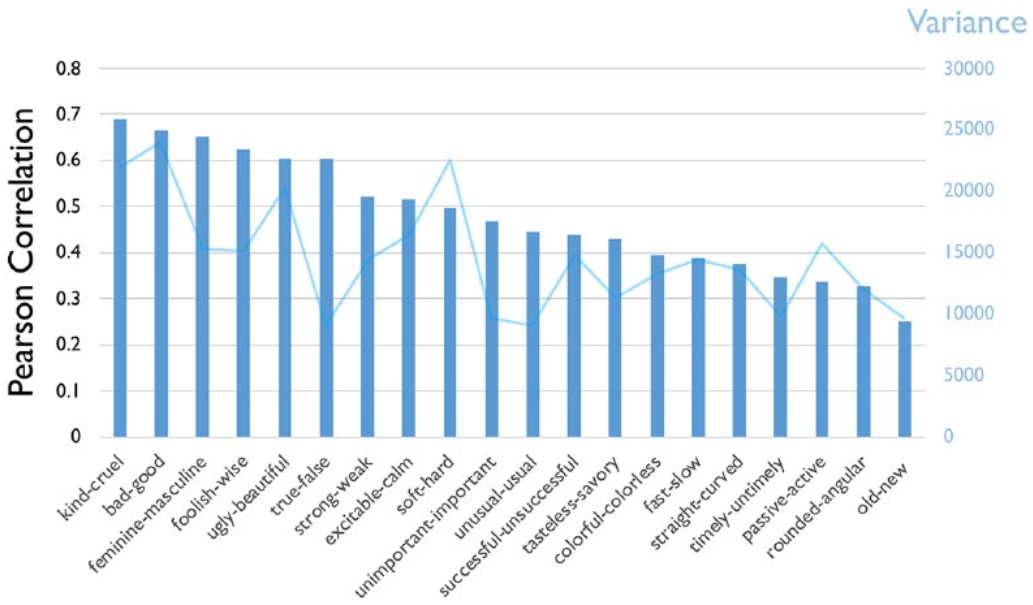


Figure 4. Correlations between Word Embedding Projections and Human-Rated Associations on 20 Semantic Dimensions, Alongside Variance of Average Human-Ratings on Those Dimensions; 1950 to 1959 Google Ngrams Corpus

means dimensions with many strongly-rated words on both ends of the spectrum are more successfully captured by word embedding models. For example, subjects tended to rate most words near the middle on the *rounded-angular* dimension, suggesting they do not register strong associations. Unsurprisingly, these subtle and potentially more noisy associations are more difficult to capture from text. We engage semantic differential theory more deeply with supplemental analyses in Appendix Part G, showing that subspaces of word embeddings can reproduce the dimension reduction typical of semantic differential

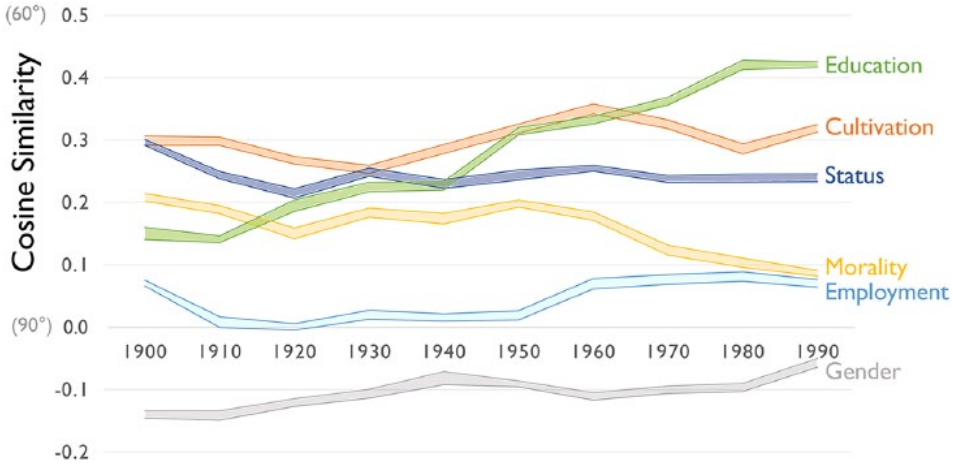


Figure 5. Cosine Similarity between the Affluence Dimension and Six Other Cultural Dimensions of Class by Decade; 1900 to 1999 Google Ngrams Corpus
Note: Bands represent 90 percent bootstrapped confidence intervals produced by subsampling.

analysis, but the full spaces cannot be represented in lower dimensionality without considerable loss of information.

Meanings of Class across the Twentieth Century

Having validated word embeddings' capacity to capture meaning along many semantic dimensions, we apply this method to unpack multiple dimensions of class and explore their interrelation in the United States over the twentieth century. Specifically, we seek to discover how shared understandings of social class evolved during a period of dramatic economic transformation and which class components remained stable in spite of these developments. We analyze five dimensions of class identified prominently in sociological theory: affluence, employment (owner/worker), status, education, and cultivation.[18] For additional comparison, we also construct dimensions for two categories that theorists have noted as deeply intertwined with class: morality (Lamont 1992; Lerner and Miller 1978; Skeggs 1997) and gender (Reay 1998; Veblen [1899] 1912).

First, we focus on the cultural dimension of affluence, the ubiquitous class marker that anchors modern understandings of socioeconomic inequality (Piketty 2014). We begin by investigating how affluence has changed its relations to the other components of class. To accomplish this, we calculate the angle between each class dimension and the other six dimensions of interest, and then we explore how these angles shift over the course of the twentieth century.[19] Figure 5 displays the angle, measured in cosine similarity, between the affluence dimension and each of our other six cultural dimensions: employment, status, cultivation, education, morality, and gender. We observe general stability in the relations between affluence and other cultural dimensions, with a few key exceptions. Interestingly, the dimensions most parallel to affluence at the start of the twentieth century are cultivation and status. These are closely followed by morality, gender, and education, respectively. Affluence notably manifests the most modest association with employment position.

It is illuminating to consider places where popular cultural associations run counter to understandings of class expressed within sociology. For example, gender's association with affluence is weakly negative within general discourse, implying an association between affluence and femininity. This finding runs contrary to the sociological expectation that masculinity would be associated with affluence, given that men in the United States earn greater income and control more wealth than women. Such disjunctions between sociological and conventional

understandings of class can be verified by comparing results from embeddings trained on Google Ngrams to those trained on sociological literature. We provide this empirical comparison in Appendix Part H.

The popular association of femininity with affluence in general discourse is less surprising when affluence is considered from a historical perspective. Veblen ([1899] 1912) documented how wives and daughters were frequently used as vessels for men's "vicarious consumption," and how women's distance from toil in the workplace served as a marker of class in affluent society. Similarly, Zelizer (1989) notes that women's money in the early twentieth century was commonly considered "pin money," earmarked for extravagant and indulgent purchases, whereas men's money was reserved for mundane necessities. Projections in historical Google Ngram embeddings reinforce this interpretation. Among the 10 nouns most highly projecting on the affluence dimension in the first decade of the twentieth century are "fragrance," "perfume," "jewels," and "gems," all of which project strongly feminine, suggesting that upper-class women were cultural mannequins for the display of wealth.

Employment position, either as a worker or owner, is similarly prominent in sociological understandings of wealth accumulation in the late twentieth century, yet its relationship with affluence in general discourse is weak. Across the entire century, the employment association is dwarfed by affluence's relationship with the symbolic factors of cultivation and status. Again, although these findings do not align with how sociologists conceive of social class, they accord with certain key theories of class representation. Bourdieu's concept of "misrecognition" and Marx's earlier concept of fetishism both describe how relations of production undergirding systems of economic stratification are obscured while the outward trappings of class, displayed through consumption patterns, remain visible and culturally salient. This perspective also anticipates the tight association between affluence and cultivated tastes in popular discourse throughout the twentieth century.

Most cultural dimensions of class remain remarkably stable over the century, yet we observe a striking change in the relationship between dimensions of affluence and education. Although their association is only weakly positive at the dawn of the twentieth century, it surpasses all other dimensions by the century's close, suggesting that education and affluence became increasingly synonymous. It is possible, however, that this relationship is mediated by notions of cultivation. Cultural capital scholars have long argued that education reproduces patterns of economic stratification by providing students with cultural knowledge and dispositions that exert signaling effects in the market (Collins 1979; Lamont and Lareau 1988; Lareau and Weininger 2003). In embedding terms, this would imply that words with strong, positive educational valence only have an association with affluence insofar as they also project strongly on cultivation. To determine the extent to which education's semantic connection to affluence is mediated by cultivation, we use regression to model their relationship and parse the geometry of this cultural space. OLS regression estimates the expected slope along one dimension of the vector space while holding others fixed. Given that nonindependence is inherent to word embedding models, we do not intend the quasi-experimental interpretation of regression common in sociological analysis.[20]

Figure 6 presents results from OLS regressions of cultivation and education projections predicting affluence projections. Interestingly, when adjusting for cultivation, projection on the education dimension actually exhibits a weakly negative association with affluence in the first half of the twentieth century. In other words, for two words with the same cultivation projection, the word with a greater education projection would have a *lower* expected affluence projection, suggesting education's cultural association with affluence was a byproduct of its association with cultivation, sophistication, and refinement. Indeed, at the beginning of the twentieth century, education at times implied a necessity to participate in the world of work rather than living comfortably on rentier income (Veblen [1899] 1912).

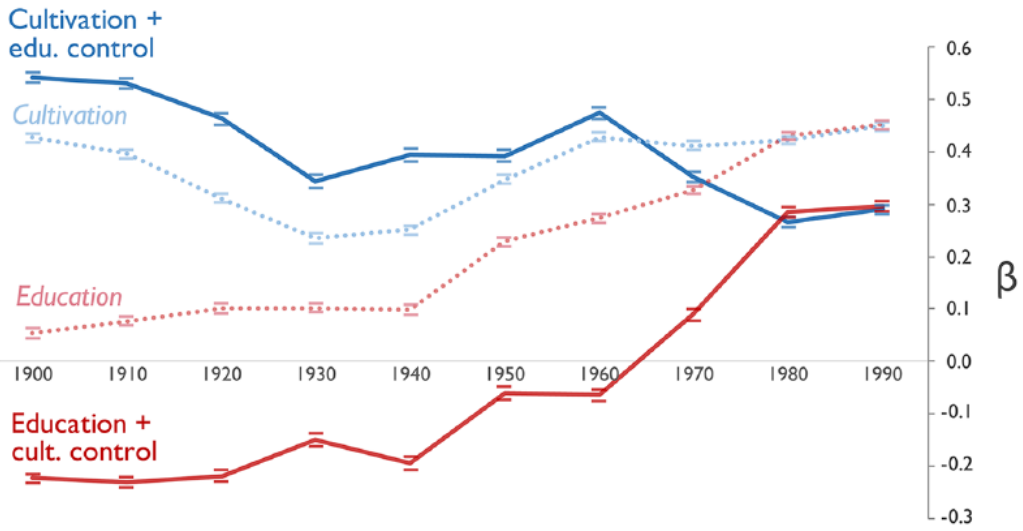


Figure 6. Standardized Coefficients from OLS Regression Models in Which Word Projections on Cultivation and Education Dimensions Predict Projection on the Affluence Dimension; 1900 to 1999 Google Ngrams Corpus
Note: A separate OLS regression model is fit for each decade; $N = 50{,}000$ most common words in each decade.

This relationship transforms over the course of the century. By the 1990s, education projections strongly associate with affluence, independent of cultivation. This finding suggests that by the end of the twentieth century, education represents a marginally distinct cultural marker of affluence, no longer redundant with cultivation. Education's cultural association with affluence was mediated by cultivation at the beginning of the twentieth century, but meanings associated with education and affluence intertwined as education became increasingly essential for socioeconomic attainment.

This finding is ironic when considered against concurrent trends in sociological theories of class. With the rise of cultural capital theory, critical scholars suggested that education influences income by bestowing forms of cultural distinction rather than by providing practical knowledge and skills (see Appendix Part H). Yet, at the moment sociologists came to see education as operating via cultivation, the opposite occurred in public perception, where education became imbued with independent connotations of affluence as its demand among elite, well-paid occupations rose (see Appendix Part I).

In Figure 7, we broaden our focus away from affluence to comprehensively view relations between the multiple dimensions of class, displaying each dimension's angles with all others. In spite of the rapid and encompassing economic transformations of the twentieth century, we find that relations between the cultural dimensions of class remain remarkably constant. Most dimensions that begin close together remain close, and those orthogonal retain their independence. The rank ordering of most angles is preserved for 100 years. Examining which cultural dimensions are correlated and which are independent, we see that cultivation, morality, and education are consistently close together, moderately related to status and affluence, and almost orthogonal to employment position. In fact, employment shows an association with morality in the opposite direction, with bosses carrying an odious cultural valence relative to workers. Despite its negative relationship with morality, however, employment shares modest but positive associations with affluence and status.

Taken together, these results demonstrate a remarkably stable and complex structure among the cultural dimensions of class, with dimensions most closely associated with social distinction—morality, cultivation, and education—clustered on one end, employment position on the other, and status and affluence mediating these otherwise unrelated domains. Observing this structure holistically

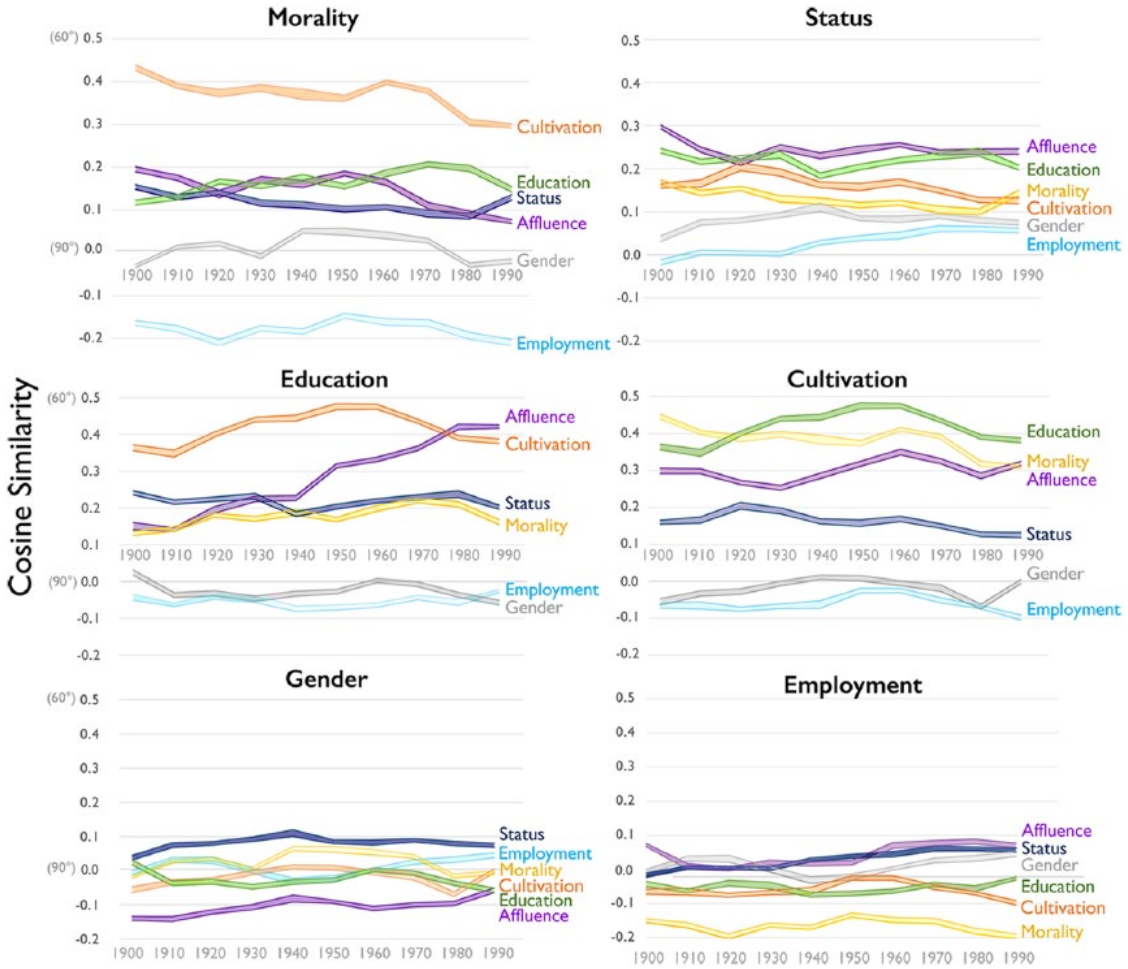


Figure 7. Cosine Similarity between Each Class Dimension and All Others by Decade; 1900 to 1999 Google Ngrams Corpus
Note: Bands represent 90 percent confidence intervals produced by subsampling.

helps clarify the cultural relationship between these dimensions, which are rarely considered simultaneously. Status and affluence are at once colored by two distinct cultural valances. On one side, they carry connotations of ownership and power. On the other, they are signaled by refinement, virtue, and edification—characteristics with little association to power and industry.

This complex semantic structure requires high dimensionality for representation. In Figure 8, conceptual diagrams illustrate that two dimensions are not enough to reproduce the angles between any three dimensions of class without significant distortion. If the relation between employment and cultivation is held at its measured value of 90°, then it is impossible to keep the angle between cultivation and status at 85.5° while also maintaining that between employment and status at 79.3°. Thus, even when considering cultural categories closely related to class, high dimensionality is necessary to preserve crucial distinctions between meanings.

Finally, we turn from relations between class dimensions to focus on the stability of meanings *within* dimensions. We operationalize stability as the correlation between words' projection on a given dimension in one decade and their projection in subsequent decades. Figure 9 displays the stability of projections for the 50,000 most common words on each class dimension. The first line represents the average correlation of word projections in the 1900s with their projections in the 1910s, 1920s, and so on through the 1990s. Similarly, the second line shows the correlation between projections in the 1920s with those in the 1930s, 1940s, and so on. For each decade, a word's projection is highly correlated with its projection the following decade, in most cases greater than .9. This correlation diminishes by

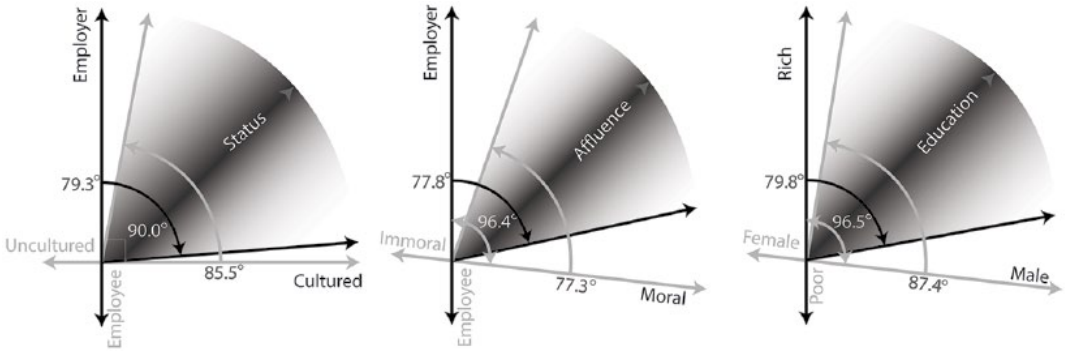


Figure 8. Conceptual Diagram of the Distortions Introduced When Reducing High-Dimensional Embeddings to Two-Dimensions

decade, however, such that the correlation between a word's projection in the 1900s and 1990s falls between .7 and .6. This pattern reveals that beneath the historic stability of class's dimensional structure, there is continuous flux in how cultural markers are positioned along these spectra.

To clarify this process of cultural circulation, we look more closely at how particular words change their projections on the employment dimension over the twentieth century. We select employment because it shows the greatest decline in correlation between the beginning and end of the century in Figure 9. Figure 10 displays the four highest and lowest loading words on the employment dimension in the beginning (1900s to 1910s) and end (1980s to 1990s) of the century, along with exemplary terms that display informative semantic trajectories. The top-left of the figure shows that many terms most strongly associated with the employer position were titles of formal office: "lords," "governor," "mayor," "earl," "bishop," and "secretary." As the century progresses, however, these titles lose ascendency to terms associated with power in an industrial and financialized economy: "promoter," "speculator," "rival," "designer," and "mogul."

The bottom of Figure 10 shows terms associated with the position of worker or employee. The strongest association at the start of the century is with "wage" and "earners"; this attenuates as a greater share of the U.S. workforce becomes contracted and salaried employees. The words "soldier," "muscle," and "bodied" project strongly on the "employee" end of the employment dimension during a period when manual labor comprised a large proportion of the workforce and World War I saw a large share of able-bodied workers enlisted into armed service. These words are displaced over time, with preeminent markers of "employee" at century's end including "retirement," "qualified," and "student." This suggests an emerging cultural image of the worker as white-collar and middle-class. Widespread perceptions of worker problems also shift with time. "Suffering" ceases to be a strong marker, but "unemployed" becomes prominent.

Other results of this analysis are not so easily interpretable. The words "patient" and "expectancy" are among the strongest negative projections on the ownership dimension at the end of the century, suggesting a powerful "employee" valence for both terms. Imaginative explanations for such findings are always conceivable—perhaps a growing recognition of workers as subject to ailment or injury led to an equivalence between "workers" and "patients." Yet this style of post hoc interpretivism is vulnerable to misleading conclusions drawn from statistical flukes. These ambiguous findings provide an instructive example of how inductive approaches must be applied cautiously to word embedding analyses.

DISCUSSION

Summary of the Argument and Results

In this article we introduce word embedding models as a productive method for the analysis of cultural categories and associations. By

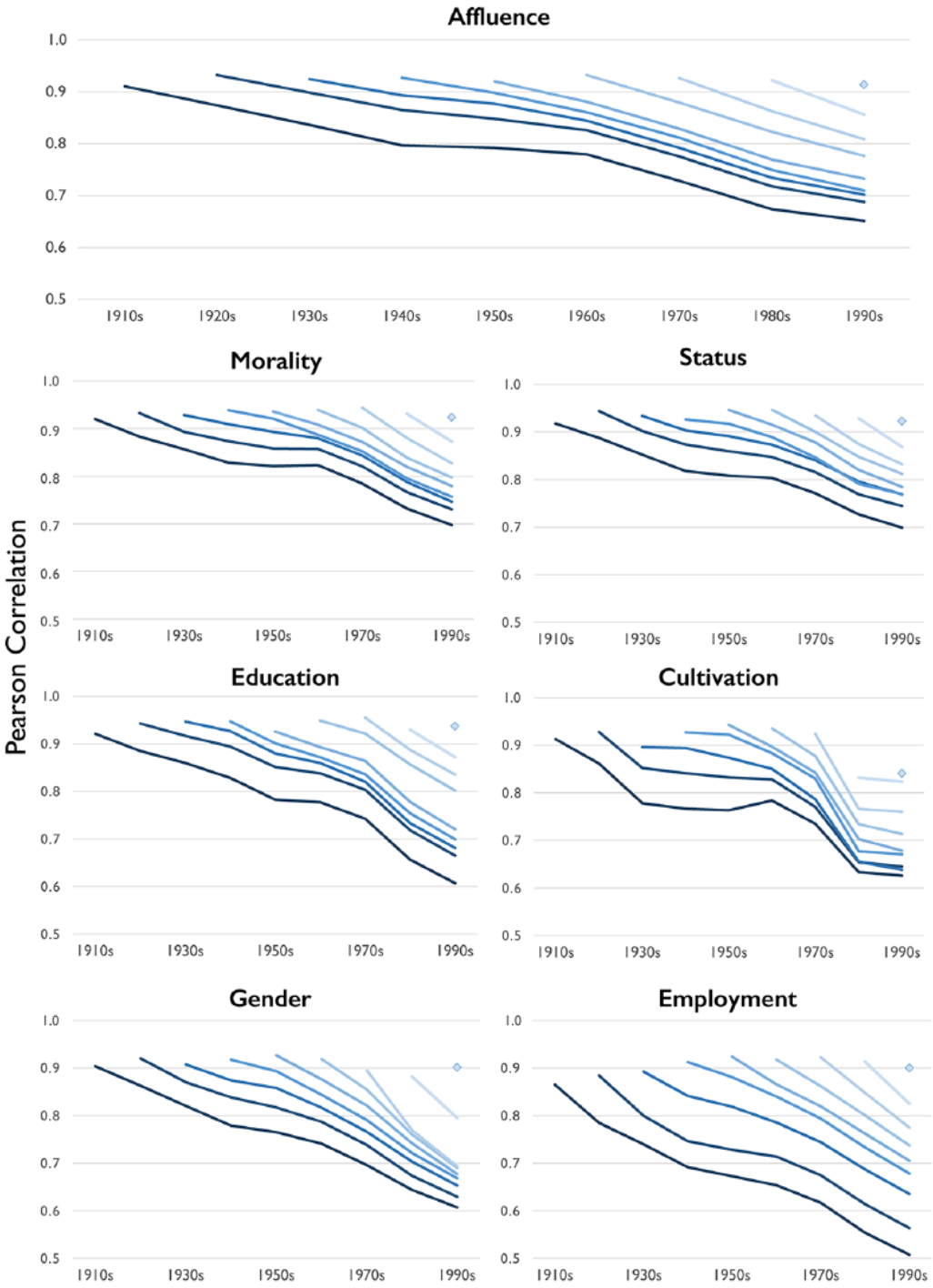


Figure 9. Correlation of 50,000 Most Common Words' Projection in One Decade with Their Projection in Each Subsequent Decade for Seven Cultural Dimensions of Class; 1900 to 1999 Google Ngrams Corpus

representing the relationship between words as the relationship between vectors in a high-dimensional vector space, word embedding models distill vast collections of text into a singular representation while preserving much of the richness and complexity of their semantic relations. We describe how dimensions of word embedding models correspond closely to "cultural dimensions" such as *rich-poor, good-evil*, and *masculine-feminine*, and how the positions of words arrayed on salient cultural dimensions of a word embedding

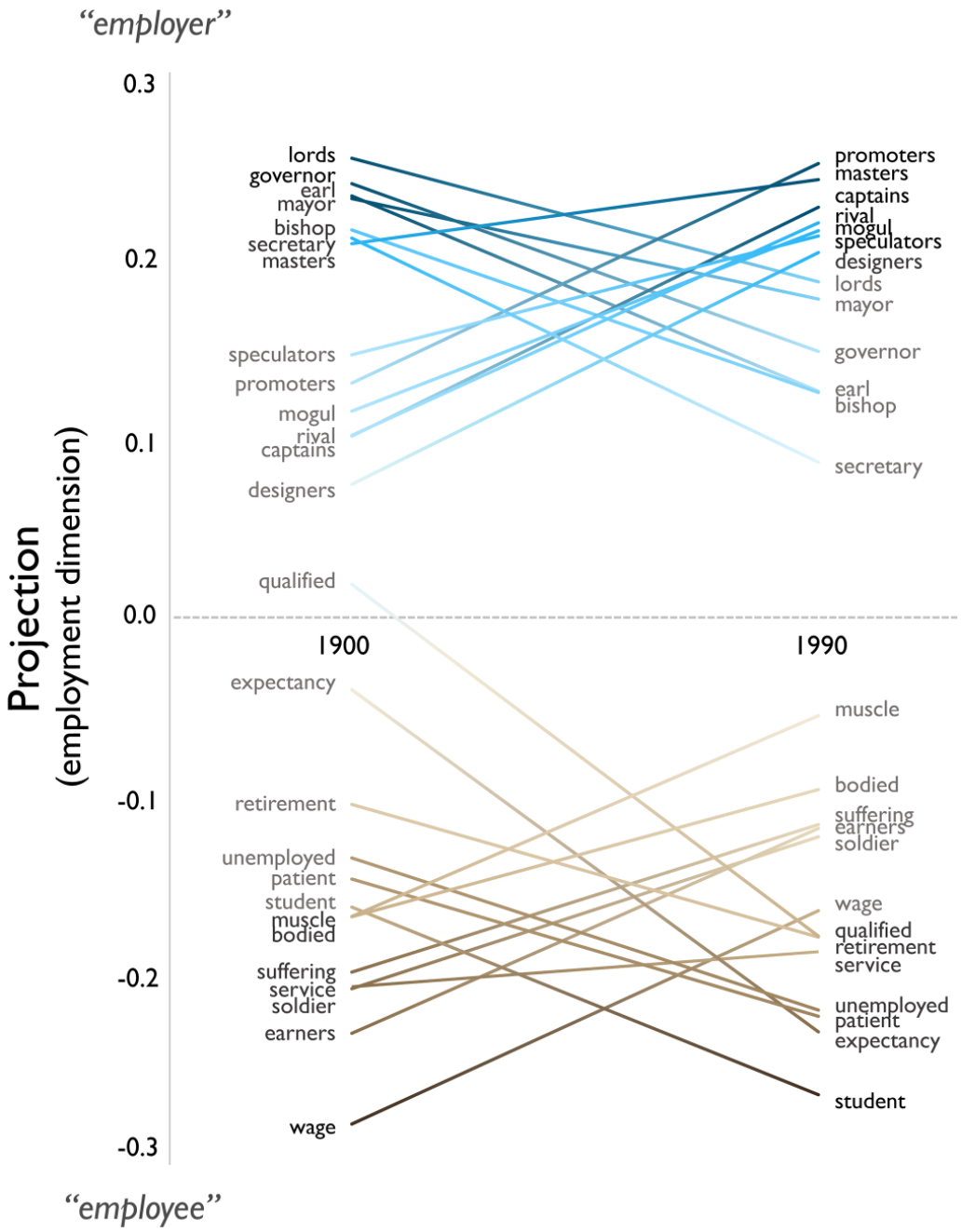


Figure 10. Words That Project High and Low on the Employment Dimension of Word Embedding Models Trained on Texts Published at the Beginning and End of the Twentieth Century; 1900–1919 and 1980–1999 Google Ngrams Corpus

reflect patterns of association and classification within a given cultural system. Furthermore, by calculating angles between cultural dimensions, we are able to investigate relationships between the axes of classification themselves.

After validating our method by comparing multiple word embeddings to contemporary and historical surveys of cultural associations, we apply it to a macro-historical investigation of shared understandings about social class in the United States over the twentieth century. We take up five facets of class and two related cultural dimensions that have been extensively theorized in the past. For each, we identify corresponding dimensions in word embedding models trained on texts produced over the twentieth century. We then measure relations between these class dimensions, bringing to light their dynamics,

but also their stability, in the face of economic and industrial transformation. Our findings reveal that the multiple dimensions of class identified in sociological theory comprise a complex yet stable semantic structure that can only be represented faithfully in high dimensionality. We find persistent, close relations between dimensions of cultivation, morality, and education, and these interrelated spectra are nearly orthogonal or negatively associated with cultural conceptions of the classic Marxian owner/worker relation. Nevertheless, both share a connection to status and affluence, which intermediate them, serving as a cultural nexus between the outward trappings of class and the social relations that produce and reproduce class in the modern world.

The relationships between the cultural dimensions of class remain stable over the century, but locations of individual words on those dimensions are in constant flux. Collectively, these findings suggest that many of the basic dimensions through which class is understood were robust against the twentieth century's tectonic shifts in the organization of economy, industry, and employment. What evolved were symbols used to signify locations in the multidimensional architecture of class.

General Implications of the Study of Culture

The full range of potential applications for word embedding models reaches far beyond the class example presented in this article. Following the general approach piloted here, analysts could use word embedding models to compare the cultural systems represented by literary genres, texts produced by distinct authors, or texts written in different languages (Lev, Klein, and Wolf 2015). A wide array of social collectives, including scientific disciplines, political elites, and contributors to online forums, can be analyzed and compared by training word embedding models on the text they produce. Furthermore, while this article focused on insights produced by identifying, extracting, or comparing "cultural dimensions" from the vector space, we do not maintain this is the only method for utilizing word embedding models to advance social science. Simply calculating the proximity of word vectors can also provide a strong indicator of the similarity or distance between word meanings (Kulkarni et al. 2015).

Word embedding models can further be used to classify and predict which group produced a text, given multiple corpora produced by distinct social groups (Taddy 2015a). Finally, future word embeddings that use hyperbolic or elliptical geometries could be used to systematically capture nonlinear relations in language, such as hierarchy or clustering (Chamberlain, Clough, and Deisenroth 2017; Nickel and Kiela 2017; see Appendix Part A). We argue that a wide range of techniques for productively developing and applying word embedding models to social and cultural inquiry are possible but yet to be developed. Nevertheless, Euclidean word embeddings are conducive to modeling and evaluating intersecting dimensions of culture in a way that maps onto a wide range of cultural theory.

Caveats and Limitations

As well as identifying broad potential, our investigation exposed clear limitations of word embedding models for cultural analysis. First, word embeddings must be trained on very large corpora if the output vector space is to capture subtle and complex associations of interest to culture analysts. Previous studies indicate that analogy tests can only be reliably solved when input text comprises several million words or more (Hill et al. 2014). As a result, groups that do not leave extensive textual records are difficult to study with word embeddings.

Second, the exact algorithmic processes undergirding the training of word embedding models can be highly complex and therefore elude theoretically parsimonious description. Although the word embedding models we present (*word2vec* and *GLoVe*) rely on two-layered neural networks with a single hidden layer, state-of-the-art deep-learning models

deploy many-layered neural architectures with hundreds of millions of parameters for improved performance on natural language and intelligence tasks like question-answering (e.g., Devlin et al. 2018). Added algorithmic complexity can produce more sensitive and informative models, but it may also diminish the researcher's understanding of how the model is generated and what distortions it is likely to produce.[21]

Moreover, word embeddings are not able to adjudicate the suitability of a given corpus for an investigation. Just as rigorous sampling is crucial in interview-based methods, the ability to make cultural inferences about a given group with word embeddings depends on the sample of text utilized in model training. In our analysis, we opted to use a broad sampling of U.S. texts over time. The magnitude of our corpora enables recovery of subtle and diffuse semantic relations, but it requires combining texts produced by very different groups across vastly different social and cultural contexts. The resulting model captures broadly shared meanings that characterize U.S. culture in a given decade, but it levels the cultural heterogeneity of the individuals and groups that articulated them.

Furthermore, we acknowledge that the voices and worldviews published in books digitized by Google are not a random sample of U.S. culture. Poor and marginalized populations are unlikely to have their discourses published in the books, periodicals, and pamphlets that comprise the Google Ngrams corpus. We therefore must limit our population of inference to the U.S. "literary public" in any given decade. The correspondence between these word embedding findings and surveyed Americans on Mechanical Turk suggests the models' associations are prevalent in the general public, but identifying exactly where this generalization succeeds and fails falls beyond the scope of this investigation.

A set of texts should not be taken as a pure or complete reflection of the culture that produced it. Authors may strategically emphasize or obscure semantic associations depending on their goals in producing the text (Jakobson 1960). Factors including the genre, purpose, and audience must be considered when utilizing texts for analysis and inference. Moreover, various elements of culture, such as tacit knowledge and embodied practices, are not inscribed in written discourse and therefore remain overlooked by formal models of text such as word embeddings (Lizardo 2017).

Finally, word embeddings cannot identify the cultural dimensions most important for a given semantic system or social process. Analysts can identify the cultural dimensions of the model that explain the most variance, either across the entire semantic space or within a circumscribed vocabulary. But although high explained variance indicates that terms have strong positive and negative valences along the dimension, it reveals little about how these valences are deployed in social life and to what ends. It is possible that subtle cultural associations may be deeply consequential for action, and thus explained variance could be misleading as an indicator of social significance. We argue that selection of cultural dimensions for analysis should be motivated by theoretical considerations, as we ultimately did here with class, rather than emergent and sometimes arbitrary qualities of the embedding space.[22]

Concluding Remarks

Despite their limitations, word embedding models can serve as a powerful tool for analysts of culture. Word embedding algorithms require large corpora to create informative models, but the amount of digitized text produced and available to analysts is growing exponentially (Evans and Aceves 2016; Salganik 2017). Although social scientists have widely recognized that these vast archives contain a cache of cultural information with great potential for analysis, scholars have remained limited by the available tools for integrating and analyzing large-scale text. Neural word embeddings present a method for producing rich models of semantic relationships from corpora too large for techniques such as topic modeling or semantic network

analysis. When adequate text is available, contemporary word embedding approaches can distill detailed and precise semantic information and relationships with a fidelity that exceeds prior methods and approaches human performance (e.g., Devlin et al. 2018).

Cultural dimensions from word embeddings do not simply provide a means of deriving textual measures compatible with intersectionality, affect control, and field theory. The very success of these models provides suggestive validation for relational approaches to cultural theorizing. Furthermore, by operationalizing a high dimensional model of culture, word embedding models allow researchers to extend and contend with theories of culture in novel ways. Embeddings present a much vaster set of potential axes along which individuals and social groups may compete, cooperate, fracture, or coalesce than low-dimension theories of cultural constraint allow. By simultaneously capturing the multiplicity of associations expressed in language, word embedding models are able to represent a complex geometry of culture, which, like a many-faceted crystal, amplifies the subtle and shifting framings that enable coordinated and spontaneous social action.

APPENDIX

Part A: Word Embeddings: History, Varieties, and Implementations

Word embedding models are naïve as to what words signify, lacking intrinsic word referents. They position words relative to one another based purely on how they are used in relation to one another. This process of identifying a word's meaning from context resonates with a tradition of practice-oriented theories of language in which word meanings are always understood through usage (Searle 1969; Wittgenstein 1953). The theory of meaning implicit in word embedding algorithms is well summarized by linguist J. R. Firth's (1957) dictum: "you shall know a word by the company it keeps."

Early approaches to word embedding, including Latent Semantic Analysis (LSA) or Indexing (LSI), have existed since the 1970s (Dumais 2004), but they initially involved the factorization of word-document matrices with singular value decomposition (SVD). Recent breakthroughs in autoencoding neural networks and advances in computational power have enabled a new class of word embedding models (Mikolov, Sutskever, et al. 2013) that can heuristically factorize much larger matrices (Levy and Goldberg 2014). This allows them to incorporate information about local semantic contexts from surrounding word windows rather than the entire documents. This one change, shifting from global to local context, has resulted in a punctuated increase in their accuracy on a wide range of tasks, from the analogy tests we detail to word classification (Taddy 2015b), question-answering (Zhou et al. 2015), and automated translation (Johnson et al. 2017). As a result, contemporary neural word embedding models distil an encyclopedic breadth of subtle and complex cultural associations from large collections of text by training the embedding model with local word associations a human might learn through ambient exposure to the same collection of language (Nagy, Herman, and Anderson 1985).

Word2vec can operate under two distinct model architectures: continuous bag-of-words (CBOW) or skip-gram. Under the CBOW architecture, the corpus is read line-by-line in a sliding window of k words, with k determined by the analyst. Previous studies have found windows of ~8 words produce the most consistent results (Le and Mikolov 2014). For each word in the corpus, the algorithm aims to maximize classification of the center word n, given its surrounding words within a context window of size k. The skip-gram architecture works similarly, except instead of predicting a word with context, it predicts context given a word. Related embedding approaches take into account additional information such as the "global" proximity of words within an overarching document (*GLoVe*, Pennington et al. 2014) or even

sub-word letter sequences in surrounding words (*fastText*, Joulin et al. 2016).[23]

To test the robustness of Google Ngrams and the *word2vec* algorithm, we evaluated our approach across different corpora and word embedding algorithms. Specifically, we compared the results from our survey of cultural associations with two widely used, publicly-available pre-trained embedding models. The first model we use to represent contemporary cultural associations is trained using the *word2vec* algorithm with CBOW architecture on 100 billion words scraped from Google News articles published by U.S. news outlets (Mikolov, Chen, et al. 2013). The second model was produced with the *GloVe* algorithm, which accounts for local and global dependencies, and is trained on a corpus collected as part of the Common Crawl, a broad scraping of millions of webpages (Pennington et al. 2014). A weakness of both of these publicly-available embeddings is that they lack satisfying documentation regarding the exact conditions of inclusion for texts in the corpus. This is unfortunately common among pre-trained embeddings, and it limits their utility for analysis of culture. Nevertheless, we include results from validations on these models to allow comparability with contemporary research in natural language processing and because embeddings greatly benefit from training upon massive quantities of text, and both of these embeddings are exceptional in this regard.

We note that the varieties of word embeddings analyzed and discussed here are all "Euclidean," meaning the space is defined by non-intersecting, parallel dimensions. Embedding in other geometries is possible and may be better suited for modeling semantic patterns other than "cultural dimensions." In a hyperbolic space, infinitely many lines may go through p without intersecting ℓ, and a central node may be close to many peripheral nodes without those nodes being close to each other. Embedding corpora in a hyperbolic geometry makes discovery of the semantic dimensions underlying them less straightforward, but it facilitates modeling semantic hierarchy.[24] Embedding semantic networks in hyperbolic space has facilitated automatic discovery of hypernyms—words with broad meanings under which specific instance words lie—such as the relationship between *animal, rodent*, and *rat*, or *color* and *red, green*, and *blue* (Chamberlain et al. 2017; Handler 2014; Nickel and Kiela 2017; Rei and Briscoe 2014). This might also enable discovery of holonyms and meronyms—words constituting wholes and their parts—like *hand, flesh*, and *fingers*. In this way, Euclidean word embeddings are tuned to capture semantic dimensions, but altering hidden parameters, such as the curvature of the underlying geometry, would allow them to capture other associations, like semantic hierarchy.[25]

Part B: Survey of Cultural Associations

Here we detail the Survey of Cultural Associations we fielded to produce a set of current, human-rated cultural evaluations for comparison against results from word embedding models. The survey was fielded through Amazon Mechanical Turk, an online service through which "requesters" can post a task and workers find and select tasks to complete in exchange for monetary compensation. Our survey was listed as "Sociological Survey," with the description "a fifteen-minute survey of cultural associations" and compensation of $1.75. The task was only available to Mechanical Turk workers located in the United States. The survey was fielded in two waves, October 2016 and December 2017 to samples of 206 and 200, respectively, of which a total of 398 respondents completed the survey. We pool the two waves in our analyses. Respondents were posed with the task of rating words on three scales, gender (very masculine to very feminine), race (very African American to very white), and class (very upper-class to very working-class). The set of 59 words they rated are listed in Table B1.

A number of previous studies have found that Mechanical Turk surveys fare well when compared to surveys with probability sampling, particularly when researchers measure and account for the sociodemographic

characteristics of the sample (Levay et al. 2016). Although Mechanical Turk's population of workers cannot be said to represent the general U.S. population, it is characterized by considerable diversity along racial, gender, and socioeconomic lines (Huff and Tingley 2015).

To mitigate any bias in estimates due to disproportionate representation of sociodemographic groups in the sample, we use post-stratification weighting to make our sample match the U.S. general population. We took population estimates from the U.S. Census Current Population Survey (CPS) of 2017 as population estimates for weighting our sample. We weighted along three strata: sex, education, and race. Sex is treated as two categories: male and female; education is divided into two categories: bachelor's degree or less than bachelor's degree; and race is divided into three strata: white, African American, or other. Results presented in this article include post-stratification weighting; however, additional analyses available upon request confirm that the inclusion of weights does not substantively alter results. Table B2 displays basic demographic characteristics of the sample.

In Table B3 we provide a more detailed summary of the correspondence between associations produced in Google News word embedding models and those reported by survey respondents, and we examine differences in performance between word domains. For all pairs of words that have a statistically significant difference in mean survey rating ($p < .01$) for class, race, and gender associations within a substantive domain, we calculate the proportion of pairs that are correctly ordered by the word embedding model trained on Google News text. For instance, if "steak" is significantly more upper-class than "hamburger" in the survey, we test if *steak* projects more masculine than *hamburger* in the embedding, and we then calculate the percentage of all such pairs of words that are correctly matched.

Table B3 shows that within most substantive domains, the rate of correct classification is above 80 percent and in many cases above 90 percent. It is also clear that embedding does a better job in domains with stronger cultural associations. For instance, there is very little difference in racial association between the clothing items included in the survey (standard deviation of 4.68), and in this domain the embedding has a low 55.0 percent rate of matching the survey. In first names, however, where signals are stronger (standard deviation of 32.46), the same dimension of the word embedding correctly matches 94.7 percent of differences in the survey.

Table B1. List of Words Rated in Cultural Associations Survey

Occupations	Clothing	Sports	Music Genres	Vehicles	Food	First Names
Banker	Blouse	Baseball	Bluegrass	Bicycle	Beer	Aaliyah
Carpenter	Briefcase	Basketball	Hip hop	Limousine	Cheesecake	Amy
Doctor	Dress	Boxing	Jazz	Minivan	Hamburger	Connor
Engineer	Necklace	Golf	Opera	Motorcycle	Pastry	Jake
Hairdresser	Pants	Hockey	Punk	Skateboard	Salad	Jamal
Journalist	Shirt	Soccer	Rap	SUV	Steak	Molly
Lawyer	Shorts	Softball	Techno	Truck	Wine	Shanice[a]
Nanny	Socks	Tennis				Tyrone
Nurse	Suit	Volleyball				
Plumber	Tuxedo					
Scientist						

[a]Word did not appear frequently enough in the 2000 to 2012 Google Ngrams to appear in the embedding model and is therefore excluded from 2000 to 2012 Google Ngrams analyses.

Table B2. Descriptive Statistics for Mechanical Turk Sample and Census CPS Sample

	Mechanical Turk	Census CPS
Gender (1 = female)	43.47%	51.76%
Education		
High school, GED, or less	12.31%	39.99%
Some college	26.88%	18.83%
Associate's degree	10.05%	9.75%
Bachelor's degree	43.47%	20.03%
Graduate degree	7.29%	11.39%
Race/Ethnicity		
African American	6.53%	12.52%
White	79.15%	78.22%
Other	14.32%	9.26%
Hispanic	9.82%	15.92%
Age (mean)	34.40	47.20
N	398	135,137

Table B3. Percentage of Statistically Significant ($p < .01$) Survey Differences Correctly Classified in Google News Word Embedding Model

	Sports	Food	Music	Occupations	Vehicles	Clothes	Names	All Domains
Gender	87.9%	88.2%	72.2%	93.6%	82.4%	74.4%	95.2%	84.8%
Class	96.3%	93.8%	88.9%	60.9%	94.1%	90.0%	77.3%	75.3%
Race	90.0%	68.8%	100%	51.5%	87.5%	55.0%	94.7%	69.1%

Part C: Statistical Significance of Distances and Associations

We propose well-established nonparametric bootstrapping and subsampling methods to show the stability and significance of word associations within our embedding model. This approach allows us to establish conservative confidence or credible intervals for both (a) distances between words in a model and (b) projections of words onto an induced dimension (e.g., *affluence-poverty*). If we assume the texts underlying our word embedding model are observations drawn from an independent and identically distributed (i.i.d.) population of cultural observations, then bootstrapping allows us to estimate the variance of word distances and projections by measuring those properties through sampling the empirical distribution of texts with replacement (Efron 2003; Efron and Tibshirani 1994).

To estimate bootstrapped 90 percent confidence intervals, the analyst draws documents with replacement from the corpus to construct 20 new corpora, each the size of the original corpus. The analyst then estimates either word similarities or angles between vectors on all 20 of these new corpora. The 2nd order (2nd smallest) estimated statistic $s_{(2)}$ is taken as the confidence interval's lower bound and the 19th order statistic $s_{(19)}$ as its upper bound. The distance between $s_{(2)}$ and $s_{(19)}$ across 20 bootstrap samples span the 5th to the 95th percentiles of the statistic's variance, bounding the 90th confidence interval. A 95 percent confidence interval would span $s_{(2)}$ and $s_{(39)}$ in word embedding distances or projections estimated on 40 bootstrap samples of a corpus, tracing the 2.5th to 97.5th percentiles. Due to the limits of corpus size, we use this bootstrapping approach to conduct statistical significance tests for our JSTOR models.

If the corpus is very large, however, we may take a subsampling approach, which randomly partitions the corpus into non-overlapping samples, then estimates the word

embedding models on these subsets and calculates confidence or credible intervals as a function of the empirical distribution of distance or projection statistics and number of texts in the subsample (Politis, Romano, and Wolf 1997). Subsampling relies on the same i.i.d. assumption as the bootstrap (Politis and Romano 1992, 1994). For 90 percent confidence intervals, we randomly partition the corpus into 20 subcorpora, then calculate the error of our embedding distance or projection statistic s for each subsample k as $B^k = \sqrt{\tau_\kappa}(s^k - \overline{s})$, where τ_κ is the number of texts in subsample k, s^k is the embedding distance or projection for the k_{th} sample, and $\overline{s}$ is the mean of the 20 estimates. The 90 percent confidence interval spans the 5th to 95th percentile variances, inscribed by $\overline{s} - \frac{B^K_{(19)}}{\sqrt{\tau}}$ and $\overline{s} - \frac{B^K_{(2)}}{\sqrt{\tau}}$ where τ is the number of texts in the total corpus. As with bootstrapping, a 95 percent confidence interval would require 40 subsamples; a 99 percent confidence would require 200 (.5th to 99.5th percentiles). We use this subsampling approach to construct confidence intervals for our Google Ngrams models.

A great benefit of bootstrapped and subsampled confidence intervals is that they reflect how robust an association is across texts. If a word occurs only rarely or is used in a diffuse set of very distinct contexts, the word's position in the vector space will be radically different between subsamples and therefore will produce larger confidence or credible intervals. On the other hand, words that are frequently used in consistent contexts will hold more stable positions across the subsamples and hence produce smaller confidence or credible intervals.

Part D: Word Pair Lists

Table D1. Word Pairs Used to Construct Affluence, Gender, and Race Dimensions for Amazon Mechanical Turk Survey Validation

Affluence		Gender	Race
rich-poor	precious-cheap	man-woman	black-white
richer-poorer	priceless-worthless	men-women	blacks-whites
richest-poorest	privileged-underprivileged	he-she	Black-White
affluence-poverty	propertied-bankrupt	him-her	Blacks-Whites
affluent-destitute	prosperous-unprosperous	his-her	African-European
advantaged-needy	developed-underdeveloped	his-hers	African-Caucasian
wealthy-impoverished	solvency-insolvency	boy-girl	Afro-Anglo
costly-economical	successful-unsuccessful	boys-girls	
exorbitant-impecunious	sumptuous-plain	male-female	
expensive-inexpensive	swanky-basic	masculine-feminine	
exquisite-ruined	thriving-disadvantaged		
extravagant-necessitous	upscale-squalid		
flush-skint	valuable-valueless		
invaluable-cheap	classy-beggarly		
lavish-economical	ritzy-ramshackle		
luxuriant-penurious	opulence-indigence		
luxurious-threadbare	solvent-insolvent		
luxury-cheap	moneyed-moneyless		
moneyed-unmonied	rich-penniless		
opulent-indigent	affluence-penury		
plush-threadbare	posh-plain		
luxuriant-penurious	opulence-indigence		

Table D2. Word Pairs Used to Reconstruct 20 Semantic Differential Dimensions from Jenkins and Colleagues (1958) for Historical Survey Validation

soft-hard	**foolish-wise**	**unimportant-important**	**fast-slow**
supple-tough	dumb-smart	inconsequential-consequential	quick-lagging
delicate-dense	irrational-rational	secondary-principal	rapid-unhurried
pliable-rigid	stupid-thoughtful	irrelevant-major	speedy-sluggish
fluffy-firm	unwise-sensible	trivial-crucial	swift-gradual
mushy-solid	silly-reasonable	negligible-critical	quickly-slowly
softer-harder	ridiculous-enlightened	insignificant-significant	swiftly-gradually
softest-hardest	unintelligent-intelligent	unnecessary-essential	faster-slower
		peripheral-central	fastest-slowest
unusual-usual	**excitable-calm**	**strong-weak**	**colorful-colorless**
different-customary	volatile-tranquil	powerful-powerless	brilliant-uncolored
abnormal-normal	nervous-still	muscular-frail	bright-pale
irregular-regular	tempestuous-serene	brawny-feeble	radiant-drab
odd-standard	fiery-peaceful	strapping-puny	vivid-pallid
atypical-typical	emotional-restful	sturdy-fragile	vibrant-lackluster
unexpected-expected	jumpy-sedate	robust-flimsy	colored-bleached
unconventional-conventional	unsettled-settled	vigorous-languid	
rounded-angular	**passive-active**	**true-false**	**ugly-beautiful**
circular-cornered	immobile-mobile	true-untrue	unattractive-attractive
round-pointed	lethargic-energetic	verifiable-erroneous	unsightly-pretty
dull-sharp	frail-vital	veracious-fallacious	hideous-handsome
smooth-jagged	subdued-vigorous	accurate-inaccurate	grotesque-gorgeous
spherical-edged	static-dynamic	faithful-fraudulent	repulsive-cute
	subdued-lively	correct-incorrect	
feminine-masculine	**bad-good**	**successful-unsuccessful**	**old-new**
woman-man	worst-best	victorious-failed	aged-recent
women-men	deficient-fine	triumphant-abortive	ancient-contemporary
she-he	inferior-superior	winning-losing	decrepit-fresh
her-him	unsatisfactory-satisfactory	thriving-failing	elderly-young
her-his	unacceptable-acceptable	fruitful-fruitless	historic-modern
hers-his	awful-excellent	prosperous-ineffectual	adult-child
girl-boy	terrible-superb	success-failure	older-newer
girls-boys	dreadful-outstanding	win-lose	oldest-newest
female-male	unexceptional-exceptional		
kind-cruel	**straight-curved**	**timely-untimely**	**tasteless-savory**
tender-callous	linear-nonlinear	punctual-late	bland-tasty
compassionate-heartless	unswerving-swerving	ready-unready	flavorless-flavorful
humane-inhumane	unbending-bent	prompt-delayed	unappetizing-delectable
merciful-merciless	untwisted-twisted	reliable-unreliable	mild-piquant
gentle-brutal	direct-meandering	early-late	insipid-succulent
nice-unpleasant	undeviating-serpentine	earlier-later	dull-delicious
kindest-cruelest	straighter-curvier	earliest-latest	blandest-tastiest

Note: Terms used by Jenkins and colleagues to specify dimensions are in bold.

Table D3. Word Pairs Used to Construct Class Dimensions (Along with Affluence and Gender in Table D1)

Cultivation	Employment	Education	Status	Morality
cultivated-uncultivated	employer-employee	educated-uneducated	prestigious-unprestigious	good-evil
cultured-uncultured	employers-employees	learned-unlearned	honorable-dishonorable	moral-immoral
civilized-uncivilized	owner-worker	knowledgeable-ignorant	esteemed-lowly	good-bad
courteous-discourteous	owners-worker	trained-untrained	influential-uninfluential	honest-dishonest
proper-improper	industrialist-laborer	taught-untaught	reputable-disreputable	virtuous-sinful
polite-rude	industrialists-laborers	literate-illiterate	distinguished-commonplace	virtue-vice
cordial-uncordial	proprietor-employee	schooled-unschooled	eminent-mundane	righteous-wicked
formal-informal	proprietors-employees	tutored-untutored	illustrious-humble	chaste-transgressive
courtly-uncourtly	capitalist-proletarian	lettered-unlettered	renowned-prosaic	principled-unprincipled
urbane-boorish	capitalists-proletariat		acclaimed-modest	unquestionable-questionable
polished-unpolished	manager-staff		dignitary-commoner	noble-nefarious
refined-unrefined	managers-staff		venerable-unpretentious	uncorrupt-corrupt
civility-incivility	director-employee		exalted-ordinary	scrupulous-unscrupulous
civil-uncivil	directors-employees		estimable-lowly	altruistic-selfish
urbanity-boorishness	boss-worker		prominent-common	chivalrous-knavish
politesse-rudeness	bosses-workers			honest-crooked
edified-loutish	foreman-laborer			commendable-reprehensible
mannerly-unmannerly	foremen-laborers			pure-impure
polished-gruff	supervisor-staff			dignified-undignified
gracious-ungracious	superintendent-staff			holy-unholy
obliging-unobliging				valiant-fiendish
cultured-uncultured				upstanding-villainous
genteel-ungenteel				guiltless-guilty
mannered-unmannered				decent-indecent
polite-blunt				chaste-unsavory
				righteous-odious
				ethical-unethical

Part E: Constructing Cultural Dimensions

Here we perform tests to reveal design principles for the construction of cultural dimensions. We begin by considering the set of antonym pairs needed to effectively approximate a cultural dimension. The English language contains a vast vocabulary for denoting affluence and poverty, and drawing on five thesauri, we assembled a list of 42 pairs of terms that closely correspond to this cultural dimension. The very selection of antonym pairs presents two methodological difficulties. First, it is not clear how many antonym pairs are required to approximate the cultural dimension of interest. Second, terms do not always have a single obvious antonym, so constructing pairs requires subjective judgment on the part of the researcher. We investigate both of these issues in our validation of the affluence dimension.

First, we test if using a greater number of antonym pairs in constructing a cultural dimension is associated with improvements in correlation between projections on that dimension and human-rated associations from survey data. To accomplish this, we found the average correlation between surveyed class association and projection on an affluence dimension constructed with a single antonym pair, two antonym pairs, three antonym pairs, through all 42 pairs. Results are presented in Figure E1.[26] Cultural dimensions constructed from single antonym pairs fare relatively poorly, with their projections correlating on average at .2 with surveyed response. Correlations with survey response rise as a greater number of antonym pairs are used to construct the cultural dimension, but the gains in correlation from adding additional antonym pairs shrinks. In the following analyses, we use the full 42 antonym pairs for our affluence dimensions to improve robustness and decrease chance variability.

Next we test the extent to which the precise pairing of words affects correlation with survey data. To do this, we take our sets of 42 "rich" synonyms and 42 "poor" synonyms, and we re-pair them in random permutations. For instance, "rich" may be paired with "impoverished" instead of "poor." We then construct the affluence dimension using this set of randomly paired, roughly antonym terms, and we correlate its projections with the survey data. On average, it performs only marginally worse than our curated pairs of antonyms.

Finally, to eliminate the element of analyst judgment in pairing, we try a third strategy of averaging all "rich" synonyms together and subtracting the average of all the "poor" synonyms, an approach we label the "grouped pairs" method. The result of this operation is very similar to the one we propose, but it is mathematically distinct because it involves the averaging of vectors *before* performing the nonlinear operation of cosine similarity. Once again, we find substantively similar results, as displayed in Figure E1.

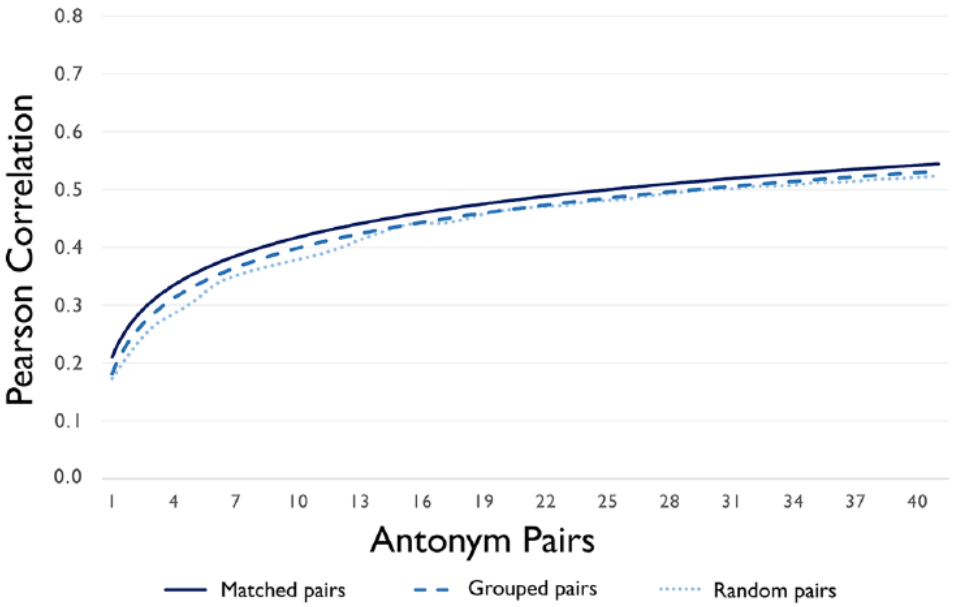


Figure E1. Average Correlation between Survey-Rated Class Associations and Word Embedding Projections on the Cultural Dimension of Affluence, Constructed with 1 to 42 Antonym Pairs; Google Ngrams 2000 to 2012 Word Embedding, Smoothed to Clarify Trend

Part F: Corpus Sizes

Table F1. Sizes of Google Ngrams and JSTOR Corpora by Decade

Decade	Google Ngram Word Count	JSTOR Word Count	JSTOR Article Count
1900s	3.0×10^{10}	4.7×10^{7}	1,294
1910s	2.9×10^{10}	5.7×10^{7}	2,020
1920s	2.4×10^{10}	8.4×10^{7}	3,266
1930s	2.1×10^{10}	1.1×10^{8}	4,228
1940s	2.2×10^{10}	1.6×10^{8}	5,923
1950s	2.9×10^{10}	2.2×10^{8}	7,442
1960s	5.0×10^{10}	3.5×10^{8}	10,152
1970s	5.9×10^{10}	6.8×10^{8}	17,855
1980s	7.2×10^{10}	8.9×10^{8}	19,830
1990s	1.2×10^{11}	1.2×10^{9}	20,698
2000–12	2.5×10^{11}		

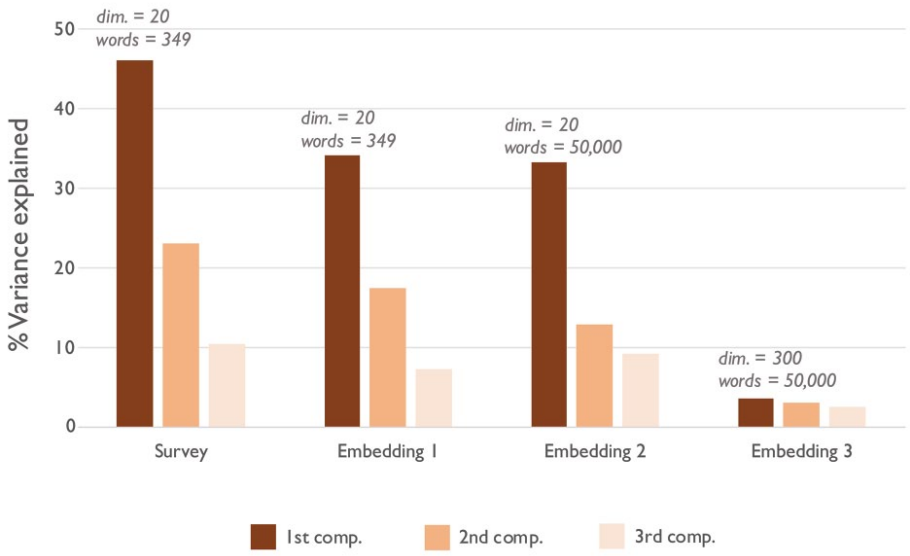


Figure G1. Variance Explained in Principal Components Analysis of 1958 Semantic Differential Survey Data and from Three Datasets of Projections from the 1950s Google Ngrams Embedding

Part G: Semantic Differential Validation

We draw on Jenkins and colleagues' (1958) semantic differential dataset again to conduct a close comparison between the semantic spaces produced by word embeddings and those produced by the semantic differential method. A major finding of Osgood and colleagues' (1957) research program was that the data produced with the semantic differential method could be reduced to relatively low dimensionality with only modest loss of information. They consistently found that when they subjected the matrix of word associations to principal components analysis (PCA), the first three components captured upward of 70 percent of the total variance of the semantic spaces. We validate this finding but show that the same kind of successful dimension reduction is not possible across all dimensions with word embedding models, suggesting the importance of higher dimensionality in analyzing culture. Results are presented in Figure G1.

First, we conduct PCA on Jenkins and colleagues' 1958 dataset of human-rated associations. As anticipated by semantic differential theory, the great majority of the variance of the 20-dimensional space is explained by the first three components. Second, we construct an embedding-derived dataset that mirrors the dataset of human ratings by projecting the same set of 349 terms onto 20 cultural dimensions corresponding to those measured by Jenkins and colleagues (1958) (see Table D2). Conducting PCA on this embedding-derived dataset, we find a comparably high percent of the total variance is explained by the first three principal components. Third, we expand the embedding-derived dataset from the set of 349 words used by Jenkins and colleagues to the set of 50,000 most commonly used words in the 1950s Ngrams corpus, while restricting to the same 20 semantic dimensions specified by Jenkins and colleagues. Again, most of the variance is explained by the first three dimensions. Finding that the projections of 50,000 common terms can similarly be reduced to low dimensionality suggests the ability to compress the space to three dimensions does not result from the particular set of terms rated.

Finally, we perform PCA on the full, 300 dimensional *word2vec* output model for the 50,000 most common words. Figure G1 shows the first component explains only 3.4 percent of the variance in the entire vector space. The stark difference in results between this and the previous analyses suggests the information in semantic spaces produced by word

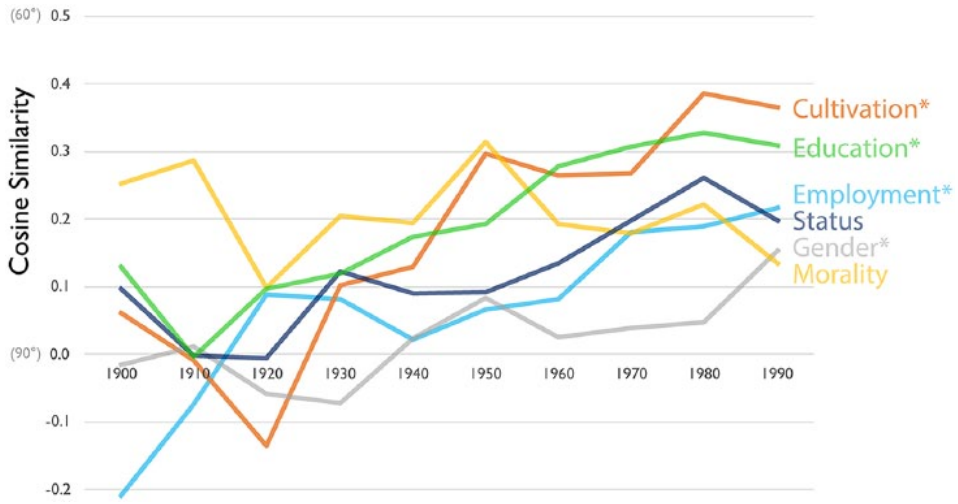


Figure H1. Cosine Similarity of the Affluence Dimension with Six Other Class Dimensions in Sociology Texts; 1900 to 1999 JSTOR Sociology Corpus
Note: Asterisks represent statistically significant difference between angles in the 1900s and 1990s ($p < .10$).

embeddings is diffusely spread across its many dimensions. This finding suggests these additional dimensions of the word embedding provide information not contained in the semantic differential spaces.[27] These additional dimensions likely make it possible to capture subtle or unusual semantic dimensions, such as *American-French* or *city-state* (Mikolov, Chen, et al. 2013), which would be missed by the standard semantic differential approach.

Part H: Embedding Sociological Discourse

We contextualize our cultural analysis of class by comparing associations held in the general public to those expressed in sociological literature. Here we display results from word embedding models trained on a corpus of all sociology articles published in the twentieth century in the JSTOR collection. This corpus includes 121 English-language periodicals, ranging from *American Sociological Review* and *Sociological Methods & Research* to *Poetics, Social Problems*, and *Symbolic Interaction*. As with the Google Ngrams, we divide the JSTOR sociology corpus into 10-year windows, training *word2vec* embedding models for each decade of the twentieth century. The class-based associations we find in this corpus generally accord with widely recognized trends in the discipline, so we use this analysis not to produce new discoveries but to allow formal comparison with the embeddings trained on general discourse. Details regarding the size of the JSTOR corpus are available in Appendix Part F.

Figure H1 is analogous to Figure 5 in the main text, but it presents results from the JSTOR corpus rather than the Google Ngrams corpus. As described in the main text, there are several places where sociological understandings of class depart from conventional associations. First, while femininity maintains a persistent association with affluence in general discourse, masculinity becomes identified with affluence in the second half of the twentieth century in sociological texts, evincing the discipline's growing concern for gender inequality. Second, within sociology, the latter half of the twentieth century witnesses a heightened association between the affluence and employment dimensions, with owners and bosses becoming increasingly marked as wealthy relative to workers and staff. This strong association between employment and affluence at the end of the century contrasts with the middling association found in the Ngrams embeddings, and it may reflect sociology's focus on structural sources of stratification and the influence of Marxian thought. Finally, cultivation only emerges as a strong

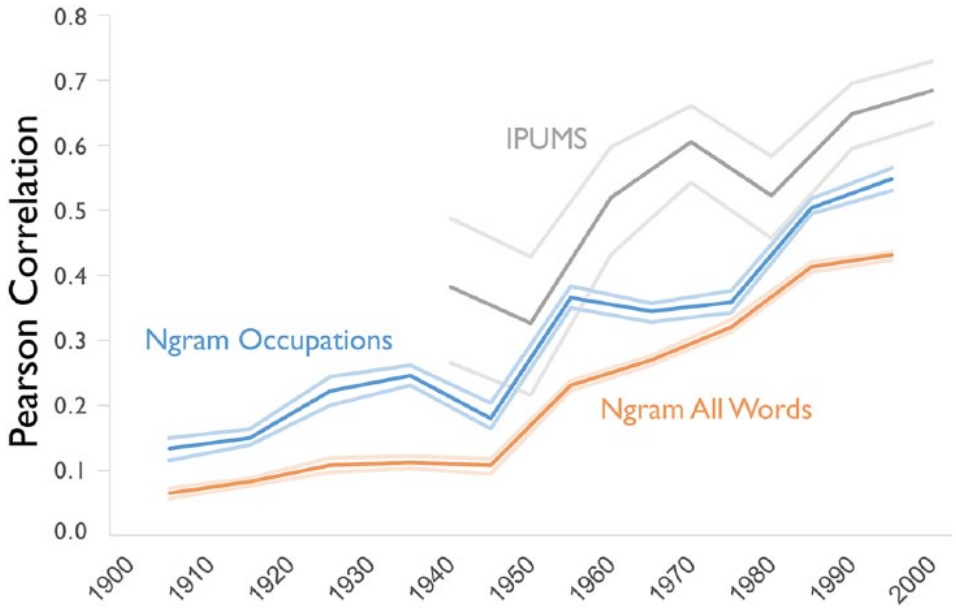


Figure I1. Correlations of Affluence and Education from IPUMS Surveys and Google Ngrams Text
Note: Correlation of occupations' average income and average education by decade; correlation of occupation names' projections on affluence and education dimensions; and correlation of all words' projections on affluence and education dimensions.

marker of affluence at the end of the century in sociological texts, whereas the two dimensions display a persistent association in general discourse. This tightening relationship coincides with the discipline's growing awareness of how self-presentation and cultural capital fuel class reproduction.

Part I: Cultural versus Material Changes in Education

The most striking change in the transformation of class associations revealed in Figures 5 and 6 occurs between the dimensions of affluence and education. Their association is weakly positive at the dawn of the twentieth century, but it becomes visibly stronger in the second half of the century. Figure I1 shines light on the growing semantic connection between affluence and education by displaying multiple indicators of this relationship over time, drawing on word embedding projections and their complex relationship with census data from IPUMS (Ruggles et al. 2019).

First, for all occupations reported in the census within a given decade, we calculate the correlation between the average reported income and average years of formal education for all individuals from that occupation in the IPUMS data. We observe a distinct spike in the correlation between an occupation's average income and education level between 1950 and 2000. Unfortunately, the census did not collect data on education level prior to 1940, but extensive historical records confirm that, while education has long shown returns to income, it played a substantially smaller role in shaping social stratification in the beginning of the twentieth century (Collins 1979; Goldin and Katz 1999).

To compare the material trend to the cultural trend, we take the names of all occupations reported by the census in a given decade and project them on both the affluence and education dimensions of the word embedding trained on the respective decade of Google Ngrams text. We then calculate the correlation between occupations' projections on education and affluence for each decade. We find an upward trend in semantic association between education and affluence among the occupations beginning in the 1950s that mirrors the socioeconomic trend. Finally, we correlate the projections on the education and

affluence dimensions of the 50,000 most common words in each decade. We find that the correlation among all words follows a parallel trend, exhibiting a steady increase that begins mid-century. Its overall levels of correlation are consistently lower than when the vocabulary is limited to occupations. These findings suggest a growing correspondence between the cultural associations of education and affluence coincided with material shifts that drew these dimensions of class materially together in the economy. Furthermore, while this semantic convergence is particularly apparent in the domain of occupations, it is also evident to a lesser extent in the general lexicon.

Editors' Note

Figure 3, Figure 10 and Table B3 were incorrect in the Online First version. They have now been corrected online and in print, along with two related values in the text.

Acknowledgments

We thank John Levi Martin, Etienne Ollion, and Marion Fourcade for their helpful comments. An earlier version of this paper was presented at the 2017 American Sociological Association Annual Meeting in Montreal, QC.

ORCID iDs

Austin C. Kozlowski https://orcid.org/0000-0001-8458-1129
James A. Evans https://orcid.org/0000-0001-9838-0707

Notes

1. Word embedding models are sometimes considered "low dimensional" relative to the number of words used in text (e.g., 50,000) because they reduce this *very* high dimensional word space. Nevertheless, considered from the perspective of one-, two-, or three-dimensional models common in the analysis of culture, these spaces are much more complex and reproduce much more accurate cultural associations, as we will show.
2. The cultural dimensions of race are also strongly linked to collective understandings of class. However, historical analyses of race with word embeddings present methodological challenges that require a unique and careful treatment that is beyond the scope of this investigation, as we will detail.
3. Such networks could be made dense and their links weighted, encoding a myriad of word collocations, but analysis of the resulting hairball would require a calculus that deviates widely from standard network analysis, such as one based on random walks (Rosvall and Bergstrom 2008; Shi, Foster, and Evans 2015) or simulated flows over implied curvature (Jost and Liu 2014).
4. Scientists have attempted to perform these parametrically, as with exponential family embedding models, but their performance has not yet approached that of autoencoders (Rudolph et al. 2016).
5. The surface area of a unit circle surpasses its volume in three dimensions. As a hypersphere's dimension approaches infinity, its volume approaches zero.
6. A shallow, two-layer neural network word embedding like *word2vec* constrains semantic dimensions to be linear as they are in PCA or SVD. Deeper neural network embedding models allow estimation of nonlinear semantic dimensions (Devlin et al. 2018).
7. Multiple word embedding approaches have become widespread in recent years, but the analyses we present here primarily utilize the skip-gram models in *word2vec*, cross-validated with those from *GloVe* (Pennington et al. 2014). The methodological principles outlined here, however, reach beyond neural-network autoencoders and are generally applicable to word embedding models constructed with other algorithms, including Latent Semantic Analysis based on SVD (Dumais 2004) and Bayesian nonparametric estimation (Rudolph et al. 2016).
8. We find that this calculus produces nearly identical results to a similar approach of first averaging the words on each side of the semantic dimension and then taking the difference between the two averages.
9. Because the cultural category of race is itself multidimensional, its representation in word embeddings is multidimensional as well. We restrict our analyses to the *black-white* dimension, but other word pairs, such as *hispanic-white* or *hispanic-black*, similarly capture meaningful semantic relations.
10. The signs may be flipped, of course, making positive values reflect low-class associations if poverty terms are subtracted from affluence terms, that is, by using *poverty – affluence* instead of *affluence – poverty*.
11. We know this from the Gaussian Annulus theorem, that two random points from a d-dimensional Gaussian with unit variance in each direction are approximately orthogonal (Blum, Hopcroft, and Kannan 2016).
12. "Bias" in this literature refers to harmful negative stereotypes, not the statistical definition of the term.
13. In a few instances, Bourdieu (1984:266, 343) notes other dimensions that structure the social topography, such as upward or downward trajectory and a preference for the traditional versus the innovative, but these dimensions have not enjoyed the same systematic treatment or theoretical elaboration as economic and cultural capital.
14. Code used in this analysis is available at: https://github.com/KnowledgeLab/GeometryofCulture.
15. First names were sampled from lists of names found to be most predictive of belonging to an African American person and most predictive of belonging

to a non-Hispanic white person for each sex from data of all children born in California from 1961 to 2000 (Fryer and Levitt 2004). Terms in other domains were selected based on known race, class, and gender markers that have been examined in previous literature.

16. In preliminary analyses, we trained word embeddings on collections of both 5-grams and 4-grams, but we found they performed poorer on survey validation than models trained on 5-grams alone. Because all information in word embedding models comes from neighboring words, it is unsurprising that smaller context windows produce weaker models.
17. We note that an analyst could discover differential weights for each word pair by estimating them with exploratory factor analysis or within a linear model that predicts surveyed cultural associations. A weighted sum necessarily improves the correlation of our dimension with surveyed associations, but it would be fragile for analysis of historical culture where the weights likely change, and we can field no surveys in the past.
18. Each of these cultural dimensions refers to a specific vector within an embedding model. Nevertheless, we reserve vector notation for individual word vectors, indexed by the precise word under the vector symbol (e.g., $\overrightarrow{man}$ refers to the vector associated with the word "man" in the embedding). For consistency with theoretical discussions earlier and later in the text, we do not use vector notation but assume it for vectors such as affluence, status, and cultivation, which comprise the average of the difference between many specific word vectors (see Tables D1 and D3).
19. Substantively similar results are produced when we correlate projections of all words on the two dimensions instead of calculating the angle between the dimensions. This correlational approach more closely derives from our validations, but we chose to display angles between dimensions to underscore the geometric rendering of cultural meaning inherent to word embedding models.
20. We acknowledge that because our embedding space was constructed with a single optimization algorithm, word projections on one semantic dimension are not independent from those on another. Nevertheless, as we show in Figure 1, there are many dimensions and degrees of freedom that limit the influence of this singular dependence even for semantically proximate associations. Moreover, because we do not seek to generalize beyond the texts within our substantial sample, we do not violate the assumptions of the OLS framework, which allows us to directly ask the degree to which word shifts in the projection along one dimension are a function of their position on another, holding constant their position on a third.
21. Complex neural network models are not statistical objects, in that their heuristic methods of optimization cannot (yet) be characterized by a σ-algebra, which details the full range of parameters searched on the path to the final, fitted model. This means fitted models, including those in this article, lack proof that they are the best models of their kind, despite successful performance on language and culture tasks.
22. Early in this project, we attempted the quixotic feat of inductively identifying the "most important" semantic dimension in word embedding space. To accomplish this, we collected every pair of antonyms in English from the digital dictionary WordNet (Miller 1995) and calculated the total variance in the vector space explained by each pair. Our analysis revealed one dimension that explained more semantic variance across the entire lexicon than any other: *steroidal–nonsteroidal*. After feeling like the crew from Douglas Adams's *The Hitchhiker's Guide to the Galaxy* when they find that the "Answer to the Ultimate Question of Life, the Universe, and Everything" computed by the supercomputer Deep Thought over 7.5 million years is "42" (Adams, Brett, and Perkings 1978), we caution against "theory-free" approaches to meaning discovery.
23. Although we do not use them in this analysis, the m contexts by k dimensions matrix in Figure 1 also retains a great deal of semantic information and has been used in concert with word embeddings to identify words that are complements versus substitutes in text (Nalisnick et al. 2016; Ruiz, Athey, and Blei 2017).
24. Words have highly unequal frequencies, with some central to common language and others peripheral (Zipf 1932).
25. Embedding networks in hyperbolic geometry typically requires fewer dimensions than in Euclidean space, both because the space may better reflect the intrinsic geometry of the data, and because there is "more space" in a d-dimensional Poincare ball, where the volume rises exponentially relative to the surface area, than in a Euclidean hypersphere of the same dimension.
26. Calculating the average correlation for dimensions constructed from all possible combinations of antonym pairs, although possible, is computationally impractical. For example, there are more than 500 billion ways to select 21 of the 42 antonym pairs. Instead, we sample 400 randomly selected combinations of pairs and calculate the average correlation in the sample for each number of antonym pairs.
27. It is possible that the variation on many of the dimensions of word embeddings is composed of noise and lacks meaningful semantic information. However, prior studies that have attempted to train embedding models with fewer than 200 dimensions display substantially lower performance in semantic benchmarking tasks (Mikolov, Chen, et al. 2013). Conclusively determining how many dimensions are required to faithfully reproduce a system of cultural associations is beyond our scope, but the evidence we provide suggests it is much greater than three.

References

Accominotti, Fabien, Shamus R. Khan, and Adam Storer. 2018. "How Cultural Capital Emerged in Gilded Age America: Musical Purification and Cross-Class Inclusion at the New York Philharmonic." *American Journal of Sociology* 123(6):1743–83.

Adams, Douglas, S. Brett, and G. Perkings. 1978. *The Hitchhiker's Guide to the Galaxy*. London, UK: BBC Radio 4.

Bartlett's Roget's Thesaurus. 1996. Edited by A. Grometstein, P. B. Hansen, K. W. McManus, R. G. Pustell, S. W. Reinecke, and J. C. Ritchie. Boston, MA: Little, Brown, and Company.

Bearman, Peter S., and Katherine Stovel. 2000. "Becoming a Nazi: A Model for Narrative Networks." *Poetics* 27(2):69–90.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55(4):77–84.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.

Blum, Avrim, John Hopcroft, and Ravindran Kannan. 2016. "Foundations of Data Science." *Vorabversion Eines Lehrbuchs*.

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." Pp. 4349–57 in *Advances in Neural Information Processing Systems*.

Bourdieu, Pierre. 1984. *Distinction: A Social Critique of the Judgement of Taste*. Cambridge, MA: Harvard University Press.

Bourdieu, Pierre. 1989. "Social Space and Symbolic Power." *Sociological Theory* 7(1):14–25.

Bourgois, Philippe. 2003. *In Search of Respect: Selling Crack in El Barrio*. Cambridge, UK: Cambridge University Press.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases." *Science* 356(6334):183–6.

Carley, Kathleen. 1994. "Extracting Culture through Textual Analysis." *Poetics* 22(4):291–312.

Cha, Youngjoo, and Kim A. Weeden. 2014. "Overwork and the Slow Convergence in the Gender Gap in Wages." *American Sociological Review* 79(3):457–84.

Chamberlain, Benjamin Paul, James Clough, and Marc Peter Deisenroth. 2017. "Neural Embeddings of Graphs in Hyperbolic Space" (arXiv:1705.10359).

Chan, Tak Wing, and John H. Goldthorpe. 2004. "Is There a Status Order in Contemporary British Society? Evidence from the Occupational Structure of Friendship." *European Sociological Review* 20(5):383–401.

Chan, Tak Wing, and John H. Goldthorpe. 2007. "Class and Status: The Conceptual Distinction and Its Empirical Relevance." *American Sociological Review* 72(4):512–32.

Clark, Terry N. 2018. *The New Political Culture*. New York: Routledge.

Cohen, Lizabeth. 2003. *A Consumers' Republic: The Politics of Mass Consumption in Postwar America*. New York: Knopf.

Collins, Randall. 1979. *The Credential Society: An Historical Sociology of Education and Stratification*. New York: Academic Press.

Corman, Steven R., Timothy Kuhn, Robert D. McPhee, and Kevin J. Dooley. 2002. "Studying Complex Discursive Systems: Centering Resonance Analysis of Communication." *Human Communication Research* 28(2):157–206.

Crenshaw, Kimberle. 1991. "Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color." *Stanford Law Review* 43(6):1241–99.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" (arXiv:1810.04805).

DiMaggio, Paul. 1982. "Cultural Capital and School Success: The Impact of Status Culture Participation on the Grades of U.S. High School Students." *American Sociological Review* 47(2):189–201.

DiMaggio, Paul. 1997. "Culture and Cognition." *Annual Review of Sociology* 23(1):263–87.

DiMaggio, Paul. 2011. "Cultural Networks." Pp. 286–310 in *Sage Handbook of Social Network Analysis*, edited by J. Scott and P. J. Carrington. Thousand Oaks, CA: Sage Publications (http://dx.doi.org/10.4135/9781446294413.n20).

DiMaggio, Paul, and John Mohr. 1985. "Cultural Capital, Educational Attainment, and Marital Selection." *American Journal of Sociology* 90(6):1231–61.

DiMaggio, Paul, Manish Nag, and David Blei. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41(6):570–606.

Douglas, Mary. 1966. *Purity and Danger: An Analysis of Concepts of Pollution and Taboo*. New York: Routledge.

Dumais, Susan T. 2004. "Latent Semantic Analysis." *Annual Review of Information Science and Technology* 38(1):188–230.

Efron, Bradley. 2003. "Second Thoughts on the Bootstrap." *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 18(2):135–40.

Efron, Bradley, and R. J. Tibshirani. 1994. *An Introduction to the Bootstrap*. London, UK: CRC Press.

Elias, Norbert. 1978. *The History of Manners,* Vol. 1, *The Civilizing Process*. New York: Pantheon.

Emirbayer, Mustafa. 1997. "Manifesto for a Relational Sociology." *American Journal of Sociology* 103(2):281–317.

Evans, James A., and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42:21–50.

Firth, John R. 1957. "A Synopsis of Linguistic Theory, 1930–1955." *Studies in Linguistic Analysis*. Oxford, UK: Blackwell.

Fischer, Claude S., and Michael Hout. 2006. *Century of Difference: How America Changed in the Last One Hundred Years*. New York: Russell Sage Foundation.

Fourcade, Marion. 2011. "Cents and Sensibility: Economic Valuation and the Nature of 'Nature.'" *American Journal of Sociology* 116(6):1721–77.

Fourcade, Marion, and Kieran Healy. 2007. "Moral Views of Market Society." *Annual Review of Sociology* 33(1):285–311.

Franzosi, Roberto. 2004. *From Words to Numbers: Narrative, Data, and Social Science*. Cambridge, UK: Cambridge University Press.

Freeland, Robert E., and Jesse Hoey. 2018. "The Structure of Deference: Modeling Occupational Status Using Affect Control Theory." *American Sociological Review* 83(2):243–77.

Fryer, Roland G., and Steven D. Levitt. 2004. "The Causes and Consequences of Distinctively Black Names." *Quarterly Journal of Economics* 119(3):767–805.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." *Proceedings of the National Academy of Sciences* 115(16):E3635–44.

Gilman, Nils. 1999. "Thorstein Veblen's Neglected Feminism." *Journal of Economic Issues* 33(3):689–711.

Glaser, Barney, and Anselm Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago, IL: Aldine.

Goldin, Claudia, and Lawrence F. Katz. 1999. "Human Capital and Social Capital: The Rise of Secondary Schooling in America, 1910–1940." *Journal of Interdisciplinary History* 29(4):683–723.

Gramsci, Antonio. 1992. *Prison Notebooks*. New York: Columbia University Press.

Greenacre, Michael. 2017. "Ordination with Any Dissimilarity Measure: A Weighted Euclidean Solution." *Ecology* 98(9):2293–300.

Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74(6):1464–80.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change" (arXiv:1605.09096).

Handler, Abram. 2014. "An Empirical Study of Semantic Similarity in WordNet and Word2Vec." PhD dissertation, University of New Orleans, New Orleans, LA.

Heise, David R. 1979. *Understanding Events: Affect and the Construction of Social Action*. Cambridge, UK: Cambridge University Press Archive.

Heise, David R. 1987. "Affect Control Theory: Concepts and Model." *Journal of Mathematical Sociology* 13(1–2):1–33.

Hill, Felix, Kyunghyun Cho, Sebastien Jean, Coline Devin, and Yoshua Bengio. 2014. "Not All Neural Embeddings Are Born Equal" (arXiv:1410.0718).

Hochschild, Arlie Russell. 2012. *The Managed Heart: Commercialization of Human Feeling*. Berkeley: University of California Press.

Hoffman, Mark Anthony, Jean-Philippe Cointet, Philipp Brandt, Newton Key, and Peter Bearman. 2017. "The (Protestant) Bible, the (Printed) Sermon, and the Word(s): The Semantic Structure of the Conformist and Dissenting Bible, 1660–1780." *Poetics* 68:89–103.

Hout, Michael. 2012. "Social and Economic Returns to College Education in the United States." *Annual Review of Sociology* 38:379–400.

Huff, Connor, and Dustin Tingley. 2015. "'Who Are These People?' Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents." *Research & Politics* 2(3)(https://doi.org/10.1177/2053168015604648).

Hunter, James Davison. 1992. *Culture Wars: The Struggle to Control the Family, Art, Education, Law, and Politics in America*. New York: Basic Books.

Illouz, Eva. 1997. *Consuming the Romantic Utopia: Love and the Cultural Contradictions of Capitalism*. Berkeley: University of California Press.

Jakobson, Roman. 1960. "Linguistics and Poetics." Pp. 350–77 in *Style in Language*. Cambridge, MA: MIT Press.

Jenkins, James J., Wallace A. Russell, and George J. Suci. 1958. "An Atlas of Semantic Profiles for 360 Words." *American Journal of Psychology* 71(4):688–99.

Ji, Shihao, Nadathur Satish, Sheng Li, and Pradeep Dubey. 2016. "Parallelizing Word2Vec in Shared and Distributed Memory" (arXiv:1604.04661).

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation." *Transactions of the Association for Computational Linguistics* 5:339–51.

Jost, Jürgen, and Shiping Liu. 2014. "Ollivier's Ricci Curvature, Local Clustering and Curvature-Dimension Inequalities on Graphs." *Discrete & Computational Geometry* 51(2):300–322.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. "Bag of Tricks for Efficient Text Classification" (arXiv:1607.01759).

Kaufer, David S., and Kathleen M. Carley. 1993. "Condensation Symbols: Their Variety and Rhetorical Function in Political Discourse." *Philosophy & Rhetoric* 26(3):201–26.

Khan, Shamus Rahman. 2010. *Privilege: The Making of an Adolescent Elite at St. Paul's School*. Princeton, NJ: Princeton University Press.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. "Statistically Significant Detection of Linguistic Change." Pp. 625–35 in *Proceedings of the 24th International Conference on World Wide Web, WWW '15*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Lamont, Michèle. 1992. *Money, Morals, and Manners: The Culture of the French and the American Upper-Middle Class*. Chicago: University of Chicago Press.

Lamont, Michèle. 2000. *The Dignity of Working Men: Morality and the Boundaries of Race, Class, and Immigration*. Cambridge, MA: Harvard University Press.

Lamont, Michèle, and Annette Lareau. 1988. "Cultural Capital: Allusions, Gaps and Glissandos in Recent Theoretical Developments." *Sociological Theory* 6(2):153–68.

Lareau, Annette, and Elliot B. Weininger. 2003. "Cultural Capital in Educational Research: A Critical Assessment." *Theory and Society* 32(5):567–606.

Le, Quoc, and Tomas Mikolov. 2014. "Distributed Representations of Sentences and Documents." *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China.

Lee, Monica, and John Levi Martin. 2015. "Coding, Counting and Cultural Cartography." *American Journal of Cultural Sociology* 3(1):1–33.

Lerner, Melvin J., and Dale T. Miller. 1978. "Just World Research and the Attribution Process: Looking Back and Ahead." *Psychological Bulletin* 85(5):1030–51.

Lev, Guy, Benjamin Klein, and Lior Wolf. 2015. "In Defense of Word Embedding for Generic Text Representation." Pp. 35–50 in *Natural Language Processing and Information Systems*, edited by C. Biemann, S. Handschuh, A. Freitas, F. Meziane, and E. Métais. New York: Springer International Publishing.

Levay, Kevin E., Jeremy Freese, and James N. Druckman. 2016. "The Demographic and Political Composition of Mechanical Turk Samples." *SAGE Open* (https://doi.org/10.1177/2158244016636433).

Lévi-Strauss, Claude. 1963. *Structural Anthropology*. New York: Basic Books.

Levy, Omer, and Yoav Goldberg. 2014. "Neural Word Embedding as Implicit Matrix Factorization." Pp. 2177–85 in *Advances in Neural Information Processing Systems 27*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. Red Hook, NY: Curran Associates, Inc.

Levy, Omer, Yoav Goldberg, and Ido Dagan. 2015. "Improving Distributional Similarity with Lessons Learned from Word Embeddings." *Transactions of the Association for Computational Linguistics* 3:211–25.

Lin, Yuri, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. 2012. "Syntactic Annotations for the Google Books Ngram Corpus." Pp. 169–74 in *Proceedings of the ACL 2012 System Demonstrations, ACL '12*. Stroudsburg, PA: Association for Computational Linguistics.

Lizardo, Omar. 2017. "Improving Cultural Analysis: Considering Personal Culture in Its Declarative and Nondeclarative Modes." *American Sociological Review* 82(1):88–115.

Marx, Karl. [1867] 2004. *Capital: A Critique of Political Economy*. London, UK: Penguin.

Marx, Karl, and Friedrich Engels. 1970. *The German Ideology*. New York: International Publishers.

McCall, Leslie. 2005. "The Complexity of Intersectionality." *Signs: Journal of Women in Culture and Society* 30(3):1771–800.

Mears, Ashley. 2010. "Size Zero High-End Ethnic: Cultural Production and the Reproduction of Culture in Fashion Modeling." *Poetics* 38(1):21–46.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331(6014):176–82.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space" (arXiv:1301.3781).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." Pp. 3111–9 in *Advances in Neural Information Processing Systems 26*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Red Hook, NY: Curran Associates, Inc.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013. "Linguistic Regularities in Continuous Space Word Representations." *Proceedings of NAACL-HLT 2013*, 746–51.

Miller, George A. 1995. "WordNet: A Lexical Database for English." *Communications of the ACM* 38(11):39–41.

Mische, Ann. 2011. "Relational Sociology, Culture, and Agency." Pp. 80–97 in *Sage Handbook of Social Network Analysis*, edited by J. Scott and P. J. Carrington. Thousand Oaks, CA: Sage Publications.

Mohr, John W., and Petko Bogdanov. 2013. "Introduction—Topic Models: What They Are and Why They Matter." *Poetics* 41(6):545–69.

Mohr, John W., Robin Wagner-Pacifici, and Ronald L. Breiger. 2015. "Toward a Computational Hermeneutics." *Big Data & Society* 2(2)(https://doi.org/10.1177/2053951715613809).

Nagy, William E., Patricia A. Herman, and Richard C. Anderson. 1985. "Learning Words from Context." *Reading Research Quarterly* 20(2):233–53.

Nalisnick, Eric, Bhaskar Mitra, Nick Craswell, and Rich Caruana. 2016. "Improving Document Ranking with Dual Word Embeddings." Pp. 83–84 in *Proceedings of the 25th International Conference Companion on World Wide Web, WWW '16 Companion*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Nickel, Maximillian, and Douwe Kiela. 2017. "Poincaré Embeddings for Learning Hierarchical Representations." Pp. 6338–47 in *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Red Hook, NY: Curran Associates, Inc.

Osgood, Charles E. 1969. "On the Whys and Wherefores of E, P, and A." *Journal of Personality and Social Psychology* 12(3):194–9.

Osgood, Charles Egerton, George J. Suci, and Percy H. Tannenbaum. 1957. *The Measurement of Meaning*. Urbana: University of Illinois Press.

Pachucki, Mark A., and Ronald L. Breiger. 2010. "Cultural Holes: Beyond Relationality in Social Networks and Culture." *Annual Review of Sociology* 36(1):205–24.

Pakulski, Jan, and Malcolm Waters. 1996. *The Death of Class*. London, UK: Sage.

Pechenick, Eitan Adam, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. "Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution." *PloS One* 10(10):e0137041 (https://doi.org/10.1371/journal.pone.0137041).

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "Glove: Global Vectors for Word Representation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–43.

Piketty, Thomas. 2014. "Capital in the Twenty-First Century: A Multidimensional Approach to the History of Capital and Social Classes." *British Journal of Sociology* 65(4):736–47.

Politis, Dimitris N., and Joseph P. Romano. 1992. "A Circular Block-Resampling Procedure for Stationary Data." *Exploring the Limits of Bootstrap* 263–70.

Politis, Dimitris N., and Joseph P. Romano. 1994. "The Stationary Bootstrap." *Journal of the American Statistical Association* 89(428):1303–13.

Politis, Dimitris N., Joseph P. Romano, and Michael Wolf. 1997. *Subsampling*. New York: Springer.

Reay, Diane. 1998. "Rethinking Social Class: Qualitative Perspectives on Class and Gender." *Sociology* 32(2):259–75.

Rei, Marek, and Ted Briscoe. 2014. "Looking for Hyponyms in Vector Space." Pp. 68–77 in *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, Baltimore, MD.

Ricoeur, Paul. 1981. *Hermeneutics and the Human Sciences: Essays on Language, Action and Interpretation*. Cambridge, UK: Cambridge University Press.

Ridgeway, Cecilia L. 2011. *Framed by Gender: How Gender Inequality Persists in the Modern World*. New York: Oxford University Press.

Roget, Peter M. 1912. *Thesaurus of English Words and Phrases*, edited by A. Boyle. New York: E. P. Dutton.

Rosch, Eleanor, and Carolyn B. Mervis. 1975. "Family Resemblances: Studies in the Internal Structure of Categories." *Cognitive Psychology* 7(4):573–605.

Rosvall, Martin, and Carl T. Bergstrom. 2008. "Maps of Random Walks on Complex Networks Reveal Community Structure." *Proceedings of the National Academy of Sciences of the United States of America* 105(4):1118–23.

Rudolph, Maja, Francisco Ruiz, Stephan Mandt, and David Blei. 2016. "Exponential Family Embeddings." Pp. 478–86 in *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Red Hook, NY: Curran Associates, Inc.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2019. *IPUMS USA: Version 9.0*. Minneapolis, MN: IPUMS.

Ruiz, Francisco J. R., Susan Athey, and David M. Blei. 2017. "SHOPPER: A Probabilistic Model of Consumer Choice with Substitutes and Complements" (arXiv:1711.03560).

Salganik, Matthew J. 2017. *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

Salzinger, Leslie. 2003. *Genders in Production: Making Workers in Mexico's Global Factories*. Berkeley: University of California Press.

de Saussure, Ferdinand. 1916. *Course in General Linguistics*. New York: Columbia University Press.

Schröder, Tobias, Jesse Hoey, and Kimberly B. Rogers. 2016. "Modeling Dynamic Identities and Uncertainty in Social Interactions: Bayesian Affect Control Theory." *American Sociological Review* 81(4):828–55.

Searle, John R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge, UK: Cambridge University Press.

Shi, Feng, Jacob G. Foster, and James A. Evans. 2015. "Weaving the Fabric of Science: Dynamic Network Models of Science's Unfolding Structure." *Social Networks* 43(Supplement C):73–85.

Simmel, Georg. [1900] 2004. *The Philosophy of Money*. New York: Routledge.

Skeggs, Beverley. 1997. *Formations of Class & Gender: Becoming Respectable*. London, UK: Sage.

Smith, Charles J. 1903. *Synonyms Discriminated: A Dictionary of Synonymous Words in the English Language*, edited by P. Smith. New York: Henry Holt.

Svallfors, Stefan. 2006. *The Moral Economy of Class: Class and Attitudes in Comparative Perspective*. Stanford, CA: Stanford University Press.

Taddy, Matt. 2015a. "Document Classification by Inversion of Distributed Language Representations" (arXiv:1504.07295).

Taddy, Matt. 2015b. "Document Classification by Inversion of Distributed Language Representations." Pp. 45–49 in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Stroudsburg, PA: Association for Computational Linguistics.

Tversky, Amos, and Itamar Gati. 1978. "Studies of Similarity." *Cognition and Categorization* 1:79–98.

Veblen, Thorstein. [1899] 1912. *The Theory of the Leisure Class: An Economic Study of Institutions*. New York: B. W. Huebsch.

Vilhena, Daril A., Jacob G. Foster, Martin Rosvall, Jevin D. West, James Evans, and Carl T. Bergstrom. 2014. "Finding Cultural Holes: How Structure and Culture Diverge in Networks of Scholarly Communication." *Sociological Science* 1:221–38.

Warner, W. Lloyd, Marchia Meeker, and Kenneth Eells. 1949. *Social Class in America: A Manual of Procedure for the Measurement of Social Status*. Chicago: Science Research Associates.

Weber, Max. 1978. *Economy and Society*. Berkeley: University of California Press.

Webster's Collegiate Thesaurus. 1976. Springfield, MA: Merriam-Webster.

Weeden, Kim A., and David B. Grusky. 2005. "The Case for a New Class Map." *American Journal of Sociology* 111(1):141–212.

Whorf, Benjamin Lee. 1956. *Language, Thought and Reality, Selected Writings of Benjamin Lee Whorf*, edited by J. B. Carroll. Cambridge, MA: MIT Press.

Willis, Paul. 1977. *Learning to Labor: How Working Class Kids Get Working Class Jobs*. New York: University of Columbia Press.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Oxford, UK: Blackwell.

Wright, Erik Olin. 1979. *Class Structure and Income Determination*. New York: Academic Press.

Wright, Erik Olin. 2000. *Class Counts: Student Edition*. Cambridge, UK: Cambridge University Press.

Zelizer, Viviana A. 1979. *Morals and Markets: The Development of Life Insurance in the United States*. New York: Columbia University Press.

Zelizer, Viviana A. 1989. "The Social Meaning of Money: 'Special Monies.'" *American Journal of Sociology* 95(2):342–77.

Zhou, Guangyou, Tingting He, Jun Zhao, and Po Hu. 2015. "Learning Continuous Word Embedding with Metadata for Question Retrieval in Community Question Answering." Pp. 250–9 in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Zipf, George Kingsley. 1932. *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA: Harvard University Press.

Austin C. Kozlowski is a doctoral student at the University of Chicago, Department of Sociology. His research applies a diverse set of methodological tools to the analysis of contemporary American culture.

Matt Taddy is Vice President for Economic Technology and Chief Economist at Amazon. He previously was professor of economics and statistics at the University of Chicago Booth School of Business. His research focuses on statistics, machine learning, and their application to a wide range of problems and large-scale data in economics, social science, and business.

James A. Evans is Professor of sociology at the University of Chicago, Department of Sociology, and External Professor at the Santa Fe Institute. His research uses large-scale data, machine learning, and generative models to understand how collectives think, what they know, and what they create. He is especially interested in innovation and the emergence of ideas, shared patterns of reasoning, and processes of attention, communication, agreement, and certainty in science, technology, society, and culture.

Chapter 6

Garg et al.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018) "Word embeddings quantify 100 years of gender and ethnic stereotypes" PNAS 115:16.

Word embeddings quantify 100 years of gender and ethnic stereotypes

Nikhil Garg[a,1], Londa Schiebinger[b], Dan Jurafsky[c,d], and James Zou[e,f,1]

[a]Department of Electrical Engineering, Stanford University, Stanford, CA 94305; [b]Department of History, Stanford University, Stanford, CA 94305; [c]Department of Linguistics, Stanford University, Stanford, CA 94305; [d]Department of Computer Science, Stanford University, Stanford, CA 94305; [e]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305; and [f]Chan Zuckerberg Biohub, San Francisco, CA 94158

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 12, 2018 (received for review November 22, 2017)

Word embeddings are a powerful machine-learning framework that represents each English word by a vector. The geometric relationship between these vectors captures meaningful semantic relationships between the corresponding words. In this paper, we develop a framework to demonstrate how the temporal dynamics of the embedding helps to quantify changes in stereotypes and attitudes toward women and ethnic minorities in the 20th and 21st centuries in the United States. We integrate word embeddings trained on 100 y of text data with the US Census to show that changes in the embedding track closely with demographic and occupation shifts over time. The embedding captures societal shifts—e.g., the women's movement in the 1960s and Asian immigration into the United States—and also illuminates how specific adjectives and occupations became more closely associated with certain populations over time. Our framework for temporal analysis of word embedding opens up a fruitful intersection between machine learning and quantitative social science.

word embedding | gender stereotypes | ethnic stereotypes

The study of gender and ethnic stereotypes is an important topic across many disciplines. Language analysis is a standard tool used to discover, understand, and demonstrate such stereotypes (1–5). Previous literature broadly establishes that language both reflects and perpetuates cultural stereotypes. However, such studies primarily leverage human surveys (6–16), dictionary and qualitative analysis (17), or in-depth knowledge of different languages (18). These methods often require time-consuming and expensive manual analysis and may not easily scale across types of stereotypes, time periods, and languages. In this paper, we propose using word embeddings, a commonly used tool in natural language processing (NLP) and machine learning, as a framework to measure, quantify, and compare beliefs over time. As a specific case study, we apply this tool to study the temporal dynamics of gender and ethnic stereotypes in the 20th and 21st centuries in the United States.

In word-embedding models, each word in a given language is assigned to a high-dimensional vector such that the geometry of the vectors captures semantic relations between the words—e.g., vectors being closer together has been shown to correspond to more similar words (19). These models are typically trained automatically on large corpora of text, such as collections of Google News articles or Wikipedia, and are known to capture relationships not found through simple co-occurrence analysis. For example, the vector for France is close to vectors for Austria and Italy, and the vector for XBox is close to that of PlayStation (19). Beyond nearby neighbors, embeddings can also capture more global relationships between words. The difference between London and England—obtained by simply subtracting these two vectors—is parallel to the vector difference between Paris and France. This pattern allows embeddings to capture analogy relationships, such as London to England is as Paris to France.

Recent works demonstrate that word embeddings, among other methods in machine learning, capture common stereotypes because these stereotypes are likely to be present, even if subtly, in the large corpora of training texts (20–23). For example, the vector for the adjective honorable would be close to the vector for man, whereas the vector for submissive would be closer to woman. These stereotypes are automatically learned by the embedding algorithm and could be problematic if the embedding is then used for sensitive applications such as search rankings, product recommendations, or translations. An important direction of research is to develop algorithms to debias the word embeddings (20).

In this paper, we take another approach. We use the word embeddings as a quantitative lens through which to study historical trends—specifically trends in the gender and ethnic stereotypes in the 20th and 21st centuries in the United States. We develop a systematic framework and metrics to analyze word embeddings trained over 100 y of text corpora. We show that temporal dynamics of the word embedding capture changes in gender and ethnic stereotypes over time. In particular, we quantify how specific biases decrease over time while other stereotypes increase. Moreover, dynamics of the embedding strongly correlate with quantifiable changes in US society, such as demographic and occupation shifts. For example, major transitions in the word embedding geometry reveal changes in the descriptions of genders and ethnic groups during the women's movement in the 1960s–1970s and Asian-American population growth in the 1960s and 1980s. We validate our findings on external metrics and show that our results are robust to the different algorithms for training the word embeddings. Our framework reveals and quantifies how stereotypes toward women and ethnic groups have evolved in the United States.

Significance

Word embeddings are a popular machine-learning method that represents each English word by a vector, such that the geometry between these vectors captures semantic relations between the corresponding words. We demonstrate that word embeddings can be used as a powerful tool to quantify historical trends and social change. As specific applications, we develop metrics based on word embeddings to characterize how gender stereotypes and attitudes toward ethnic minorities in the United States evolved during the 20th and 21st centuries starting from 1910. Our framework opens up a fruitful intersection between machine learning and quantitative social science.

Author contributions: N.G., L.S., D.J., and J.Z. designed research; N.G. and J.Z. performed research; and N.G. and J.Z. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Data and code related to this paper are available on GitHub (https://github.com/nikhgarg/EmbeddingDynamicStereotypes).

[1]To whom correspondence may be addressed. Email: nkgarg@stanford.edu or jamesz@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1720347115/-/DCSupplemental.

Published online April 3, 2018.

Our results demonstrate that word embeddings are a powerful lens through which we can systematically quantify common stereotypes and other historical trends. Embeddings thus provide an important quantitative metric which complements existing, more qualitative, linguistic and sociological analyses of biases. In *Embedding Framework Overview and Validations*, we validate that embeddings accurately capture sociological trends by comparing associations in the embeddings with census and other externally verifiable data. In *Quantifying Gender Stereotypes* and *Quantifying Ethnic Stereotypes* we apply the framework to quantify the change in stereotypes of women, men, and ethnic minorities. We further discuss our findings in *Discussion* and provide additional details in *Materials and Methods*.

Embedding Framework Overview and Validations

In this section, we briefly describe our methods and data and then validate our findings. We focus on showing that word embeddings are an effective tool to study historical biases and stereotypes by relating measurements from these embeddings to historical census and survey data. The consistent replication of such historical data, both in magnitude and in direction of biases, validates the use of embeddings in such work. This section extends the analysis of refs. 20 and 21 in showing that embeddings can also be used as a comparative tool over time as a consistent metric for various biases.

Summary of Data and Methods. We now briefly describe our datasets and methods, leaving details to *Materials and Methods* and *SI Appendix*, section A. All of our code and embeddings are available publicly*. For contemporary snapshot analysis, we use the standard Google News word2vec vectors trained on the Google News dataset (24, 25). For historical temporal analysis, we use previously trained Google Books/Corpus of Historical American English (COHA) embeddings, which are a set of nine embeddings, each trained on a decade in the 1900s, using the COHA and Google Books (26). As additional validation, we train, using the GLoVe algorithm (27), embeddings from the *New York Times* Annotated Corpus (28) for every year between 1988 and 2005. We then collate several word lists to represent each gender† (men, women) and ethnicity‡ (White, Asian, and Hispanic), as well as neutral words (adjectives and occupations). For occupations, we use historical US census data (29) to extract the percentage of workers in each occupation that belong to each gender or ethnic group and compare it to the bias in the embeddings.

Using the embeddings and word lists, one can measure the strength of association (embedding bias) between neutral words and a group. As an example, we overview the steps we use to quantify the occupational embedding bias for women. We first compute the average embedding distance between words that represent women—e.g., she, female—and words for occupations—e.g., teacher, lawyer. For comparison, we also compute the average embedding distance between words that represent men and the same occupation words. A natural metric for the embedding bias is the average distance for women minus the average distance for men. If this value is negative, then the embedding more closely associates the occupations with men. More generally, we compute the representative group vector by taking the average of the vectors for each word in the given gender/ethnicity group. Then we compute the average Euclidean distance between each representative group vector and each vector in the neutral word list of interest, which could be occupations or adjectives. The difference of the average distances is our metric for bias—we call this the relative norm difference or simply embedding bias.

We use ordinary least-squares regressions to measure associations in our analysis. In this paper, we report r^2 and the coefficient P value for each regression, along with the intercept confidence interval when relevant.

Validation of the Embedding Bias. To verify that the bias in the embedding accurately reflects sociological trends, we compare the trends in the embeddings with quantifiable demographic trends in the occupation participation, as well as historical surveys of stereotypes. First, we use women and minority ethnic participation statistics (relative to men and Whites, respectively) in different occupations as a benchmark because it is an objective metric of social changes. We show that the embedding accurately captures both gender and ethnic occupation percentages and consistently reflects historical changes.

Next, we validate that the embeddings capture personality trait stereotypes. A difficulty in social science is the relative dearth of historical data to systematically quantify stereotypes, which highlights the value of our embedding framework as a quantitative tool but also makes it challenging to directly confirm our findings on adjectives. Nevertheless, we make use of the best available data from historical surveys, gender stereotypes from 1977 and 1990 (6, 7) and ethnic stereotypes from the Princeton trilogy from 1933, 1951, and 1969 (8–10).

Comparison with women's occupation participation. We investigate how the gender bias of occupations in the word embeddings relates to the empirical percentage of women in each of these occupations in the United States. Fig. 1 shows, for each occupation, the relationship between the relative percentage (of women) in the occupation in 2015 and the relative norm distance between words associated with women and men in the Google News embeddings. (Occupations whose 2015 percentage is not available, such as midwife, are omitted. We further note that the Google News embedding is trained on a corpus

Fig. 1. Women's occupation relative percentage vs. embedding bias in Google News vectors. More positive indicates more associated with women on both axes. $P < 10^{-10}$, $r^2 = 0.499$. The shaded region is the 95% bootstrapped confidence interval of the regression line. In this single embedding, then, the association in the embedding effectively captures the percentage of women in an occupation.

*All of our own data and analysis tools are available on GitHub at https://github.com/nikhgarg/EmbeddingDynamicStereotypes. Census data are available through the Integrated Public Use Microdata Series (29). We link to the sources for each embedding used in *Materials and Methods*.

†There is an increasingly recognized difference between sex and gender and thus between the words male/female and man/woman, as well as nonbinary categories. We limit our analysis to the two major binary categories due to technical limitations, and we use male and female as part of the lists of words associated with men and women, respectively, when measuring gender associations. We also use results from refs. 6 and 7 which study stereotypes associated with sex.

‡When we refer to Whites or Asians, we specifically mean the non-Hispanic subpopulation. For each ethnicity, we generate a list of common last names among the group. Unfortunately, our present methods do not extend to Blacks due to large overlaps in common last names among Whites and Blacks in the United States.

over time, and so the 2015 occupations are not an exact comparison.) The relative distance in the embeddings significantly correlates with the occupation percentage ($P < 10^{-10}$, $r^2 = 0.499$). It is interesting to note that the regression line nearly intersects the origin [intercept in $(-0.021, -0.002)$]: Occupations that are close to 50–50 in gender participation have small embedding bias. These results suggest that the embedding bias correctly matches the magnitude of the occupation frequency, along with which gender is more common in the occupation.

We ask whether the relationship between embedding and occupation percentage holds true for specific occupations. We perform the same embedding bias vs. occupation frequency analysis on a subset of occupations that are deemed "professional" (e.g., nurse, engineer, judge; full list in *SI Appendix*, section A.3) and find nearly identical correlation [$P < 10^{-5}$, $r^2 = 0.595$, intercept in $(-0.026, 0)$]. We further validate this association using different embeddings trained on Wikipedia and Common Crawl texts instead of Google News; see *SI Appendix*, section B.1 for details.

The Google News embedding reveals one aggregate snapshot of the bias since it is trained over a pool of news articles. We next analyze the embedding of each decade of COHA from 1910 to 1990 separately to validate that for a given historical period, the embedding bias from data in that period accurately reflects occupation participation. For each decade, the embedding gender bias is significantly correlated with occupation frequency ($P \leq 0.003$, $r^2 \geq 0.123$), as in the case with the Google News embedding; however, we note that the intercepts here show a consistent additional bias against women for each decade; i.e., even occupations with the same number of men and women are closer to words associated with men.

More importantly, these correlations are very similar over the decades, suggesting that the relationship between embedding bias score and "reality," as measured by occupation participation, is consistent over time. We measure this consistency in several ways. We first train a single model for all (occupation percentage, embedding bias) pairs across time. We compare this model to a model where there is an additional term for each year and show that the models perform similarly ($r^2 = 0.236$ vs. $r^2 = 0.298$). Next, we compare the performance of the model without terms for each year to models trained separately for each year, showing that the single model both has similar parameters and performance to such separate models. Finally, for each embedding year, we compare performance of the model trained for that embedding vs. a model trained using all other data (leave-one-out validation). We repeat the entire analysis with embeddings trained using another algorithm on the same dataset [singular value decomposition (SVD)]. See *SI Appendix*, section B.3.1 for details.

This consistency makes the interpretation of embedding bias more reliable; i.e., a given bias score corresponds to approximately the same percentage of the workforce in that occupation being women, regardless of the embedding decade.

Next, we ask whether the changes in embeddings over decades capture changes in the women's occupation participation. Fig. 2 shows the average embedding bias over the occupations over time, overlaid with the average women's occupation relative percentage over time. [We include only occupations for which census data are available for every decade and which are frequent enough in all embeddings. We use the linear regression mapping inferred from all of the data across decades to align the scales for the embedding bias and occupation frequency (the two y axes in the plot).] The average bias closely tracks with the occupation percentages over time. The average bias is negative, meaning that occupations are more closely associated with men than with women. However, we see that the bias steadily moves closer to 0 from the 1950s to the 1990s, suggesting that the bias

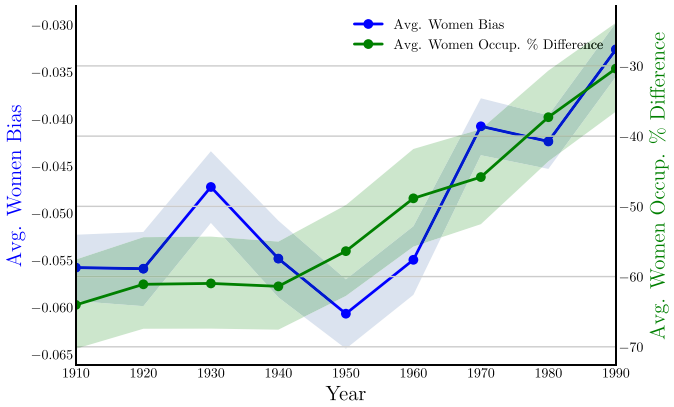


Fig. 2. Average gender bias score over time in COHA embeddings in occupations vs. the average percentage of difference. More positive means a stronger association with women. In blue is relative bias toward women in the embeddings, and in green is the average percentage of difference of women in the same occupations. Each shaded region is the bootstrap SE interval.

is decreasing. This trend tracks with the proportional increase in women's participation in these occupations.

Comparison with ethnic occupation participation. Next, we compare ethnic bias in the embeddings to occupation participation rates and stereotypes. As in the case with gender, the embeddings capture externally validated ethnic bias. Table 1 shows the 10 occupations that are the most biased toward Hispanic, Asian, and White last names[§]. The Asian-American "model minority" (30, 31) stereotype appears predominantly; academic positions such as professor, scientist, and physicist all appear among the top Asian-biased occupations. Similarly, White and Hispanic stereotypes also appear in their respective lists. [Smith, besides being an occupation, is a common White-American last name. It is thus excluded from regressions, as are occupations such as conductor, which have multiple meanings (train conductors as well as music conductors).] As in the case with gender, the embedding bias scores are significantly correlated with the ethnic group's relative percentage of the occupation as measured by the US Census in 2010. For Hispanics, the bias score is a significant predictor of occupation percentage at $P < 10^{-5}$, $r^2 = 0.279$ and, for Asians, at $P = 0.041$, $r^2 = 0.065$. Due to the large population discrepancy between Whites and each respective minority group, the intercept values for these plots are large and are difficult to interpret and so are excluded from the main exposition (see *Discussion* for further details). The corresponding scatter plots and regression tables of embedding bias vs. occupation relative percentage are in *SI Appendix*, section C.1.

Similarly, as for gender, we track the occupation bias score over time and compare it to the occupation relative percentages; Fig. 3 does so for Asian Americans, relative to Whites, in the COHA embeddings. The increase in occupation relative percentage across all occupations is well tracked by the bias in the embeddings. More detail and a similar plot with Hispanic Americans are included in *SI Appendix*, section C.3.

Comparison with surveys of gender stereotypes. Now, we validate that the historical embeddings also capture gender stereotypes of personality traits. We leverage sex stereotype scores assigned to a set of 230 adjectives (300 adjectives are in the original studies; 70 adjectives are discarded due to low frequencies

[§]We adapt the relative norm distance in Eq. **3** for three groups. For each group, we compare its norm bias with the average bias of the other groups; i.e., bias(group 1) $= \sum_w \left[\frac{1}{2}(\|w - v_2\| + \|w - v_3\|) - \|w - v_1\|\right]$. This method can lead to the same occupation being highly ranked for multiple groups, such as happens for mason.

in the COHA embeddings) by human participants (6, 7). Participants scored each word for its association with men or women (example words: headstrong, quarrelsome, effeminate, fickle, talkative). This human subject study was first performed in 1977 and then repeated in 1990. We compute the correlation between the adjective embedding biases in COHA 1970s and 1990s with the respective decade human-assigned scores. In each case, the embedding bias score is significantly correlated with the human-assigned scores [$P < 0.0002, r^2 \geq 0.095$, intercepts in $(-0.017, -0.012)$ and $(-0.029, -0.024)$, respectively). *SI Appendix*, section B.3 contains details of the analysis. These analyses suggest that the embedding gender bias effectively captures both occupation frequencies as well as human stereotypes of adjectives, although noisily.

Comparison with surveys of ethnic stereotypes. We validate that the embeddings capture historical personality stereotypes toward ethnic groups. We leverage data from the well-known Princeton trilogy experiments (8–10), published in 1933, 1951, and 1969, respectively. These experiments have sparked significant discussion, follow-up work, and methodological criticism (11–16), but they remain our best method to validate our quantification of historical stereotypes.

These works surveyed stereotypes among Princeton undergraduates toward 10 ethnic groups, including Chinese people. (Other groups include Germans, Japanese, and Italians. We focus on Chinese stereotypes due to the ability to distinguish last names and a sufficient quantity of data in the embeddings.) Katz and Braly in 1933 reported the top 15 stereotypes attached to each group from a larger list of words (8) (example stereotypes: industrious, superstitious, nationalistic). (Each stereotype score is the percentage of respondents who indicated that the stereotype applies to the group. Note that these scores are not comparative across groups; i.e., a stereotype's score for one group does not directly imply its score for any other group, and so the regression intercepts are not meaningful.) In 1969, Karlins et al. reported scores for the same 15 stereotypes, among others (10). Scores for a subset of these adjectives were also reported in 1951 (9).

Using the stereotypes of Chinese people and our list of Chinese last names, we conduct two tests: First, using all reported scores for which there is sufficient text data, we correlate the stereotype scores with the given stereotype's embedding bias in the corresponding decade; second, using the stereotypes for which both 1933 and 1969 scores are available, we correlate the change in the scores with the change in the embedding bias during the period.

The results suggest, as in the case with gender, that adjective stereotypes in the embeddings reflect attitudes of the times and that the embeddings are calibrated across time. In our first test, the studies' stereotype scores are significant predictors of the corresponding embedding biases ($r^2 = 0.146,\ P = 0.023$).

Table 1. The top 10 occupations most closely associated with each ethnic group in the Google News embedding

Hispanic	Asian	White
Housekeeper	Professor	Smith
Mason	Official	Blacksmith
Artist	Secretary	Surveyor
Janitor	Conductor	Sheriff
Dancer	Physicist	Weaver
Mechanic	Scientist	Administrator
Photographer	Chemist	Mason
Baker	Tailor	Statistician
Cashier	Accountant	Clergy
Driver	Engineer	Photographer

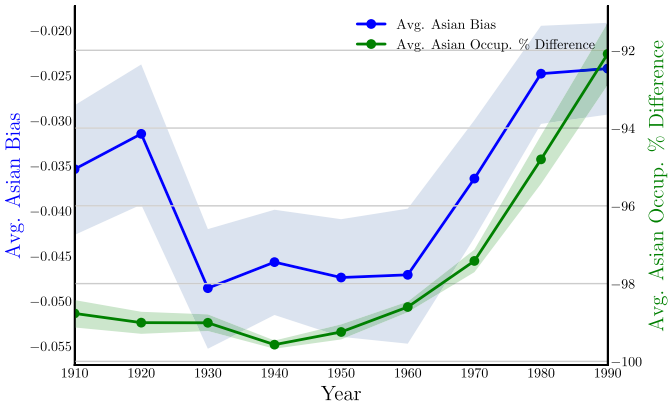


Fig. 3. Average ethnic (Asian vs. White) bias score over time for occupations in COHA (blue) vs. the average percentage of difference (green). Each shaded region is the bootstrap SE interval.

In the second test, the changes in the scores are also significant predictors of the changes in embedding biases ($r^2 = 0.472$, $P = 0.014$). See *SI Appendix*, section C.2 for regression tables and plots.

Together, the analyses in this section validate that embeddings capture historical attitudes toward both ethnic and gender groups, as well as changes in these attitudes. In the remainder of this work, we use this insight to explore such historical stereotypes to display the power of this framework.

Quantifying Gender Stereotypes

We now apply our framework to study trends in gender bias in society, both historically and in modern times. We first show that language today, such as that in the Google News corpora, is even more biased than could be accounted for by occupation data. In addition, we show that bias, as seen through adjectives associated with men and women, has decreased over time and that the women's movement in the 1960s and 1970s especially had a systemic and drastic effect in women's portrayals in literature and culture.

Due to the relative lack of systematic quantification of stereotypes in the literature, a gap that this work seeks to address, we cannot directly validate the results in this section or the next. We reference sociological literature and use statistical tests as appropriate to support the analyses.

Occupational Stereotypes Beyond Census Data. While women's occupation percentages are highly correlated with embedding gender bias, we hypothesize that the embedding could reflect additional social stereotypes beyond what can be explained by occupation participation. To test this hypothesis, we leverage the gender stereotype scores of occupations, as labeled by people on Amazon Mechanical Turk and provided to us by the authors of ref. 20¶. These crowdsource scores reflect aggregate human judgment as to whether an occupation is stereotypically associated with men or women. (A caveat here is that the US-based participants on Amazon Mechanical Turk may not represent the US population.) In separate regressions, both the crowdsourced stereotype scores [$r^2 = 0.655, P < 10^{-10}$, intercept confidence interval $(-0.281, 0.027)$] and the occupation relative percentage [$r^2 = 0.452, P < 10^{-6}$, intercept confidence

¶ List of occupations available is in *SI Appendix*, section A.3. Note that the crowdsourcing experiment collected data for a larger list of occupations; we select the occupations for which both census data and embedding orientation are also available. For this reason, the regressions with just the occupation percentage score are slightly different from those in Fig. 1.

interval $(-0.027, -0.001)$] are significantly correlated with the embedding bias.

Next, we conduct two additional separate regressions to test that the embedding bias captures the same extra stereotype information as do the crowdsource scores, information that is missing in the census data. In each regression, the occupation percentage difference is the independent covariate. In one, the embedding bias is the dependent variable; in the other, stereotype score is. In these regressions, a negative (positive) residual indicates that the embedding bias or stereotype score is closer to words associated with women (men) than is to be expected given the gender percentages in the occupation. We find that the residuals between the two regressions correlate significantly (Pearson coefficient 0.811, $P < 10^{-10}$). This correlation suggests that the embedding bias captures the crowdsource human stereotypes beyond that which can be explained by empirical differences in occupation proportions.

Where such crowdsourcing is not possible, such as in studying historical biases, word embeddings can thus further serve as an effective measurement tool. Further, although the analysis in the previous section shows a strong relationship between census data and embedding bias, it is important to note that biases beyond census data also appear in the embedding.

Quantifying Changing Attitudes with Adjective Embeddings. We now apply the insight that embeddings can be used to make comparative statements over time to study how the description of women—through adjectives—in literature and the broader culture has changed over time. Using word embeddings to analyze biases in adjectives could be an especially useful approach because the literature is lacking systematic and quantitative metrics for adjective biases. We find that—as a whole—portrayals have changed dramatically over time, including for the better in some measurable ways. Furthermore, we find evidence for how the women's movement in the 1960s and 1970s led to a systemic change in such portrayals.

How overall portrayals change over time. We first establish that comparing the embeddings over time could reveal global shifts in society in regard to gender portrayals. Fig. 4 shows the Pearson correlation in embedding bias scores for adjectives over time between COHA embeddings for each pair of decades. As expected, the highest correlation values are near the diagonals; embeddings (and attitudes) are most similar to those from adjacent decades. More strikingly, the matrix exhibits two clear blocks. There is a sharp divide between the 1960s and 1970s, the height of the women's movement in the United States, during which there was a large push to end legal and social barriers for women in education and the workplace (32, 33). The transition in the gender embeddings from 1960 to 1970 is statistically significant ($P < 10^{-4}$, Kolmogorov–Smirnov two-sample test) and is larger than the change between any two other adjacent decades. See *SI Appendix*, section B.3.3 for a more detailed description of the test and all statistics.

Fig. 4. Pearson correlation in embedding bias scores for adjectives over time between embeddings for each decade. The phase shift in the 1960s–1970s corresponds to the US women's movement.

We note that the effects of the women's movement, including on inclusive language, are well documented (18, 33–36); this work provides a quantitative way to measure the rate and extent of the change. A potential extension and application of this work would be to study how various narratives and descriptions of women developed and competed over time.

Individual words whose biases changed over time. As an example of such work, we consider a subset of the adjectives describing competence, such as intelligent, logical, and thoughtful (see *SI Appendix*, section A.3 for a full list of words; these words were curated from various online sources). Since the 1960s, this group of words on average has increased in association with women over time (from strongly biased toward men to less so): In a regression with embedding bias from each word as the dependent variable and years from 1960 to 1990 as the covariate, the coefficient is positive; i.e., there is a (small) positive trend (0.005 increase in women association per decade, $P = 0.0036$). At this rate, such adjectives would be equally associated with women as with men a little after the year 2020.

As a comparison, we also analyze a subset of adjectives describing physical appearance—e.g., attractive, ugly, and fashionable—and the bias of these words did not change significantly since the 1960s (null hypothesis of no trend not rejected with $P > 0.2$). Although the trend regarding intelligence is encouraging, the top adjectives are still potentially problematic, as displayed in Table 2.

We note that this analysis is an exploration; perceived competence and physical appearance are just two components of gender stereotypes. Models in the literature suggest that stereotypes form along several dimensions, e.g., warmth and competence (16). A more complete analysis would first collect externally validated lists of words that describe each such dimension and then measure the embedding association with respect to these lists over time.

The embedding also reveals interesting patterns in how individual words evolve over time in their gender association. For example, the word hysterical used to be, until the mid-1900s, a catchall term for diagnosing mental illness in women but has since become a more general word (37); such changes are clearly reflected in the embeddings, as hysterical fell from a top 5 woman-biased word in 1920 to not in the top 100 in 1990 in the COHA embeddings[#]. On the other hand, the word emotional becomes much more strongly associated with women over time in the embeddings, reflecting its current status as a word that is largely associated with women in a pejorative sense (38).

These results together demonstrate the value and potential of leveraging embeddings to study biases over time. The embeddings capture subtle individual changes in association, as well as larger historical changes. Overall, they paint a picture of a society with decreasing but still significant gender biases.

Quantifying Ethnic Stereotypes

We now turn our attention to studying ethnic biases over time. In particular we show how immigration and other 20th-century trends broadly influenced how Asians were viewed in the United States. We also show that embeddings can serve as effective tools to analyze finer-grained trends by analyzing the portrayal of Islam in the *New York Times* from 1988 to 2005 in the context of terrorism.

[#]We caution that due to the noisy nature of word embeddings, dwelling on individual word rankings in isolation is potentially problematic. For example, hysterical is more highly associated with women in the Google News vectors than emotional. For this reason we focus on large shifts between embeddings.

Table 2. Top adjectives associated with women in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

Trends in Asian Stereotypes. To study Asian stereotypes in the embeddings, we use common and distinctly Asian last names, identified through a process described in *SI Appendix*, section A.2. This process results in a list of 20 last names that are primarily but not exclusively Chinese last names.

The embeddings illustrate a dramatic story of how Asian-American stereotypes developed and changed in the 20th century. Fig. 5 shows the Pearson correlation coefficient of adjective biases for each pair of embeddings over time. As with gender, the analysis shows how external events changed attitudes. There are two phase shifts in the correlation: in the 1960s, which coincide with a sharp increase in Asian immigration into the United States due to the passage of the 1965 Immigration and Nationality Act, and in the 1980s, when immigration continued and the second-generation Asian-American population emerged (39). Using the same Kolmogorov–Smirnov test on the correlation differences described in the previous section, the phase shifts between the 1950s–1960s ($P = 0.011$) and 1970s–1980s ($P < 10^{-3}$) are significant, while the rest are not ($P > 0.070$).

We extract the most biased adjectives toward Asians (when compared with Whites) to gain more insights into factors driving these global changes in the embedding. Table 3 shows the most Asian-biased adjectives in 1910, 1950, and 1990. Before 1950, strongly negative words, especially those often used to describe outsiders, are among the words most associated with Asians: barbaric, hateful, monstrous, bizarre, and cruel. However, starting around 1950 and especially by 1980, with a rising Asian population in the United States, these words are largely replaced by words often considered stereotypic (40–42) of Asian Americans today: sensitive, passive, complacent, active, and hearty, for example. See *SI Appendix*, Table C.8 for the complete list of the top 10 most Asian-associated words in each decade.

Using our methods regarding trends, we can quantify this change more precisely: Fig. 6 shows the relative strength of the Asian association for words used to describe outsiders over time. As opposed to the adjectives overall, which see two distinct phase shifts in Asian association, the words related to outsiders steadily decrease in Asian association over time—except around World War II—indicating that broader globalization trends led to changing attitudes with regard to such negative portrayals. Overall, the word embeddings exhibit a remarkable change in adjectives and attitudes toward Asian Americans during the 20th century.

Trends in Other Ethnic and Cultural Stereotypes. Similar trends appear in other datasets as well. Fig. 7 shows, in the *New York Times* over two decades, how words related to Islam (vs. those related to Christianity) associate with terrorism-related words. Similar to how we measure occupation-related bias, we create a list of words associated with terrorism, such as terror, bomb, and violence. We then measure how associated these words appear to be in the text to words representing each religion, such as mosque and church, for Islam and Christianity, respectively. (Full word lists are available in *SI Appendix*, section A.) Throughout the time period in the *New York Times*, Islam is more associated with terrorism than is Christianity. Furthermore, an increase in the association can be seen both after the 1993 World Trade Center bombings and after September 11, 2001. With a more recent dataset and using more news outlets, it would be useful to study how such attitudes have evolved since 2005.

We illustrate how word embeddings capture stereotypes toward other ethnic groups. For example, *SI Appendix*, Fig. C.4, with Russian names, shows a dramatic shift in the 1950s, the start of the Cold War, and a minor shift during the initial years of the Russian Revolution in the 1910s–1920s. Furthermore, *SI Appendix*, Fig. C.5, the correlation over time plot with Hispanic names, serves as an effective control group. It shows more steady changes in the embeddings rather than the sharp transitions found in Asian and Russian associations. This pattern is consistent with the fact that numerous events throughout the 20th century influenced the story of Hispanic immigration into the United States, with no single event playing too large a role (43).

These patterns demonstrate the usefulness of our methods to study ethnic as well as gender bias over time; similar analyses can be performed to examine shifts in the attitudes toward other ethnic groups, especially around significant global events. In particular, it would be interesting to more closely measure dehumanization and "othering" of immigrants and other groups using a suite of linguistic techniques, validating and extending the patterns discovered in this work.

Discussion

In this work, we investigate how the geometry of word embeddings, with respect to gender and ethnic stereotypes, evolves over time and tracks with empirical demographic changes in the United States. We apply our methods to analyze word embeddings trained over 100 y of text data. In particular, we quantify the embedding biases for occupations and adjectives. Using occupations allows us to validate the method when the embedding associations are compared with empirical participation rates for each occupation. We show that both gender and ethnic occupation biases in the embeddings significantly track with the actual occupation participation rates. We also show that adjective associations in the embeddings provide insight into how different groups of people are viewed over time.

As in any empirical work, the robustness of our results depends on the data sources and the metrics we choose to represent bias or association. We choose the relative norm difference metric for its simplicity, although many such metrics are reasonable. Refs. 20 and 21 leverage alternate metrics, for example. Our metric agrees with other possible metrics—both

Fig. 5. Pearson correlation in embedding Asian bias scores for adjectives over time between embeddings for each decade.

Table 3. Top Asian (vs. White) adjectives in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

1910	1950	1990
Irresponsible	Disorganized	Inhibited
Envious	Outrageous	Passive
Barbaric	Pompous	Dissolute
Aggressive	Unstable	Haughty
Transparent	Effeminate	Complacent
Monstrous	Unprincipled	Forceful
Hateful	Venomous	Fixed
Cruel	Disobedient	Active
Greedy	Predatory	Sensitive
Bizarre	Boisterous	Hearty

qualitatively through the results in the snapshot analysis for gender, which replicates prior work, and quantitatively as the metrics correlate highly with one another, as shown in *SI Appendix*, section A.5.

Furthermore, we primarily use linear models to fit the relationship between embedding bias and various external metrics; however, the true relationships may be nonlinear and warrant further study. This concern is especially salient when studying ethnic stereotypes over time in the United States, as immigration drastically shifts the size of each group as a percentage of the population, which may interact with stereotypes and occupation percentages. However, the models are sufficient to show consistency in the relationships between embedding bias and external metrics across datasets over time. Further, the results do not qualitatively change when, for example, population logit proportion instead of raw percentage difference is used, as in ref. 44; we reproduce our primary figures with such a transformation in *SI Appendix*, section A.6.

Another potential concern may be the dependency of our results on the specific word lists used and that the recall of our methods in capturing human biases may not be adequate. We take extensive care to reproduce similar results with other word lists and types of measurements to demonstrate recall. For example, in *SI Appendix*, section B.1, we repeat the static occupation analysis using only professional occupations and reproduce an identical figure to Fig. 1 in *SI Appendix*, section B.1. Furthermore, the plots themselves contain bootstrapped confidence intervals; i.e., the coefficients for random subsets of the occupations/adjectives and the intervals are tight. Similarly, for adjectives, we use two different lists: one list from refs. 6 and 7 for which we have labeled stereotype scores and then a larger one for the rest of the analysis where such scores are not needed. We note that we do not tune either the embeddings or the word lists, instead opting for the largest/most general publicly available data. For reproducibility, we share our code and all word lists in a repository. That our methods replicate across many different embeddings and types of biases measured suggests their generalizability.

A common challenge in historical analysis is that the written text in, say 1910, may not completely reflect the popular social attitude of that time. This is an important caveat to consider in interpreting the results of the embeddings trained on these earlier text corpora. The fact that the embedding bias for gender and ethnic groups does track with census proportion is a positive control that the embedding is still capturing meaningful patterns despite possible limitations in the training text. Even this control may be limited in that the census proportion does not fully capture gender or ethnic associations, even in the present day. However, the written text does serve as a window into the attitudes of the day as expressed in popular culture, and this work allows for a more systematic study of such text.

Another limitation of our current approach is that all of the embeddings used are fully "black box," where the dimensions have no inherent meaning. To provide a more causal explanation of how the stereotypes appear in language, and to understand how they function, future work can leverage more recent embedding models in which certain dimensions are designed to capture various aspects of language, such as the polarity of a word or its parts of speech (45). Similarly, structural properties of words—beyond their census information or human-rated stereotypes—can be studied in the context of these dimensions. One can also leverage recent Bayesian embeddings models and train more fine-grained embeddings over time, rather than a separate embedding per decade as done in this work (46, 47). These approaches can be used in future work.

We view the main contribution of our work as introducing and validating a framework for exploring the temporal dynamics of stereotypes through the lens of word embeddings. Our framework enables the computation of simple but quantitative measures of bias as well as easy visualizations. It is important to note that our goal in *Quantifying Gender Stereotypes* and *Quantifying Ethnic Stereotypes* is quantitative exploratory analysis rather than pinning down specific causal models of how certain stereotypes arise or develop, although the analysis in *Occupational Stereotypes Beyond Census Data* suggests that common language is more biased than one would expect based on external, objective metrics. We believe our approach sharpens the analysis of large cultural shifts in US history; e.g., the women's movement of the 1960s correlates with a sharp shift in the encoding matrix (Fig. 4) as well as changes in the biases associated with specific occupations and gender-biased adjectives (e.g., hysterical vs. emotional).

In standard quantitative social science, machine learning is used as a tool to analyze data. Our work shows how the artifacts of machine learning (word embeddings here) can themselves be interesting objects of sociological analysis. We believe this paradigm shift can lead to many fruitful studies.

Materials and Methods

In this section we describe the datasets, embeddings, and word lists used, as well as how bias is quantified. More detail, including descriptions of additional embeddings and the full word lists, are in *SI Appendix*, section A. All of our data and code are available on GitHub (https://github.com/nikhgarg/EmbeddingDynamicStereotypes), and we link to external data sources as appropriate.

Embeddings. This work uses several pretrained word embeddings publicly available online; refer to the respective sources for in-depth discussion of their training parameters. These embeddings are among the most commonly used English embeddings, vary in the datasets on which they were

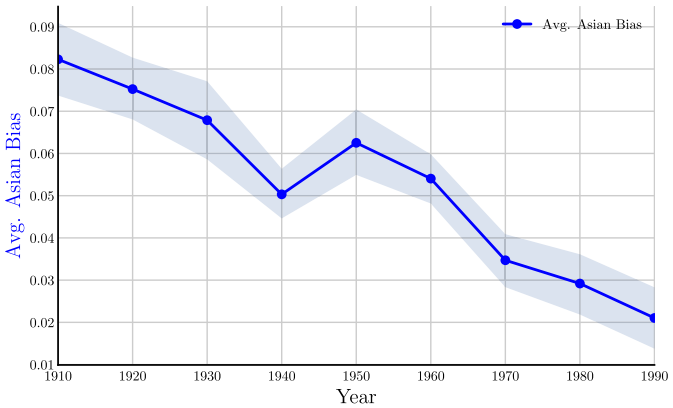


Fig. 6. Asian bias score over time for words related to outsiders in COHA data. The shaded region is the bootstrap SE interval.

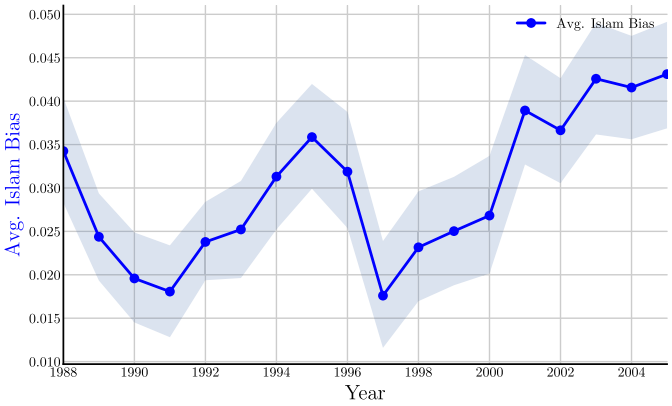


Fig. 7. Religious (Islam vs. Christianity) bias score over time for words related to terrorism in *New York Times* data. Note that embeddings are trained in 3-y windows, so, for example, 2000 contains data from 1999–2001. The shaded region is the bootstrap SE interval.

trained, and between them cover the best-known algorithms to construct embeddings. One finding in this work is that, although there is some heterogeneity, gender and ethnic bias is generally consistent across embeddings. Here we restrict descriptions to embeddings used in the main exposition. For consistency, only single words are used, all vectors are normalized by their l_2 norm, and words are converted to lowercase.

Google News word2vec vectors. Vectors trained on about 100 billion words in the Google News dataset (24, 25). Vectors are available at https://code.google.com/archive/p/word2vec/.

Google Books/COHA. Vectors trained on a combined corpus of genre-balanced Google Books and the COHA (48) by the authors of ref. 26. For each decade, a separate embedding is trained from the corpus data corresponding to that decade. The dataset is specifically designed to enable comparisons across decades, and the creators take special care to avoid selection bias issues. The vectors are available at https://nlp.stanford.edu/projects/histwords/, and we limit our analysis to the SVD and skip-gram with negative sampling (SGNS) (also known as word2vec) embeddings in the 1900s. Note that the Google Books data may include some non-American sources and the external metrics we use are American. However, this does not appreciably affect results. In the main text, we exclusively use SGNS embeddings; results with SVD embeddings are in *SI Appendix* and are qualitatively similar to the SGNS results. Unless otherwise specified, COHA indicates these embeddings trained using the SGNS algorithm.

New York Times. We train embeddings over time from *The New York Times* Annotated Corpus (28), using 1.8 million articles from the *New York Times* between 1988 and 2005. We use the GLoVe algorithm (27) and train embeddings over 3-y windows (so the 2000 embeddings, for example, contain articles from 1999 to 2001).

In *SI Appendix* we also use other embeddings available at https://nlp.stanford.edu/projects/glove/.

Related Work. Word embedding was developed as a framework to represent words as a part of the artificial intelligence and natural language processing pipeline (25). Ref. 20 demonstrated that word embeddings capture gender stereotypes, and ref. 21 additionally verified that the embedding accurately reflects human biases by comparing the embedding results with that of the implicit association test. While these two papers analyzed the bias of the static Google News embedding, our paper investigates the temporal changes in word embeddings and studies how embeddings over time capture historical trends. Our paper also studies attitudes toward women and ethnic minorities by quantifying the embedding of adjectives. The focus of ref. 20 is to develop algorithms to reduce the gender stereotype in the embedding, which is important for sensitive applications of embeddings. In contrast, our aim is not to debias, but to leverage the embedding bias to study historical changes that are otherwise challenging to quantify. Ref. 21 shows that embeddings contain each of the associations commonly found in the implicit association test. For example, European-American names are more similar to pleasant (vs. unpleasant) words than are African-American names, and male names are more similar to career (vs. family) words than are female names. Similarly, they show that, in the Google News embeddings, census data correspond to bias in the embeddings for gender.

The study of gender and ethnic stereotypes is a large focus of linguistics and sociology and is too extensive to be surveyed here (1–5). Our main innovation is the use of word embeddings, which provides a unique lens to measure and quantify biases. Another related field in linguistics studies how language changes over time and has also recently used word embeddings as a tool (49–51). However, this literature primarily studies semantic changes, such as how the word gay used to primarily mean cheerful and now means predominantly means homosexual (26, 52), and does not investigate bias.

Word Lists and External Metrics. Two types of word lists are used in this work: group words and neutral words. Group words represent groups of people, such as each gender and ethnicity. Neutral words are those that are not intrinsically gendered or ethnic (for example, fireman or mailman would be gendered occupation titles and so are excluded); relative similarities between neutral words and a pair of groups (such as men vs. women) are used to measure the strength of the association in the embeddings. In this work, we use occupations and various adjective lists as neutral words.

Gender. For gender, we use noun and pronoun pairs (such as he/she, him/her, etc.).

Race/ethnicity. To distinguish various ethnicities, we leverage the fact that the distribution of last names in the United States differs significantly by ethnicity, with the notable exception of White and Black last names. Starting with a breakdown of ethnicity by last name compiled by ref. 53, we identify 20 last names for each ethnicity as detailed in *SI Appendix*, section A.2. Our procedure, however, produces almost identical lists for White and Black Americans (with the names being mostly White by percentage), and so the analysis does not include Black Americans.

Occupation census data. We use occupation words for which we have gender and ethnic subgroup information over time. Group occupation percentages are obtained from the Integrated Public Use Microdata Series (IPUMS), part of the University of Minnesota Historical Census Project (29). Data coding and preprocessing are done as described in ref. 44, which studies wage dynamics as women enter certain occupations over time. The IPUMS dataset includes a column, OCC1950, coding occupation census data as it would have been coded in 1950, allowing accurate interyear analysis. We then hand map the occupations from this column to single-word occupations (e.g., chemical engineer and electrical engineer both become engineer, and chemist is counted as both chemist and scientist) and hand code a subset of the occupations as professional. In all plots containing occupation percentages for gender, we use the percentage difference between women and men in the occupation:

$$p_{\text{women}} - p_{\text{men}}$$
$$\text{where } p_{\text{women}} = \%\text{ of occupation that is women}$$
$$p_{\text{men}} = \%\text{ of occupation that is men.}$$

For ethnicity, we similarly report the percentage difference, except we first condition on the workers being in one of the groups in question:

$$\frac{p_{\text{min}} - p_{\text{white}}}{p_{\text{min}} + p_{\text{white}}}$$
$$\text{where } p_{\text{min}} = \%\text{ of occupation that is minority group in question}$$
$$p_{\text{white}} = \%\text{ of occupation that is White.}$$

In each case, a value of 0 indicates an equal number of each group in the occupation. We note that the results do not qualitatively change if instead the logit proportion (or conditional logit proportion) of the minority group is used, as in ref. 44 (*SI Appendix*, section A.6).

Occupation gender stereotypes. For a limited set of occupations, we use gender stereotype scores collected from users on Amazon Mechanical Turk by ref. 20. These scores are compared with embedding gender association.

Adjectives. To study associations with adjectives over time, several separate lists are used. To compare gender adjective embedding bias to external metrics, we leverage a list of adjectives labeled by how stereotypically associated with men or women they are, as determined by a group of subjects in 1977 and 1990 (6, 7). For Chinese adjective embedding bias, we use a list of stereotypes from the Princeton trilogy (8–10). For all other analyses using adjectives, a larger list of adjectives is used, primarily from ref. 54. Except when otherwise specified, adjectives are used to refer to this larger list.

Metrics. Given two vectors, their similarity can be measured either by their negative difference norm, as in Eq. **1**, or by their cosine similarity, as in Eq. **2**. The denominators are omitted because all vectors have norm 1:

$$\text{neg-norm-dif}(u, v) = -\|u - v\|_2 \quad [1]$$

$$\text{cos-sim}(u, v) = u \cdot v. \quad [2]$$

The association between the group words and neutral words is calculated as follows: Construct a group vector by averaging the vectors for each word in the group; then calculate the similarity between this average vector and each word in the neutral list as above.

The relative norm distance, which captures the relative strength of association of a set of neutral words with respect to two groups, is as described in Eq. **3**, where M is the set of neutral word vectors, v_1 is the average vector for group one, and v_2 is the average vector for group two. The more positive (negative) that the relative norm distance is, the more associated the neutral words are toward group two (one). In this work, when we say that a word is biased toward a group with respect to another group, we specifically mean in the context of the relative norm distance. Bias score also refers to this metric:

$$\text{relative norm distance} = \sum_{v_m \in M} \|v_m - v_1\|_2 - \|v_m - v_2\|_2. \quad [3]$$

We can also use cosine similarity rather than the Euclidean 2-norm. *SI Appendix*, section A.5 shows that the choice of similarity measure is not important; the respective metrics using each similarity measure correlate highly with one another (Pearson coefficient > 0.95 in most cases). In the main text, we exclusively use the relative norm.

ACKNOWLEDGMENTS. J.Z. is supported by a Chan–Zuckerberg Biohub Investigator grant and National Science Foundation (NSF) Grant CRII 1657155. N.G. is supported by the NSF Graduate Research Fellowship under Grant DGE-114747.

1. Hamilton DL, Trolier TK (1986) *Stereotypes and Stereotyping: An Overview of the Cognitive Approach in Prejudice, Discrimination, and Racism* (Academic, San Diego), pp 127–163.
2. Basow SA (1992) *Gender: Stereotypes and Roles* (Thomson Brooks/Cole Publishing Co, Belmont, CA), 3rd Ed.
3. Wetherell M, Potter J (1992) *Mapping the Language of Racism: Discourse and the Legitimation of Exploitation* (Columbia Univ Press, New York).
4. Holmes J, Meyerhoff M, eds (2004) *The Handbook of Language and Gender* (Blackwell Publishing Ltd, Oxford).
5. Coates J (2016) *Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language* (Routledge, London).
6. Williams JE, Best DL (1977) Sex stereotypes and trait favorability on the adjective check list. *Educ Psychol Meas* 37:101–110.
7. Williams JE, Best DL (1990) *Measuring Sex Stereotypes: A Multination Study* (Sage Publications, Thousand Oaks, CA), Rev Ed.
8. Katz D, Braly K (1933) Racial stereotypes of one hundred college students. *J Abnorm Soc Psychol* 28:280–290.
9. Gilbert GM (1951) Stereotype persistence and change among college students. *J Abnorm Soc Psychol* 46:245–254.
10. Karlins M, Coffman TL, Walters G (1969) On the fading of social stereotypes: Studies in three generations of college students. *J Pers Soc Psychol* 13:1–16.
11. Devine PG, Elliot AJ (1995) Are racial stereotypes really fading? The Princeton trilogy revisited. *Pers Soc Psychol Bull* 21:1139–1150.
12. Diekman AB, Eagly AH (2000) Stereotypes as dynamic constructs: Women and men of the past, present, and future. *Pers Soc Psychol Bull* 26:1171–1188.
13. Bergsieker HB, Leslie LM, Constantine VS, Fiske ST (2012) Stereotyping by omission: Eliminate the negative, accentuate the positive. *J Pers Soc Psychol* 102:1214–1238.
14. Madon S, et al. (2001) Ethnic and national stereotypes: The Princeton trilogy revisited and revised. *Pers Soc Psychol Bull* 27:996–1010.
15. Gaertner SL, Dovidio JF (1986) *The Aversive Form of Racism* (Academic, San Diego).
16. Fiske ST, Cuddy AJC, Glick P, Xu J (2002) A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *J Pers Soc Psychol* 82:878–902.
17. Henley NM (1989) Molehill or mountain? What we know and don't know about sex bias in language. *Gender and Thought: Psychological Perspectives*, eds Crawford M, Gentry M (Springer, New York), pp 59–78.
18. Hellinger M, Bußmann H eds (2001) *Gender Across Languages: The Linguistic Representation of Women and Men,* IMPACT: Studies in Language and Society (John Benjamins Publishing Company, Amsterdam), Vol 9.
19. Collobert R, et al. (2011) Natural language processing (almost) from scratch. *J Machine Learn Res* 12:2493–2537.
20. Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT (2016) Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems 29,* eds Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R (Curran Associates, Inc, Barcelona), pp 4349–4357.
21. Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* 356:183–186.
22. Zhao J, Wang T, Yatskar M, Ordonez V, Chang KW (2017) Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, eds Palmer M, Hwa R (Association for Computational Linguistics, Copenhagen), pp 2979–2989.
23. van Miltenburg E (2016) Stereotyping and bias in the Flickr30k dataset. arXiv: 1605.06083.
24. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781.
25. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, eds Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (Curran Associates, Inc, Lake Tahoe, NV), pp 3111–3119.
26. Hamilton WL, Leskovec J, Jurafsky D (2016) Diachronic word embeddings reveal statistical laws of semantic change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds Erk K, Smith NA (Association for Computational Linguistics, Berlin), Vol 1, pp 1489–1501.
27. Pennington J, Socher R, Manning C (2014) Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), eds Moschitti A, Pang B (Association for Computational Linguistics, Doha, Qatar), pp 1532–1543.
28. Sandhaus E (2008) *The New York Times Annotated Corpus* (Linguistic Data Consortium, Philadelphia).
29. Ruggles S, Genadek K, Goeken R, Grover J, Sobek M (2015) Integrated Public Use Microdata Series: Version 6.0. Available at doi.org/10.18128/D010.V6.0. Accessed August 16, 2017.
30. Osajima K (2005) Asian Americans as the model minority: An analysis of the popular press image in the 1960s and 1980s. *A Companion to Asian American Studies*, ed Ono KA (Blackwell Publishing Ltd, Malden, MA), pp 215–225.
31. Fong TP (2002) *The Contemporary Asian American Experience: Beyond the Model Minority* (Prentice Hall, Upper Saddle River, NJ).
32. Bryson V (2016) *Feminist Political Theory* (Palgrave Macmillan, New York).
33. Rosen R (2013) *The World Split Open: How the Modern Women's Movement Changed America* (Tantor eBooks, Old Saybrook, CT).
34. Thorne B, Henley N, Kramarae C, eds (1983) *Language, Gender, and Society* (Newbury House, Rowley, MA).
35. Eckert P, McConnell-Ginet S (2003) *Language and Gender* (Cambridge Univ Press, Cambridge, UK).
36. Evans S (2010) *Tidal Wave: How Women Changed America at Century's End* (Simon and Schuster, New York).
37. Tasca C, Rapetti M, Carta MG, Fadda B (2012) Women and hysteria in the history of mental health. *Clin Pract Epidemiol Ment Health* 8:110–119.
38. Sanghani R (2016) Feisty, frigid and frumpy: 25 Words we only use to describe women. The Telegraph. Available at https://www.telegraph.co.uk/women/life/ambitious-frigid-and-frumpy-25-words-we-only-use-to-describe-wom/. Accessed August 21, 2017.
39. Zong J, Batalova J (2016) *Asian Immigrants in the United States* (Migration Policy Institute, Washington, DC).
40. Lee SJ (1994) Behind the model-minority stereotype: Voices of high- and low-achieving Asian American students. *Anthropol Educ Q* 25:413–429.
41. Kim A, Yeh CJ (2002) *Stereotypes of Asian American students* (ERIC Digest New York, NY).
42. Lee SJ (2015) *Unraveling the "Model Minority" Stereotype: Listening to Asian American Youth* (Teachers College Press, New York), 2nd Ed.
43. Gutiérrez DG (2016) A historic overview of Latino immigration and the demographic transformation of the United States. *The New Latino Studies Reader: A Twenty-First-Century Perspective*, eds Gutierrez RA, Almaguer T (Univ of California Press, Oakland, CA), pp 108–125.
44. Levanon A, England P, Allison P (2009) Occupational feminization and pay: Assessing causal dynamics using 1950–2000 U.S. Census data. *Soc Forces* 88:865–891.
45. Rothe S, Schütze H (2016) Word embedding calculus in meaningful ultradense subspaces. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, eds Erk K, Smith NA (Association for Computational Linguistics, Berlin), Vol 2, pp 512–517.
46. Rudolph M, Ruiz F, Athey S, Blei D (2017) *Structured Embedding Models for Grouped Data in Advances in Neural Information Processing Systems 30*, eds Guyon I, et al. (Curran Associates, Inc, Long Beach, CA), pp 250–260.
47. Rudolph M, Blei D (2017) Dynamic Bernoulli embeddings for language evolution. arXiv:1703.08052.
48. Davies M (2010) The 400 million word corpus of historical American English (1810–2009). *Selected Papers from the Sixteenth International Conference on English Historical Linguistics (ICEHL 16), Pécs, 23–27 August 2010*, eds Hegedűs I, Fodor A (John Benjamins Publishing, Amsterdam), Vol 325.
49. Ullmann S (1962) *Semantics: An Introduction to the Science of Meaning* (Barnes & Noble, New York).
50. Blank A (1999) *Why Do New Meanings Occur? A Cognitive Typology of the Motivations for Lexical Semantic Change in Historical Semantics and Cognition*, ed Koch P (Walter de Gruyter, New York).

51. Kulkarni V, Al-Rfou R, Perozzi B, Skiena S (2015) Statistically significant detection of linguistic change. *Proceedings of the 24th International Conference on World Wide Web*, eds Gangemi A, Leonardi S, Panconesi A (International World Wide Web Conferences Steering Committee, New York), pp 625–635.
52. Hamilton WL, Leskovec J, Jurafsky D (2016) Cultural shift or linguistic drift? comparing two computational measures of semantic change. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, eds Su J, Duh K, Carreras X (Association for Computational Linguistics, Austin, TX), pp 2116–2121.
53. Chalabi M, Flowers A (2018) Dear Mona, what's the most common name in America? Available at https://fivethirtyeight.com/features/whats-the-most-common-name-in-america/. Accessed September 3, 2017.
54. Gunkel P (2013) 638 Primary personality traits. Available at ideonomy.mit.edu/essays/traits.html. Accessed August 21, 2017.