Representing word meaning through observed contexts

Katrin Erk

Computational Semantics

You can guess at a word based on its context (sometimes)

- The maverick maestro, who prefers life in scruffy clothes, was lent the gold-trimmed _____
- Pureed fruits -- apple, pear, _____, but NO CITRUS FRUIT before 9 months.
- I could see no just cause for carrying on after arguing vehemently against the _____, then seeing it carried.
- when we asked them to do an interview about a crime or ______ they say we've not done radio interviews before,
- Mr Kennedy seems to think he was given the _______, but that's not quite right.
- And smiles her Cheshire _____ grin

Sentences from the British National Corpus

You can guess at a word based on its context (sometimes)

- The maverick maestro, who prefers life in scruffy clothes, was lent the gold-trimmed **gown**
- Pureed fruits -- apple, pear, **bananas** but NO CITRUS FRUIT before 9 months.
- I could see no just cause for carrying on after arguing vehemently against the **idea**, then seeing it carried.
- when we asked them to do an interview about a crime or **incident** they say we've not done radio interviews before,
- Mr Kennedy seems to think he was given the **gown** but that's not quite right.
- And smiles her Cheshire **cat** grin

Sentences from the British National Corpus

Describing meaning through context

- When can you guess at the missing word based on its context?
- Why can you guess it?

Describing meaning through context

- When can you guess at the missing word based on its context?
  - Collocations: Cheshire…
  - Similar words nearby: apples, pears, and …
  - Indicative words: gold-trimmed …., argued against ….

- Why can you guess at the missing word based on its context?
  - **Similar words appear in similar contexts**

Describing meaning through context: The computational idea

- Similar words appear in similar contexts
  - Gold-trimmed gown, gold-trimmed robe, but also gold-trimmed premium plastic cups (one of the highest-ranking search hits for “gold-trimmed at this point)
- Measure similarity in meaning as similarity in contexts
- How to describe the contexts of a word?
  - Count other words nearby

Describing meaning through context: The computational idea

- Similar words appear in similar contexts
  - Gold-trimmed gown, gold-trimmed robe, but also gold-trimmed premium plastic cups (one of the highest-ranking search hits for “gold-trimmed at this point)
- Measure similarity in meaning as similarity in contexts
- How to describe the contexts of a word?
  - Count other words nearby

Describing meaning through context: How can we do this in practice?

What is context?

- Target word: the word whose contexts we want to collect
- Context: simply, what appears to the left and to the right of this word in text
- How far away from the target? We choose, for example 3 words either side.

They **picked up red** apples **that had fallen** to the ground

Eating apples **is healthy**

She **ate a red** apple

Pick an apple.

Characterizing a word through its contexts

- Say our target word is “apple”
- We collect all its contexts: among all boldfaced words, count how often each word appears

They **picked up red** apples **that had fallen** to the ground

Eating apples **is healthy**

She **ate a red** apple

Pick an apple.

Counting context words

- They **picked up red** apples **that had fallen** to the ground
- **Eating** apples **is healthy**
- She **ate a red** apple
- **Pick an** apple.

Word count, 3-word context window, lemmatized

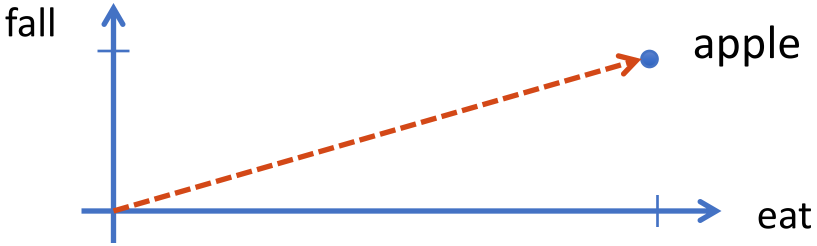
a	be	eat	fall	have	healthy	pick	red	that	up
2	1	2	1	1	1	2	2	1	1

How can we compare two context collections in their entirety?

Count how often “apple” occurs close to other words
in a large text collection (corpus):

eat	fall	ripe	slice	peel	tree	throw	fruit	pie	bite	crab
794	244	47	221	208	160	145	156	109	104	88

Interpret counts as coordinates:



Every context word
becomes a dimension.

How can we compare two context collections in their entirety?

Count how often “apple” occurs close to other words
in a large text collection (corpus):

eat	fall	ripe	slice	peel	tree	throw	fruit	pie	bite	crab
794	244	47	221	208	160	145	156	109	104	88

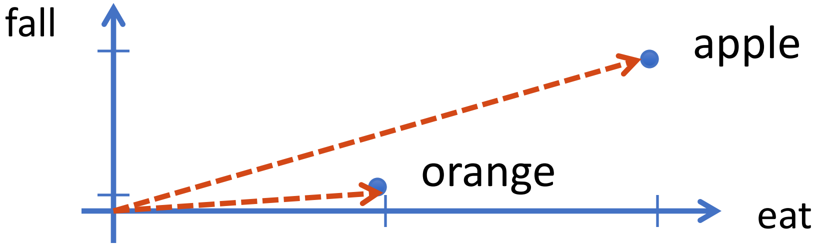
Do the same for “orange”:

eat	fall	ripe	slice	peel	tree	throw	fruit	pie	bite	crab
265	22	25	62	220	64	74	111	4	4	8

How can we compare two context collections in their entirety?

Then visualize both count tables as vectors in the same space:

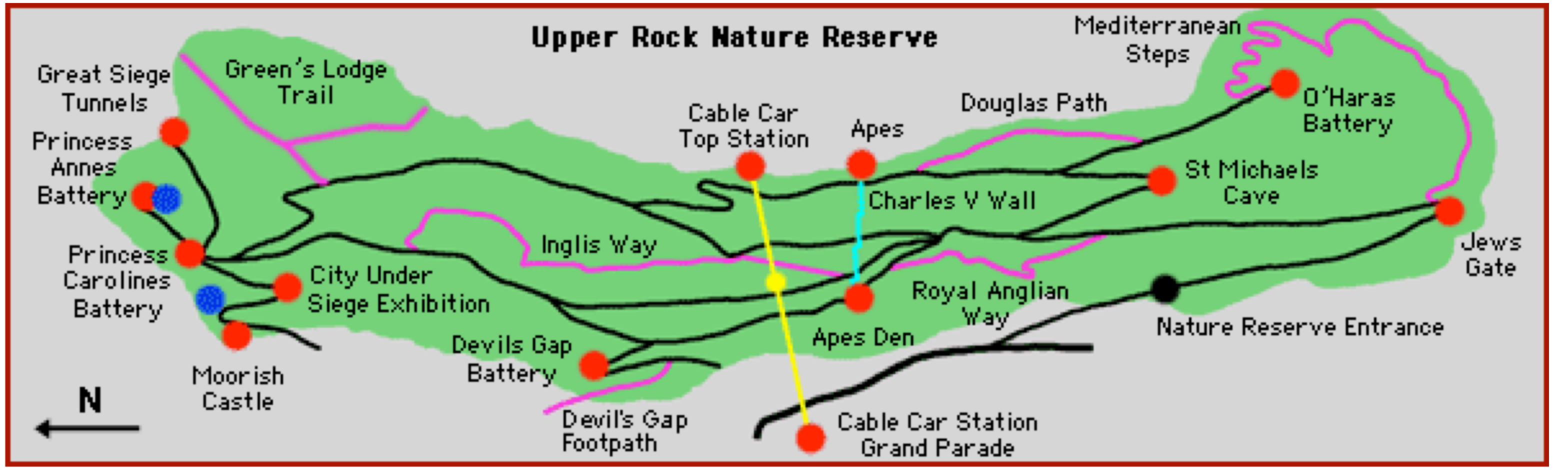
eat	fall	ripe	slice	peel	tree	throw	fruit	pie	bite	crab
794	244	47	221	208	160	145	156	109	104	88
eat	**fall**	**ripe**	**slice**	**peel**	**tree**	**throw**	**fruit**	**pie**	**bite**	**crab**
265	22	25	62	220	64	74	111	4	4	8



Similarity between two words as proximity in space

Specifying components of a context-based meaning representation: similarity

Meaning dis-similarity as distance in meaning space



Footpaths of Gibraltar: Jim Crone, CC BY-SA 3.0 <https://creativecommons.org/licenses/by-sa/3.0>, via Wikimedia Commons

If "apple" and "orange" are points in meaning space, then
I can describe their dis-similarity as the distance of going from one to the other
Distance on a map: Euclidean distance

Euclidean distance in meaning space

- Table of co-occurrence counts as a vector of numbers:
  <794, 244, 47, 221, 208, 160, 145, 156, 109, 104, 88>

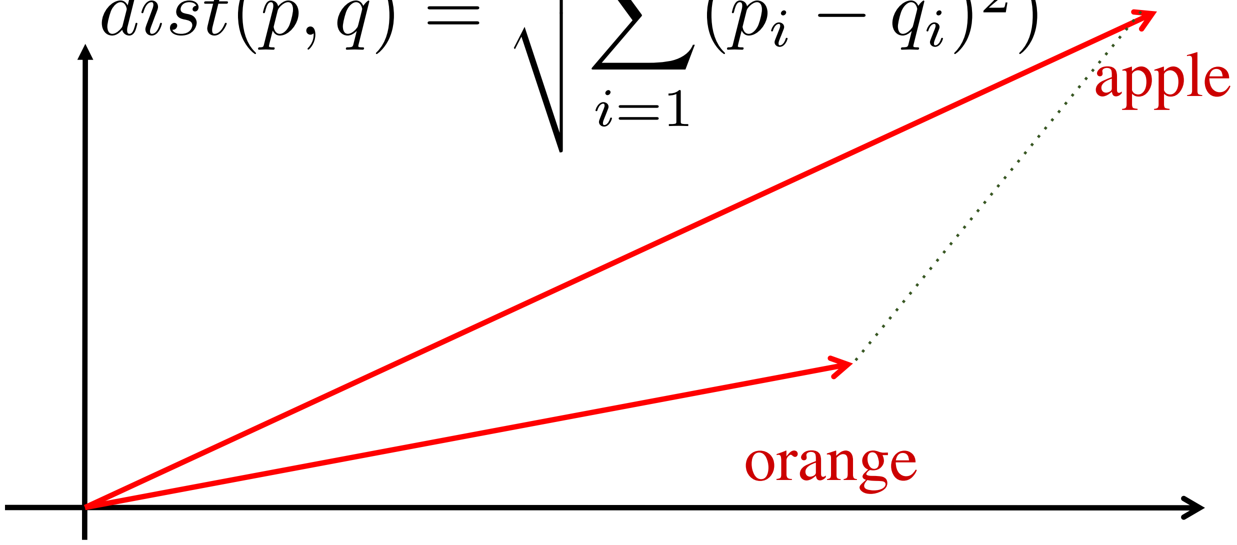
eat	fall	ripe	slice	peel	tree	throw	fruit	pie	bite	crab
794	244	47	221	208	160	145	156	109	104	88

- We also write is as: $\vec{\mathrm{apple}}$

What do we mean by “similarity” of vectors?

Euclidean distance (a dissimilarity measure!):

$$dist(\vec{p},\vec{q}) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2)}$$



Euclidean distance, taken apart

$$dist(\vec{p},\vec{q}) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2)}$$

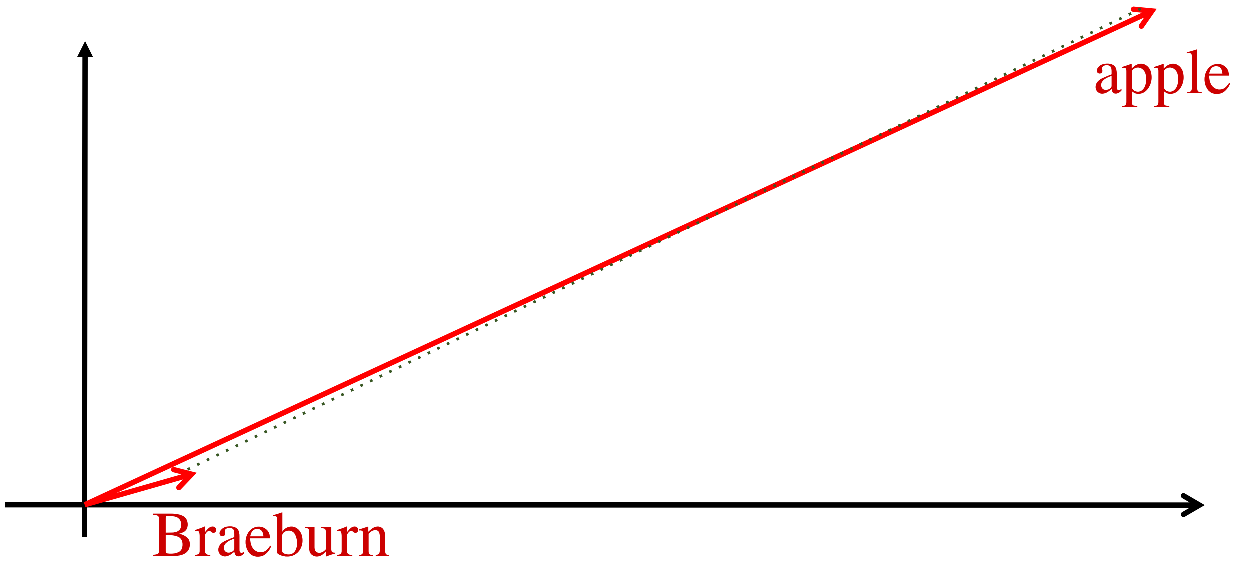
- $\vec{\text{apple}}$: <4, 5, 6>
- $\vec{\text{orange}}$: <1, 3, 5>
- Sum of squared differences:

$$(\text{apple}_1 - \text{orange}_1)^2 + (\text{apple}_2 - \text{orange}_2)^2 + (\text{apple}_3 - \text{orange}_3)^2 =$$
$$(4-1)^2 + (5-3)^2 + (6-5)^2 = 9 + 4 + 1 = 14$$

- Then take the root: $\sqrt{14} = 3.74$

- **This is a dissimilarity measure! The larger, the further apart.**

Problem with Euclidean distance: very sensitive to word frequency!

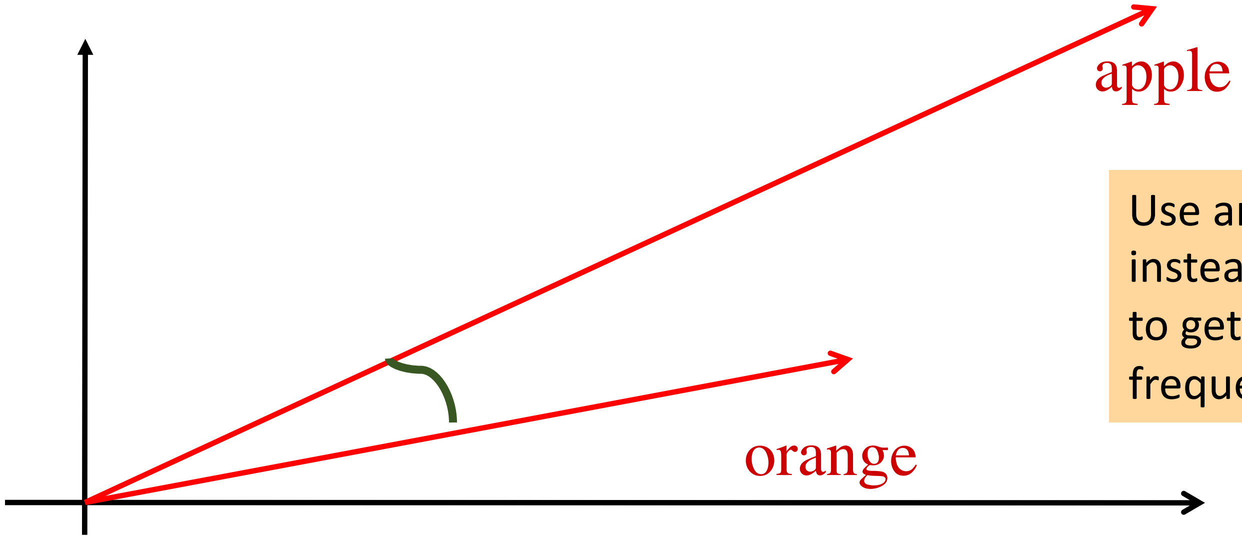


Because all counts are higher
for “apple” than “Braeburn”,
the sum of squared differences
will be very high.
But intuitively, the “arrows”
for “apple” and “Braeburn” point
in the same direction, they
are just different length.

A different formalization of similarity: The vectors need to point in the same direction

Cosine similarity:

$$cos(\vec{p}, \vec{q}) = \frac{\sum_{i=1}^{n} p_i \cdot q_i}{\sqrt{\sum_{i=1}^{n} p_i^2} \cdot \sqrt{\sum_{i=1}^{n} q_i^2}}$$



Use angle between vectors instead of point distance to get around word frequency issues

Cosine similarity, taken apart

$$cos(\vec{p},\vec{q}) = \frac{\sum_{i=1}^{n} p_i \cdot q_i}{\sqrt{\sum_{i=1}^{n} p_i^2} \cdot \sqrt{\sum_{i=1}^{n} q_i^2}}$$

- $\vec{\mathrm{apple}}$: <4, 5, 6>
- $\vec{\mathrm{orange}}$: <1, 3, 5>
- Numerator: “dot product”, a similarity measure in itself.
  - Large if we tend to multiply large apple-values with large orange-values, and small apple-values with small orange-values.
    Smaller if we multiple large apple-values with small orange-values

$$\mathrm{apple}_1 \cdot \mathrm{orange}_1 + \mathrm{apple}_2 \cdot \mathrm{orange}_2 + \mathrm{apple}_3 \cdot \mathrm{orange}_3 =$$

$$4 \cdot 1 + 5 \cdot 3 + 6 \cdot 5 = 49$$

Cosine similarity, taken apart

$$cos(\vec{p},\vec{q}) = \frac{\sum_{i=1}^{n} p_i \cdot q_i}{\sqrt{\sum_{i=1}^{n} p_i^2} \cdot \sqrt{\sum_{i=1}^{n} q_i^2}}$$

- $\overrightarrow{\text{apple}}$: <4, 5, 6>
- $\overrightarrow{\text{orange}}$: <1, 3, 5>
- Numerator: dot-product, 49
- Denominator: length of apple-vector times length of orange-vector
  - We are dividing by the vector lengths, that is, we are normalizing by vector lengths!

$$\text{Length of } \overrightarrow{\text{apple}}\text{:} \ \sqrt{\text{apple}_1^2 + \text{apple}_2^2 + \text{apple}_3^2} = \sqrt{4^2 + 5^2 + 6^2} = \sqrt{16 + 25 + 36} = 8.77$$

$$\text{Length of } \overrightarrow{\text{orange}}\text{:} \ \sqrt{\text{orange}_1^2 + \text{orange}_2^2 + \text{orange}_3^2} = \sqrt{1^2 + 3^2 + 5^2} = \sqrt{1 + 9 + 25} = 5.92$$

Cosine similarity, taken apart

$$cos(\vec{p},\vec{q}) = \frac{\sum_{i=1}^{n} p_i \cdot q_i}{\sqrt{\sum_{i=1}^{n} p_i^2} \cdot \sqrt{\sum_{i=1}^{n} q_i^2}}$$

- $\vec{\text{apple}}$: <4, 5, 6>
- $\vec{\text{orange}}$: <1, 3, 5>
- Numerator: dot-product, 49
- Denominator: length of apple-vector times length of orange-vector, 8.77 * 5.92= 51.91
- Cosine: 49 / 51.91 = 0.94
    - **Cosine is a similarity: the higher, the more similar**
    - 0 = totally different, 1 = totally the same

Specifying more details of the approach: Corpora: Data from which to compute distributional models

Corpora in which to count words

- Corpus = text collection
- What works best for computing a distributional model?
  - “Moby Dick”
  - 2 years of the Wall Street Journal
  - A collection of dating ads
  - A collection of webpage texts

Corpora in which to count words

- Need to be available electronically
- Best possible match for today’s English language in general
  - Mixture of genres
  - Mixture of authors
  - Spoken and written
- Larger is better

Corpora in which to count (English) words

- Brown Corpus
  - 1 million words
  - Balanced corpus, mixture of genres
- British National Corpus
  - 100 million words
  - Balanced corpus, mixture of genres, spoken and written

Corpora in which to count (English) words

- English Gigaword corpus
  - 1 billion words
  - Short news articles
- Wikipedia dump
  - 2 billion words
- UKWaC (UK web as corpus)
  - 2 billion words
  - Collection of webpages ending in .uk
  - Also available for other languages, including deWaC and frWaC
- Google books

Where can we find texts to use for making a distributional model?

- Text in electronic form!
- Newspaper articles
- Project Gutenberg: older books available for free
- Wikipedia
- Text collections prepared for language analysis:
  - Balanced corpora
  - WaC: Scrape all web pages in a particular domain
  - ELRA, LDC hold corpus collections
    - For example, large amounts of newswire reports
  - Google n-grams, Google books

Summary: What data should I use?

- How much data do we need?
  - The more the better
  - General purpose distributional model: at least the size of the British National Corpus, 100 million words
  - Better: add
    - UKWaC (2 billion words)
    - Wikipedia (2 billion words)
- Analyzing one specific genre: you may be able to see a signal even for much less data.
- Mostly, more data gives better models
- Watch out for genre effects

Specifying details of the approach: Targets and contexts

What is a context item, what is a target?

- Target:
    - Word form or lemma, or pieces of words
    - Part-of-speech tag?
- Context item:
    - Word:
        - Word form or lemma?
        - Part-of-speech tag?
        - Content words only? Remove stopwords?
    - Snippet of syntactic structure
    - Word pieces (morphemes, or just frequent substrings)

Snippets of syntactic structure as context items

- TypeDM:
  - shot-n aboard-1 train-n 7.3669
  - shot-n about argue-v 8.1278
  - shot-n about ask-v 28.3219
  - shot-n about bias-v 5.6922
  - shot-n about brag-v 15.7061
  - shot-n coord-1 objective-n 0.4486
  - shot-n coord-1 obstacle-n 1.8468
  - shot-n coord-1 odd-n 2.0591
  - shot-n iobj aim-v 5.2203
  - shot-n iobj ask-v 4.1064

Same corpus (BNC), different contexts (window sizes)

Nearest neighbours of *dog*

2-word window

- cat
- horse
- fox
- pet
- rabbit
- pig
- animal
- mongrel
- sheep
- pigeon

30-word window

- kennel
- puppy
- pet
- bitch
- terrier
- rottweiler
- canine
- cat
- to bark
- Alsatian

How large is the context around a target occurrence?

- Bag-of-words context:
  - Wide context or document: topical similarity
  - Narrow context: Mostly cat-dog and dog-animal, not so much dog-leash
    - Approximation of syntactic-based contexts
- Syntactic parse relations
    - Mostly cat-dog and dog-animal, not so much dog-leash
    - Sparser than bag-of-words

Improving count-based models

Some counts for “letter” in “Pride and Prejudice”. What do you notice?

the	to	of	and	a	her	she	his	is	was	in	that
102	75	72	56	52	50	41	36	35	34	34	33

had	i	from	you	as	this	mr	for	not	on	be	he
32	28	28	25	23	23	22	21	21	20	18	17

but	elizabeth	with	him	which	by	when	jane
17	17	16	16	16	15	14	12

Some counts for “letter” in “Pride and Prejudice”. What do you notice?

the	to	of	and	a	her	she	his	is	was	in	that
102	75	72	56	52	50	41	36	35	34	34	33

had	i	from	you	as	this	mr	for	not	on	be	he
32	28	28	25	23	23	22	21	21	20	18	17

but	elizabeth	with	him	which	by	when	jane
17	17	16	16	16	15	14	12

All the most frequent co-occurring words are function words.

Some words are more informative than others

- Function words co-occur frequently with all words
    - That makes them less informative
- They have much higher co-occurrence counts than content words
    - They can “drown out” more informative contexts

Using association rather than co-occurrence counts

- Degree of association between target and context:
  - High association: high co-occurrence with “letter”, lower with everything else
  - Low association: lots of co-occurrence with all words
- Many ways of implementing this
- For example Pointwise Mutual Information between target a and context b:

$$\boxed{PMI(a,b) = \log \frac{P(a,b)}{P(a) \cdot P(b)}}$$

Variants of PMI

- PPMI, positive PMI:

$$PPMI(a,b) = \begin{cases} PMI(a,b) & \text{if } \geq 0 \\ 0 & \text{else} \end{cases}$$

- Local PMI, LMI:

$$LMI(a,b) = P(a,b) \log \frac{P(a,b)}{P(a)P(b)} = Observed \log \frac{Observed}{Expected}$$

Dimensionality reduction

- Singular Value Decomposition (SVD)
- Nonnegative Matrix Factorization (NMF)
- Principal Component Analysis
- t-SNE
- Amounts to a soft clustering of dimensions
- SVD: resulting dimensions are pairwise orthogonal

Taking a step back: Where does the idea of meaning-as-context come from?

Taking a step back:
This is not just an idea in computational semantics

- **Background in philosophy of language**: Wittgenstein, “meaning” as “use”
  - Man kan für eine große Klasse von Fällen der Benützung des Wortes ‘Bedeutung’ – wenn auch nicht für alle Fälle seiner Benützung – dieses Wort so erklären: Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache.
    -- Wittgenstein, Philosophical Investigations
  - For a large class of cases – though not for all – in which we employ the word ‘meaning’ it can be explained thus: the meaning of a word is its use in the language. (translation: Anscombe/Stokhof)

Taking a step back:
This is not just an idea in computational semantics

- **Background in linguistics**:
- Zellig Harris (1957): Classify linguistic units by observing the contexts they occur in
  - phonemes
  - morphemes
  - phrase types: noun phrase, verb phrase…
  - Not specifically about semantics
- John Firth (1957): “collocations”
  - identify senses of a word by looking at groups of contexts in which it appears
  - “You shall know a word by the company it keeps”

Taking a step back:
This is not just an idea in computational semantics

- **Background in psychology**:
- Landauer/Dumais 1998: “A solution to Plato’s problem”
- How come you know so many words?
- “A typical American seventh grader knows the meaning of 10-15 words today that she did not know yesterday. She must have acquired most of them as a result of reading because (a) the majority of English words are used only in print, (b) she already knew well almost all the words she would have encountered in speech, and (c) she learned less than one word by direct instruction.”

Taking a step back:
This is not just an idea in computational semantics

- **Background in psychology**:
- Landauer and Dumais were actually using a computational model developed for information retrieval (web search)
- Their argument: Humans learn new words by seeing them used in context, and context-based computational models of lexical meaning may be a good model of what is going on
- We will get back to distributional models in psychology later