2. Feature representations, Vector Semantics

NLP for CogSci Research

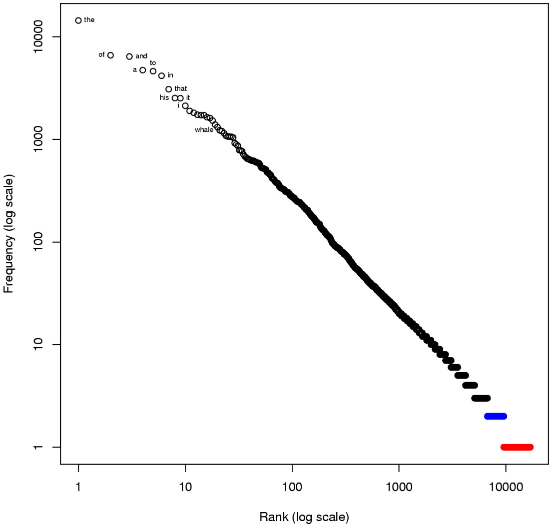
Marlene Staib

September 19, 2018

Recap:Language Data

- Sequences
- Ambiguity
- Sparsity: Zipf's Law

Recap: Zipf's Law



Words (corpus: Moby Dick - source: Wikipedia)

Recap: Sparsity

- $|V| =$ large! (e.g.,1M)
- Distribution over $\{w_1, w_2, ..., w_n\} \in$ V: most counts will be 0 (smoothing assigns non-0 prob.)

More problems...

- No relationship between words; independence
  - But: *dog - poodle; good - bad; King - Queen; good - terrific*; ...
- Paraphrases/synonyms?
  - *start the task* vs. *begin the challenge*
- No way of combining/compounding words/phrases
  - But: house boat $\approx$ house + boat ?

More problems...

Language modelling:

- "The poodle ____ at the man."
  - "poodle" is unobserved/very infrequent in training
  - "dog" is observed many times
  - can the use of "dog" tell us something about the use of "poodle"?

Semantics

Semantics


LOOK SIMBA, THIS IS LINGUISTICS. YOU MAY STUDY ANY PART OF THE LAND.
WHAT ABOUT THAT DARK SHADOWY PLACE?
THAT IS SEMANTICS. YOU MUST NEVER GO THERE.
zipmeme

Semantics

An old philosophers' joke:

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

Word Relationships

- Synonyms: *couch - sofa*
- Antonyms: *hot - cold*
- Hypernym: *dog - poodle*
- Hyponym: *poodle - dog*
- Word similarity: *car - train*
- Semantic frame: *buyer, seller, goods, money*
- Semantic field: *surgery, scalpel, doctor, nurse*

WordNet

- 3 separate databases: nouns, verbs, adjectives
- Each word has multiple senses: *bass*^{1}, *bass*^{2},...
- "Synsets": synonym sets (of specific senses)
- Links between synsets that establish hypernymy, hyponymy, antonymy, ...
- Problem: not all words/languages; costly!

Vector Semantics

John R. Firth (1957):

"You shall know a word by the company it keeps."

Ludwig Wittgenstein:

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

J. R. Firth

Knowing a word by its use

- A bottle of tesgüino is on the table.
- Everybody likes tesgüino .
- Tesgüino makes you drunk.
- We make tesgüino out of corn.

(from J&M; example modified by Lin (1998a) from Nida (1975))

Word Vectors

document	text
d1	Horses are strong and mighty animals.
d2	Unicorns are mighty and fabulous animals.
d3	Lady Gaga is a fabulous pop star.
d4	The sun is a hot, burning star.

Term-document Matrix

	d1	d2	d3	d4
horses	1	0	0	0
unicorns	0	1	0	0
are	1	1	0	0
is	0	0	1	1
and	1	1	0	0
mighty	1	1	0	0
strong	1	0	0	0
fabulous	0	1	1	0
animals	1	1	0	0
star	0	0	1	1
...	...	...	...	...

Term-term Matrix

	horses	unicorns	are	is	and	mighty	...
horses	1	0	1	0	1	1	...
unicorns	0	1	1	0	1	1	...
are	1	1	2	0	2	2	...
is	0	0	0	2	0	0	...
and	1	1	0	0	2	2	...
mighty	1	1	2	0	0	0	...
strong	1	0	0	0	0	0	...
fabulous	0	1	1	1	1	1	...
animals	1	1	0	0	0	0	...
star	0	0	0	1	0	0	...
...	...	...	...	...	...	...	...

First/second order co-occurrence

- First-order/syntagmatic: *e.g., write - book*
- Second-order/paradigmatic: *e.g., dog - poodle*

tf-idf

tf_i: term frequency of term i (original term-document matrix)

idf_i: inverse document frequency of term i

$$idf_i = log(\frac{N}{df_i})$$

N...total number of documents

df_i...number of documents in which term i occurs

tf-idf

	d1	d2	d3	d4
horses	0.6	0	0	0
unicorns	0	0.6	0	0
are	0.3	0.3	0	0
is	0	0	0.3	0.3
...	...	...	...	...

horses in d1:

$$tf \cdot idf = 1 \cdot log(\frac{4}{1}) = 0.602$$

are in d1/d2:

$$tf \cdot idf = 1 \cdot log(\frac{4}{2}) = 0.301$$

(P)PMI

Pointwise mutual information, between word w and context word c:

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

PMI $= 0 \rightarrow P(w,c) =$ chance

PMI $> 1 \rightarrow P(w,c) =$ greater than chance

PMI $< 1 \rightarrow P(w,c) =$ less than chance (*but not significant*)

PPMI ... positive PMI $\rightarrow$ set all negative values to 0

PPMI

	horses	unicorns	are	is	and	mighty	...
horses	2	0	1	0	1	1	...
unicorns	0	2	1	0	1	1	...
are	1	1	1	0	1	1	...
is	0	0	0	1	0	0	...
...	...	...	...	...	...	...	...

horses/horses:

$$PMI = log_2(\frac{0.25}{0.25 \cdot 0.25}) = 2$$

horses/are:

$$PMI = log_2(\frac{0.25}{0.25 \cdot 0.5}) = 1$$

are/and:

$$PMI = log_2(\frac{0.5}{0.5 \cdot 0.5}) = 1$$

Have we solved our problems?

- No relationship between words; independence ✓
- No way of combining/compounding words/phrases (?)
- Sparsity ✗

Dense Vectors: LSA

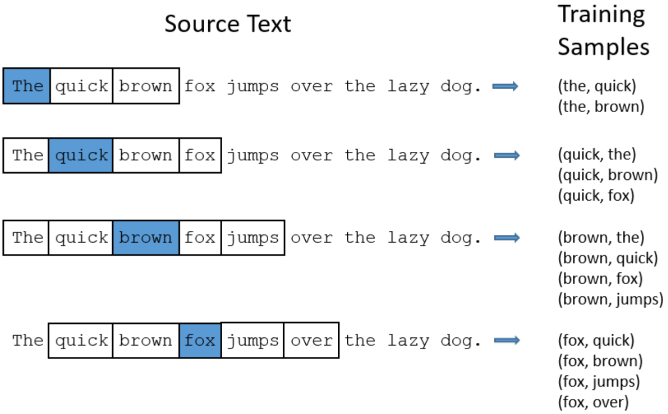
Latent Semantic Analysis: reduce term-term matrix using Singular Value Decomposition (SVD $\sim$ Factor Analysis, Principal Component Analysis):

$$\underbrace{\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}}_{|V| \times c} = \underbrace{\begin{bmatrix} & & \\ & W & \\ & & \end{bmatrix}}_{|V| \times m} \underbrace{\begin{bmatrix} \sigma_1 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_m \end{bmatrix}}_{m \times m} \underbrace{\begin{bmatrix} & & \\ & C & \\ & & \end{bmatrix}}_{m \times c}$$

(from J&M, a previous draft)
see also this amazing stackoverflow post (first answer)

Dense Vectors: Embedding Spaces

Word2Vec: Learning dense vectors with a Neural Net



(from http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/)

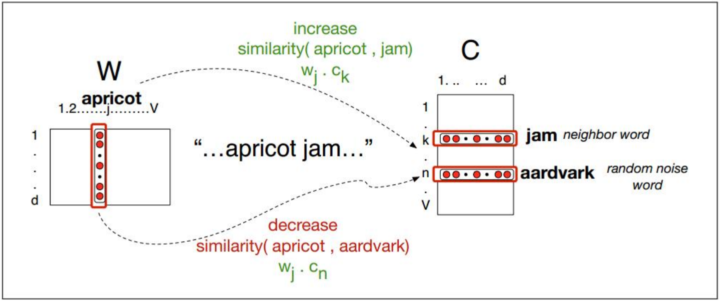
Dense Vectors: Embedding Spaces

Word2Vec: Learning dense vectors with a Neural Net

inputs	class
fox, quick	+
fox, brown	+
fox, jumps	+
fox, over	+
pineapple, helicopter	-
door, malignant	-
..., ...	...

Dense Vectors: Embedding Spaces

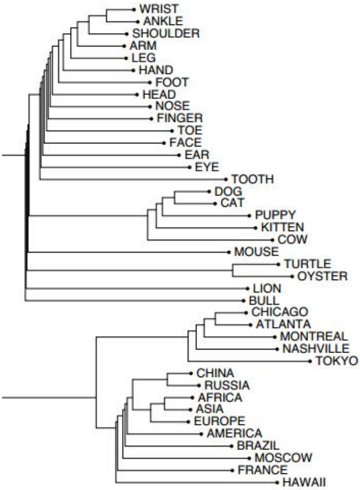
Word2Vec: Learning dense vectors with a Neural Net



(from J & M, ed. 3, chapter 6)

Measuring Word Similarity

Word Similarity - Clustering

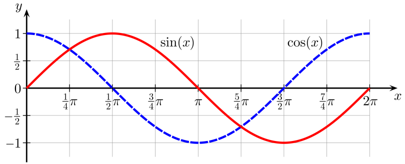


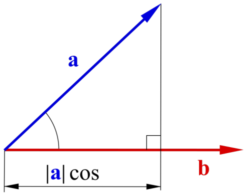
Dot product

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix} \circ \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = a_1 b_1 + a_2 b_2 + a_3 b_3 + a_4 b_4 + a_5 b_5$$

Cosine Similarity

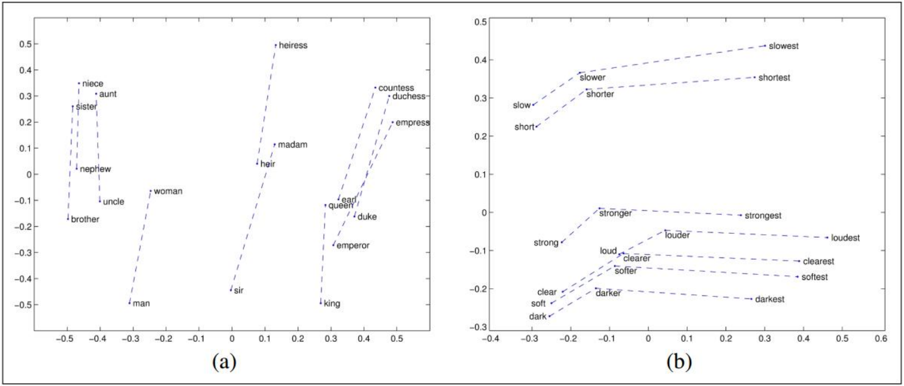
$$cos(\theta) = \frac{a \circ b}{|a|\ |b|}$$





Analogies - Maths with Words

Analogies - Projections



Analogies

Fill in the blank:

- Paris is to France as Rome is to ____
- Paris - France $+$ Rome $\approx$ Italy

Also:

- King - man $+$ woman $\approx$ Queen
- bananas - banana $+$ apple $\approx$ apples

Evaluating Word Vectors

Evaluating Similarity

- SimLex-999
- WordSim-353
- Human similarity ratings for pairs of words

WordSim-353

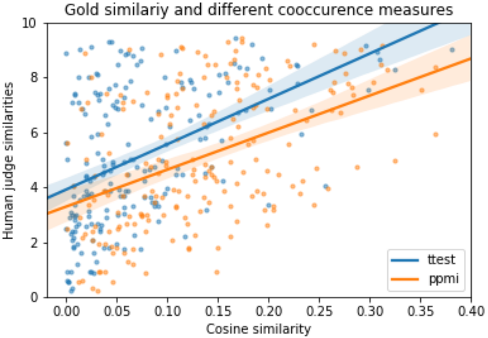
tiger	tiger
tiger	cat
plane	car
train	car
television	radio
cucumber	potato
football	soccer
football	basketball
physics	chemistry
vodka	gin
lad	wizard
sugar	approach
professor	cucumber
king	cabbage

https://goo.gl/DTgqvA

WordSim-353

		score
tiger	tiger	10.00
tiger	cat	7.35
plane	car	5.77
train	car	6.31
television	radio	6.77
cucumber	potato	5.92
football	soccer	9.03
football	basketball	6.81
physics	chemistry	7.35
vodka	gin	8.46
lad	wizard	0.92
sugar	approach	0.88
professor	cucumber	0.31
king	cabbage	0.23

WordSim-353: Correlation



Evaluating Analogies

- e.g., SemEval-2012:
  https://sites.google.com/site/semeval2012task2/download
- set of word pairs, scored on how much they share a certain *relation*

SemEval-2012

Sample questions:

Question 1: Consider the following word pairs: pilgrim:shrine, hunter:quarry, assassin:victim, climber:peak. What relation best describes these X:Y word pairs?

1. "X worships/reveres Y"
2. "X seeks/desires/aims for Y"
3. "X harms/destroys Y"
4. "X uses/exploits/employs Y"

SemEval-2012

Question 2: Consider the following word pairs: pilgrim:shrine, hunter:quarry, assassin:victim, climber:peak. These X:Y pairs share a relation, "X R Y". Now consider the following word pairs:

1. pig:mud
2. politician:votes
3. dog:bone
4. bird:worm

Which of the above numbered word pairs is the MOST illustrative example of the same relation "X R Y"?

Which of the above numbered word pairs is the LEAST illustrative example of the same relation "X R Y"?

SemEval-2012

score	
44.0	harvesting:farming
40.0	stitching:sewing
38.0	chewing:eating
32.0	stirring:cooking
30.0	mixing:baking
26.0	shampooing:bathing
12.0	talking:therapy
4.0	frying:cooking
-8.0	hearing:understanding
-10.0	bumping:volleyball
-14.0	volleying:tennis
-14.0	talking:speech
-16.0	cooking:eating

Intrinsic versus extrinsic evaluation

- **Intrinsic**: Some metric within the task itself
  - e.g., comparison to human labels created for the task
  - e.g., Mean Squared Error of a regression model
- **Extrinsic**: Using an external/downstream task
  - e.g., plugging word vectors into a language model/Naive Bayes Classifier,....

Bias in Word Vectors

Caliskan et al. 2017, *Science*

"bias refers generally to prior information, a necessary prerequisite for intelligent action. Yet bias can be problematic where such information is derived from aspects of human culture known to lead to harmful behavior."

Bias in Word Embeddings

- doctor – man $+$ woman $\approx$ nurse
- computer programmer – man $+$ woman $\approx$ homemaker

Bias in Word Embeddings

Pleasant words

caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation

Unpleasant words

abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison

(Caliskan et al. 2017, *Science*)

Bias in Word Embeddings

Findings (Caliskan et al. 2017, *Science*):

- Flowers associate with pleasant, insects with unpleasant
- Musical instruments associate with pleasant, weapons to unpleasant
- European American names associate more with pleasant, compared to African American
- Female words associate more with family than career words
- Female words associate more with arts than mathematics