# Vector Semantics and Embeddings

Kenneth Lai

Brandeis University

November 9, 2022

# Announcements

- By 11:59pm tonight
  - Final Project Idea due
- For 11/16
  - HW4 due

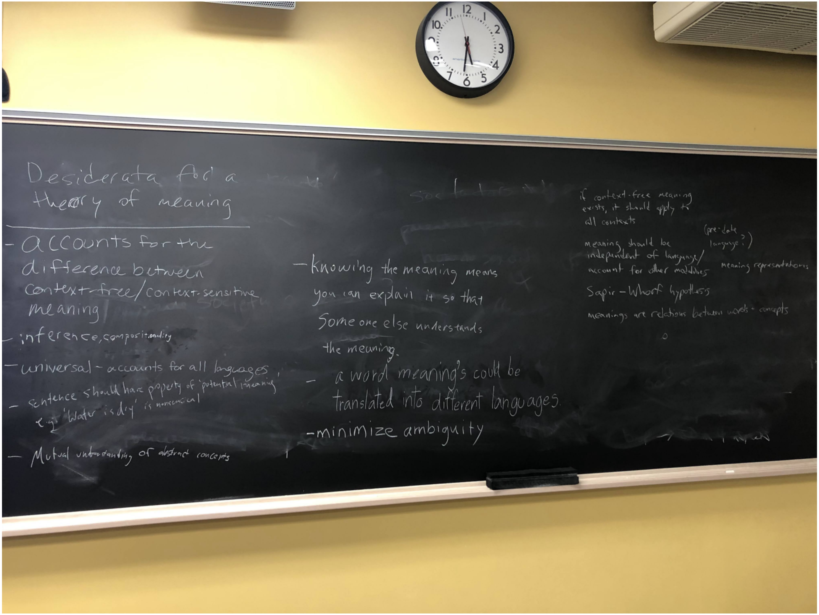
# Today's Plan

- Composition in Dynamic Semantics Handout
- Vector Semantics and Embeddings

# Today's Plan

- Composition in Dynamic Semantics Handout
- Vector Semantics and Embeddings
- (we'll see how far we get...)

# Summary of 8/29 Discussion



# Desiderata

What should a theory of word meaning do for us?

Let's look at some desiderata

From lexical semantics, the linguistic study of word meaning

# Lemmas and senses

lemma

mouse (N)

sense

1. any of numerous small rodents...
2. a hand-operated device that controls a cursor...

Modified from the online thesaurus WordNet

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

Lemmas can be polysemous (have multiple senses)

# 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 / $H_20$

# Relations between senses: Synonymy

Note that there are probably no examples of perfect synonymy.

- Even if many aspects of meaning are identical
- Still may differ based on politeness, slang, register, genre, etc.

# Relation: Synonymy?

water/$H_20$

"$H_20$" in a surfing guide?

big/large

my big sister != my large sister

# The Linguistic Principle of Contrast

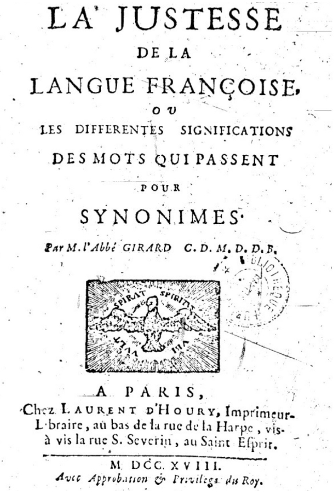
Difference in form → difference in meaning

# 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. B.



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.

Thanks to Mark Aronoff!

# 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

- `coffee, tea`: **similar**
- `coffee, cup`: **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`

# Connotation (sentiment)

- Words have **affective** meanings
  - Positive connotations (*happy*)
  - Negative connotations (*sad*)
- Connotations can be subtle:
  - Positive connotation: *copy, replica, reproduction*
  - Negative connotation: *fake, knockoff, forgery*
- Evaluation (sentiment!)
  - 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
- **arousal**: the intensity of emotion provoked by the stimulus
- **dominance**: the degree of control exerted by the stimulus

| | Word | Score | | Word | Score |
|---|---|---|---|---|---|
| **Valence** | love | 1.000 | | toxic | 0.008 |
| | happy | 1.000 | | nightmare | 0.005 |
| **Arousal** | elated | 0.960 | | mellow | 0.069 |
| | frenzy | 0.965 | | napping | 0.046 |
| **Dominance** | powerful | 0.991 | | weak | 0.045 |
| | leadership | 0.983 | | empty | 0.081 |

Values from NRC VAD Lexicon (Mohammad 2018)

# 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
- Connotation

# Vector Semantics

# Computational models of word meaning

Can we build a theory of how to represent word meaning, that accounts for at least some of the desiderata?

We'll introduce **vector semantics**

- The standard model in language processing!
- Handles many of our goals!

# 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
  - We could conclude this based on words like "leaves" and "delicious" and "sauteed"

# Ongchoi: *Ipomoea aquatica "Water Spinach"*

*空心菜*
*kangkong*
rau muống
...

Yamaguchi, Wikimedia Commons, public domain

# The Distributional Hypothesis

- "You shall know a word by the company it keeps." (Firth, 1957)
- "It may be presumed that any two morphemes A and B having different meanings, also differ somewhere in distribution: there are some environments in which one occurs and the other does not." (Harris, 1951)
- "The similarity of the contextual representations of two words contributes to the semantic similarity of those words." (Miller and Charles, 1991) (emphasis mine)

# The Distributional Hypothesis

- "You shall know a word by the company it keeps." (Firth, 1957)
- "It may be presumed that any two morphemes A and B having different meanings, also differ somewhere in distribution: there are some environments in which one occurs and the other does not." (Harris, 1951)
- "The similarity of the contextual representations of two words contributes to the semantic similarity of those words." (Miller and Charles, 1991) (emphasis mine)

- Words can be represented by (abstractions over) their contexts
  - Specifically, linguistic context

# Idea 1: Defining meaning by linguistic distribution

Let's define the meaning of a word by its distribution in language use, meaning its neighboring words or grammatical environments.

# Idea 2: Meaning as a point in space (Osgood et al. 1957)

## 3 affective dimensions for a word

- **valence**: pleasantness
- **arousal**: intensity of emotion
- **dominance**: the degree of control exerted

| | Word | Score | | Word | Score |
|---|---|---|---|---|---|
| **Valence** | love | 1.000 | | toxic | 0.008 |
| | happy | 1.000 | | nightmare | 0.005 |
| **Arousal** | elated | 0.960 | | mellow | 0.069 |
| | frenzy | 0.965 | | napping | 0.046 |
| **Dominance** | powerful | 0.991 | | weak | 0.045 |
| | leadership | 0.983 | | empty | 0.081 |

NRC VAD Lexicon
(Mohammad 2018)

## Hence the connotation of a word is a vector in 3-space

Idea 1: Defining meaning by linguistic distribution

Idea 2: Meaning as a point in multidimensional space

# Defining meaning as a point in space based on distribution

Each word = a vector (not just "good" or "$w_{45}$")

Similar words are "**nearby in semantic space**"

We build this space automatically by seeing which words are **nearby in text**

# We define meaning of a word as a vector

Called an "embedding" because it's embedded into a space (see textbook)

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
- Later we'll discuss extensions called **contextual embeddings**

# Distributed Representations of Words

- More generally, two approaches to distributed, distributional representations (Baroni et al. 2014):
  - Count-based
    - Count occurrences of words in contexts, optionally followed by some mathematical transformation (e.g., tf-idf, PPMI, SVD)
  - Prediction-based
    - Given some context vector(s) **c**, predict some word **x** (or vice versa)
    - a.k.a. language modeling-based

(e.g., word2vec,  ,  )

# From now on:
# Computing with meaning representations instead of string representations

| | |
|---|---|
| 筌者所以在鱼，得鱼而忘筌 | Nets are for fish;<br>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;<br>Once you get the meaning, you can forget the words |

庄子(Zhuangzi), Chapter 26

Vector Semantics & Embeddings

# Vector Semantics

Vector Semantics & Embeddings

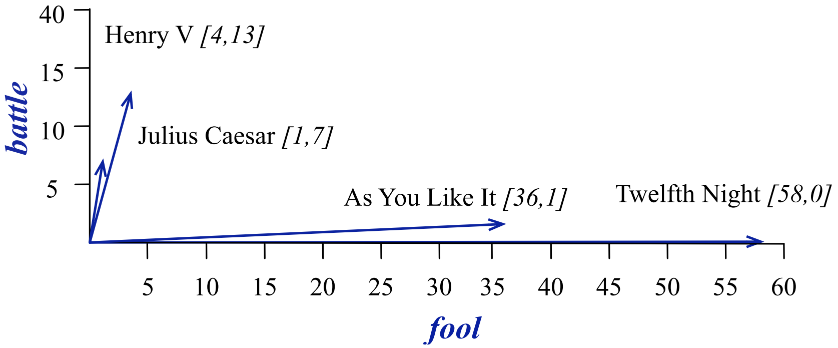
# Words and Vectors

# 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

But comedies are different than the other two

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"

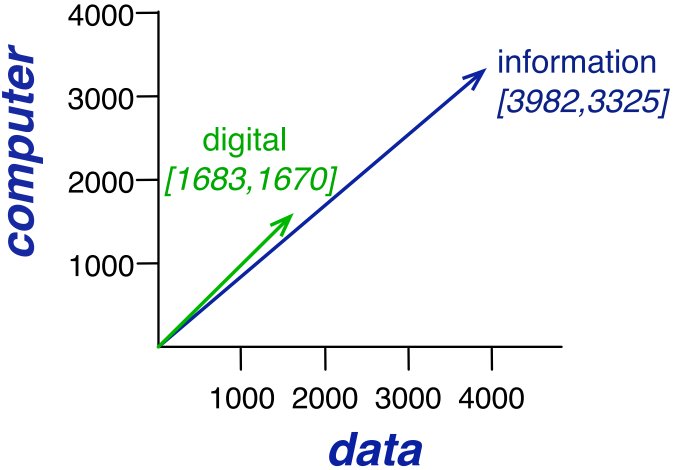
*fool* is "the kind of word that occurs in comedies, especially Twelfth Night"

# 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 | ... |



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

Vector Semantics & Embeddings

# Words and Vectors

Vector Semantics & Embeddings

# Cosine for computing word similarity

# 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 thus be a useful 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

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

Based on the definition of the dot product between two vectors a and b

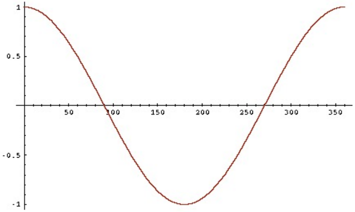
$$\begin{aligned} \mathbf{a}\cdot\mathbf{b} &= |\mathbf{a}||\mathbf{b}|\cos\theta \\ \frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}||\mathbf{b}|} &= \cos\theta \end{aligned}$$

# 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

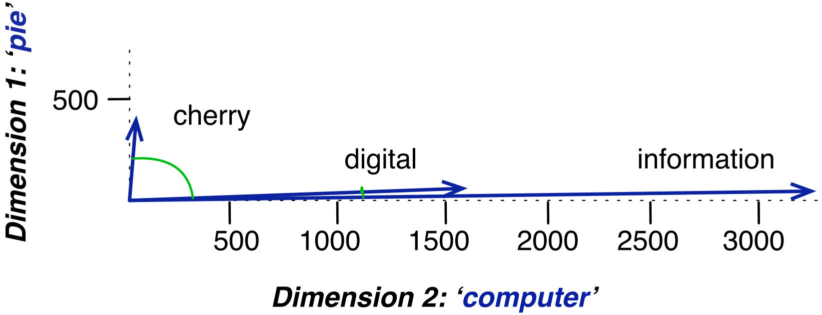
$$\cos(\vec{v},\vec{w}) = \frac{\vec{v}\bullet\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\vec{v}}{|\vec{v}|}\bullet\frac{\vec{w}}{|\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

| | 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