An introduction to distributional models

Katrin Erk

We can often infer (part of) the meaning of a word from textual context

- A hairy little wampimuk was sleeping behind a tree.

What is a wampimuk?

We can often infer (part of) the meaning of a word from textual context

- A hairy little wampimuk was sleeping behind a tree.

What is a wampimuk?

Maybe something like a squirrel or a fox.

Similar words are often found in similar linguistic contexts.

Computational models of context

Similar words are often found in similar context

So it follows that:

If we had a measure for the similarity of contexts,
we could use it as a measure for the similarity of words

How can we make this concrete?

1. What are the contexts of the word “apple”?
2. How do you measure the contextual similarity of “apple” and “orange”?

The main idea: context-word counts and a distributional space

What are the contexts of the word “apple”?

Say we have this “corpus”:

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

Now what is a context?

What are the contexts of the word “apple”?

Say we have this “corpus”:

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

Now what is a context? Let’s say:

- “Apple”, or “apples”, is our target word
- Context = three words left, three words right of the target word. Stop at sentence boundaries.
- We lemmatize all words

What are the contexts of the word “apple”?

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

3-word context, lemmatized

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

What are the contexts of the word “apple”?

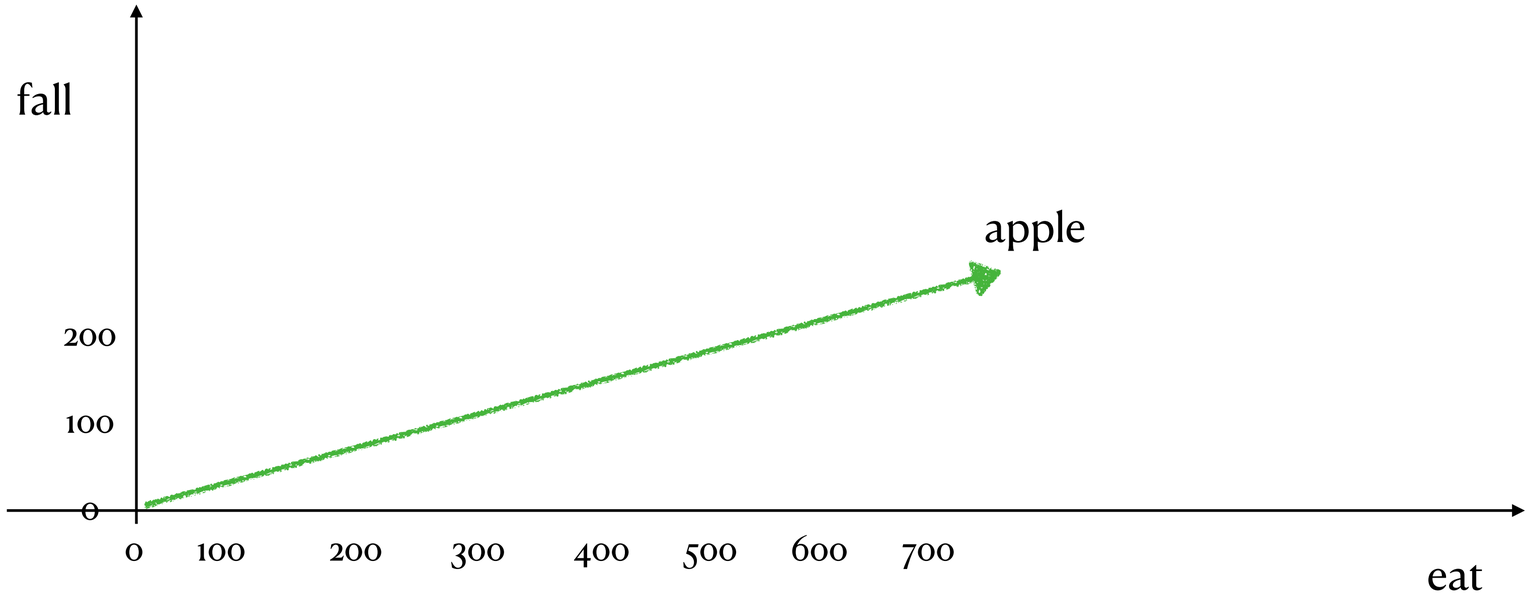
We use a bigger corpus: Context word counts for (some) context words of “apple” in the British National Corpus (100 million words)

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

The core trick: View a table of counts as a set of coordinates

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: First dimension is “eat” with a coordinate of 794. Second dimension is “fall” with a coordinate of 244.



How do you measure the contextual similarity of “apple” and “orange”?

apple

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

orange

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

Count how often “apple” occurs close to other words
in a large text collection (corpus). Do the same for “orange”.
Important: use the same context words in both cases

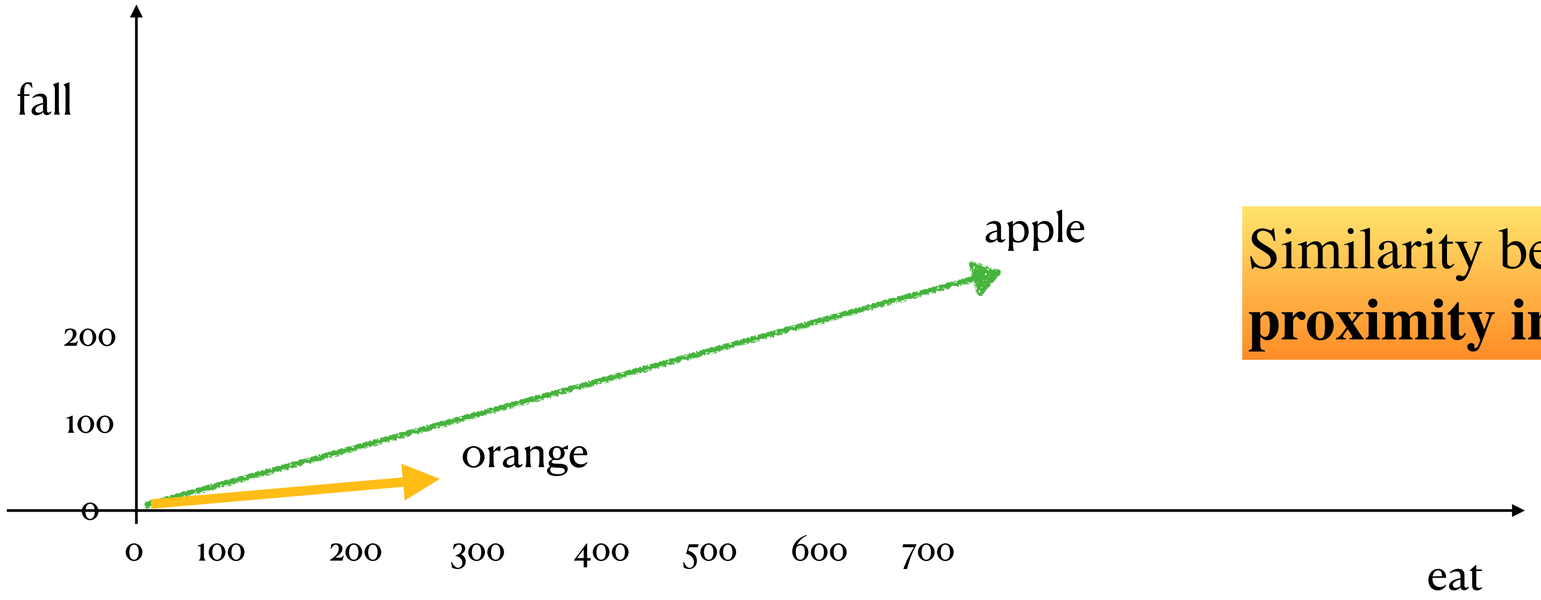
The core idea: tables of counts as coordinates in a semantic space

apple

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

orange

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

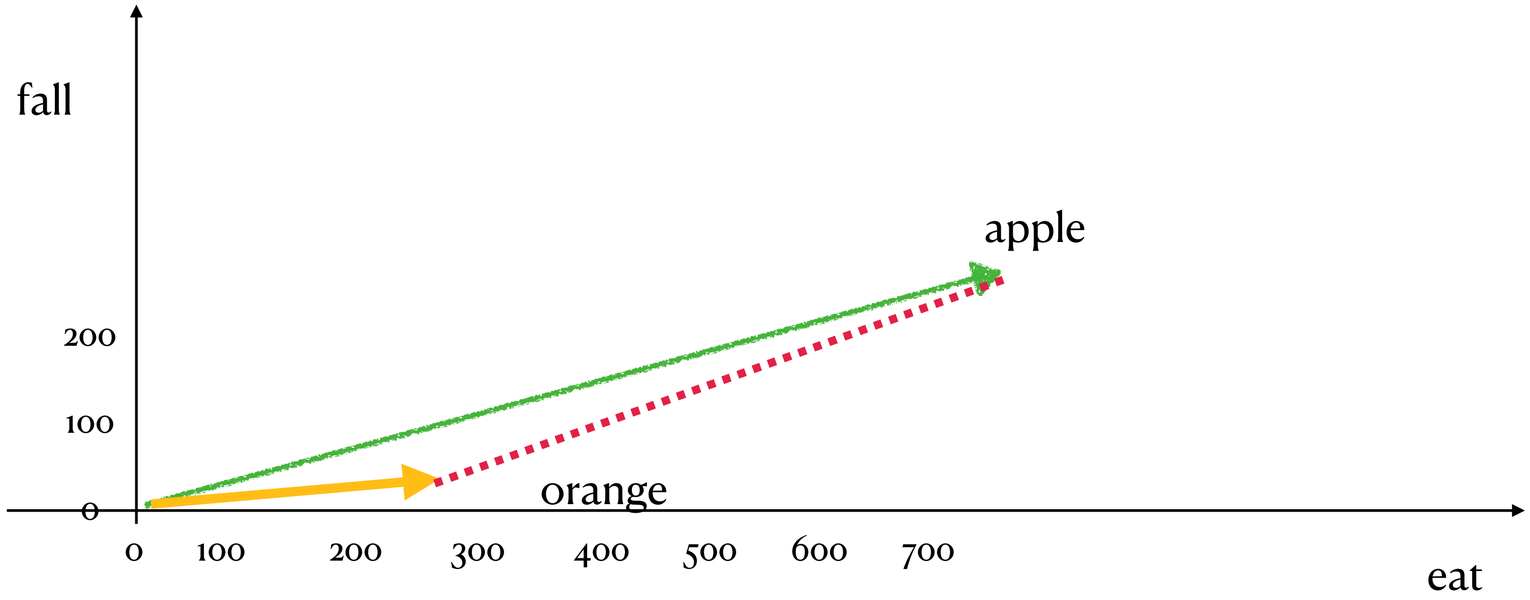


Similarity between two words as **proximity in a semantic space**

Proximity in space

How do you measure the proximity in space of “apple” and “orange”?

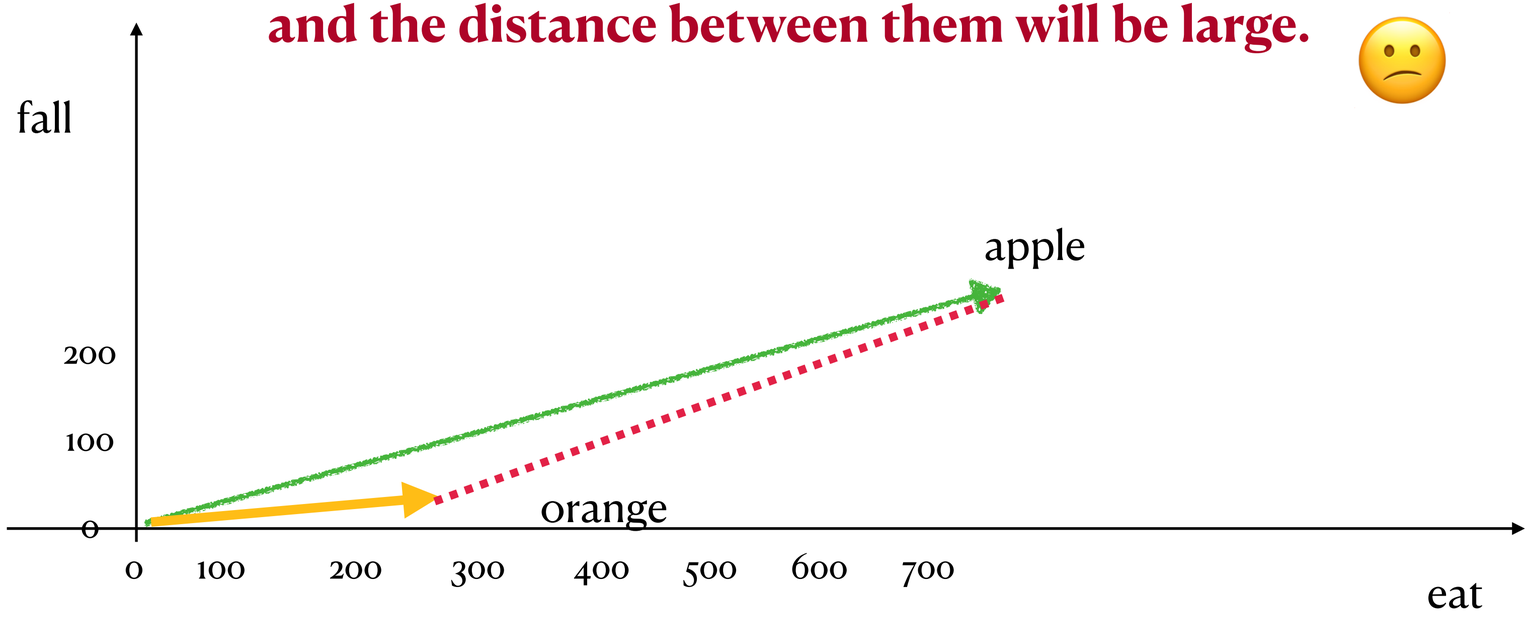
Option 1: proximity = “walking distance”: Euclidean distance



How do you measure the proximity in space of “apple” and “orange”?

Option 1: proximity = “walking distance”: Euclidean distance

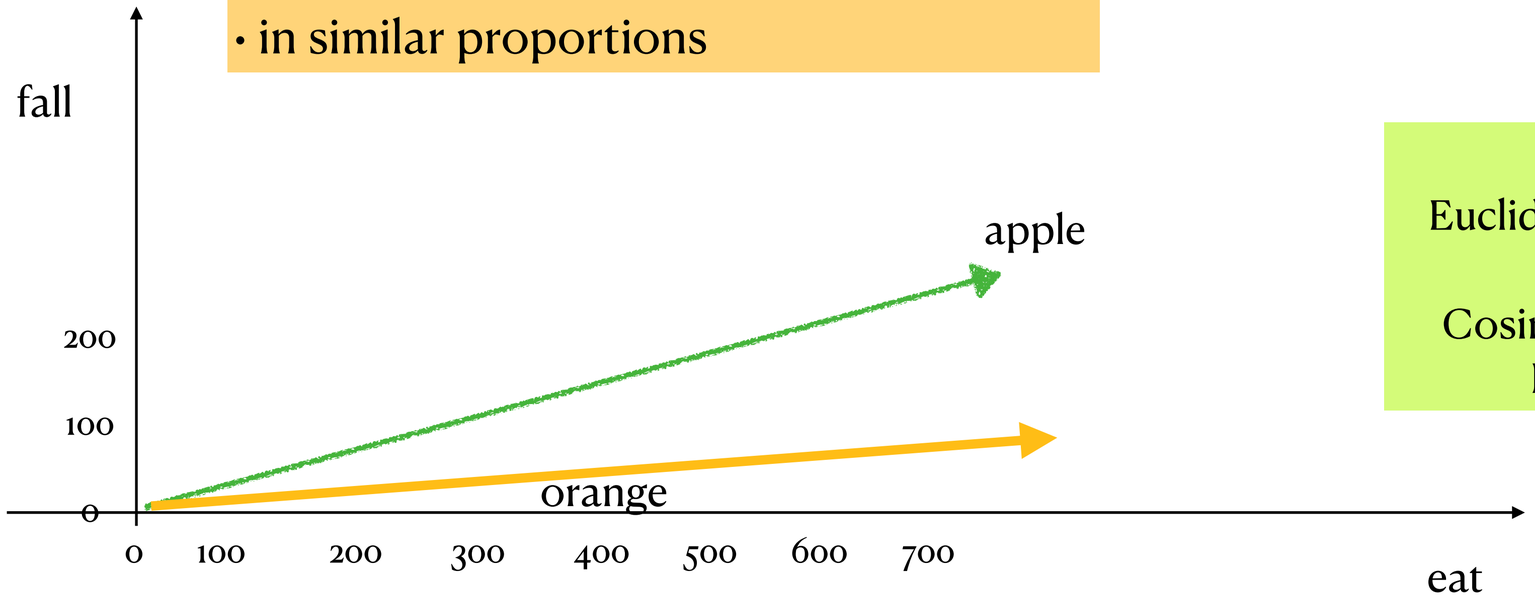
Problem with this option: Some words have all-around larger counts than others! Think “orange” versus “pomelo”.
So “orange” will have larger values in all dimensions than “pomelo”, and the distance between them will be large.



How do you measure the proximity in space of “apple” and “orange”?

Option 2: **Cosine similarity**: make all vectors equally long first!

Two vectors are similar if they point in roughly the same direction:
- similar context words
- in similar proportions



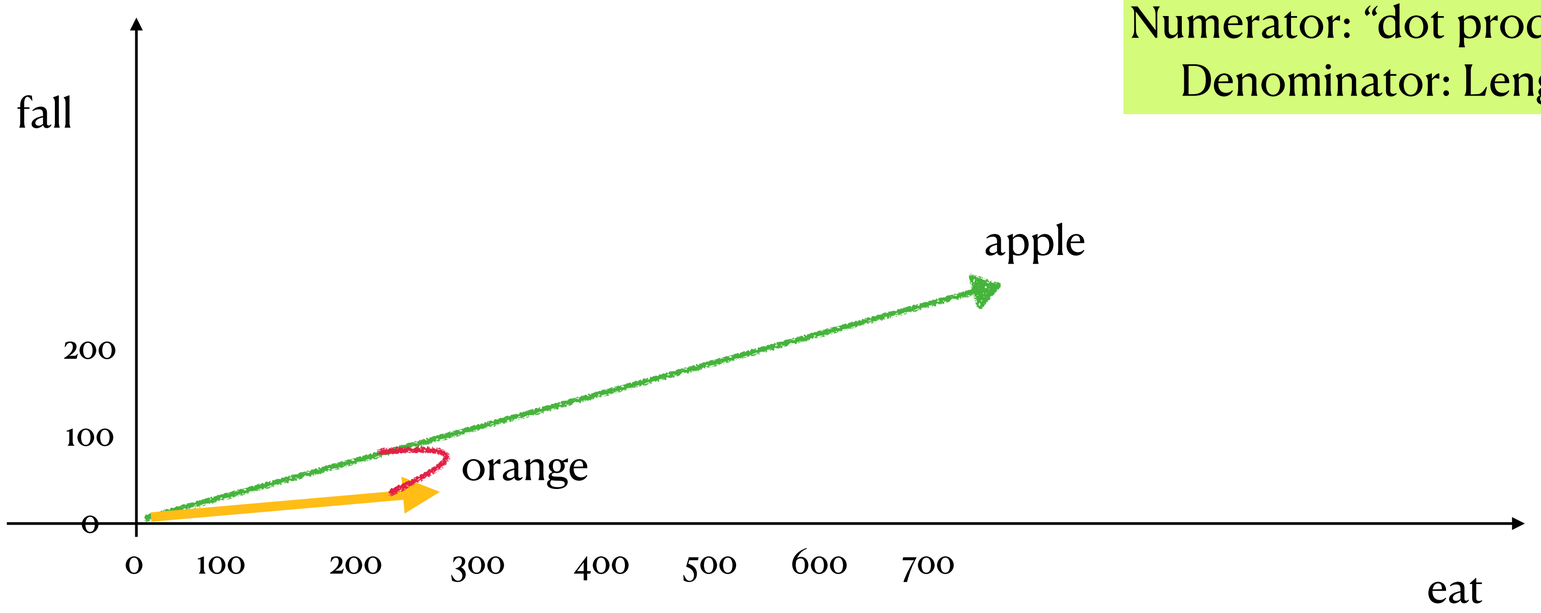
Watch out:
Euclidean distance is a distance.
Larger = less similar
Cosine similarity is a similarity.
Larger = more similar

How do you measure the proximity in space of “apple” and “orange”?

Option 2: **Cosine similarity:** cosine of the angle between the vectors

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

Numerator: “dot product”, a similarity measure
Denominator: Lengths of the two vectors



How do you measure the proximity in space of “apple” and “orange”?

Cosine similarity example:

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

- p = (1,2,3), q = (2,1,4), then we get:
- Numerator: p1 q1 + p2 q2 + p3 q3 =
- 1*2 + 2 * 1 + 3 * 4 = 16
- Denominator: sqrt(1*1 + 2*2 + 3*3) * sqrt(2*2 + 1*1 + 4*4) = sqrt(14) * sqrt(21) = 3.74 * 4.58 = 17.15
- Cosine: 16 / 17.15 = 0.93

Making it work

- **What data to use?**
  - Generally: More data = more informative vectors
  - Genre matters! You get different vectors from a news corpus than a fiction corpus
    - This is a good thing: The vector reflects the contexts in which the word has been seen in this corpus. It is a summary of that word's use in your corpus

Making it work

- **Beware the frequent words**
  - The most frequent context words are always something like ***the, a, in, of, and,*** …
  - So all words are very similar because they all occur in similar contexts 😕
  - What to do?
    - Remove stop words or most frequent words
    - Don’t use counts, use association weights.
      - Let’s say target = “letter”
      - Weight for context word “write” is high: “write” appears with “letter” more frequently than with arbitrary words
      - Weight for context word “the” is low: “the” appears with “letter” about as often as with any arbitrary word
      - Methods: pointwise mutual information, tf/idf

Some big questions

Distributional information in human concepts?

Distributional models can be used to approximate human behavior in experiments like word similarity judgments, categorization, or association. (details later)

Does that mean that human concepts involve distributional information too, or are maybe even solely about distributional information?

Strong distributional hypothesis: human concepts are distributional

Weak distributional hypothesis: distributional spaces reflect something of human concepts

Distributional information in human concepts?

Question first raised in: Landauer and Dumais, “A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge”, http://wordvec.colorado.edu/papers/Landauer_Dumais_1997.pdf

Great discussion in Lenci, “Distributional semantics in linguistic and cognitive research”, https://linguistica.sns.it/RdL/20.1/ALenci.pdf

Distributional models and formal semantics

Distributional models: fine-grained representations of word meaning, can be computed automatically from data

In formal semantics, often very impoverished representations of meanings of many words

- Use distributional vectors in formal semantics?
- In linear algebra, many operations for combining vectors. Reformulate compositionality as vector composition?

Distributional models and formal semantics

McNally, “Kinds, descriptions of kinds, concepts, and distributions”, https://core.ac.uk/download/pdf/229593426.pdf
Distributional vectors as “ersatz conceptual representations”. Nominals denote kinds, and kinds are associated with distributional vectors.

Alternative notions of compositionality:

- Baroni and colleagues, starting with Baroni and Zamparelli, “Nouns are vectors, adjectives are matrices”, https://aclanthology.org/D10-1115.pdf
- Sadrzadeh and colleagues, including Sadrzadeh and Muskens, “Static and dynamic vector semantics for lambda calculus models of natural language” , https://doaj.org/article/d2fac845a74d40198e7c87c886557e66, also Sadrzadeh et al, “Exploring Semantic Incrementality with Dynamic Syntax and Vector Space Semantics”, http://www.eecs.qmul.ac.uk/~mpurver/papers/sadrzadeh-et-al18semdial.pdf

Distributional models as data

- Distributional representation of a word: many utterances involving the word, compacted into a single vector
- Distributional model as compact record of many word uses
- What can this tell us about how people use a word?
- Distributional representations contain a mixture of information
  - word sense
  - part of speech/syntax, e.g. verbs and nouns differ in their immediate context
  - connotations, associations, cultural traces, biases
- Is some of this data, some of it noise? Or is all of it data?
  How can we bring one to the forefront and suppress the other?
  What does this mixture of information mean for what humans store in their minds about words?

How usable are vector space representations as models of meaning?

- **Word vector models conflate senses:**
  - “bank” vector will contain a mixture of financial contexts and river contexts
  - Then how can we use this to study polysemy, or anything to do with polysemous words?
- **Word vectors cannot tell antonyms apart:**
  “good” and “evil” tend to appear in the same contexts
  - How is this with contextualized embeddings? I don’t know
- **Similarity doesn’t clearly distinguish between semantic relations:**
  - synonymy, hypernymy, co-hyponymy, “context-onymy”
  - But we can distinguish “context-onymy” from taxonomic similarity to some degree, more on this later

How usable are vector space representations as models of meaning?

- **Word vector models conflate senses:**
  - “bank” vector will contain a mixture of financial contexts and river contexts
  - Then how can we use this to study polysemy, or anything to do with polysemous words?
- **Word vectors cannot tell antonyms apart:**
  “good” and “evil” tend to appear in the same contexts
  - How is this with contextualized embeddings? I don’t know
- **Similarity doesn’t clearly distinguish between semantic relations:**
  - synonymy, hypernymy, co-hyponymy, “context-onymy”
  - But we can distinguish “context-onymy” from taxonomic similarity to some degree, more on this later

How usable are vector space representations as models of meaning?

- **Word vector models conflate senses:**
  - “bank” vector will contain a mixture of financial contexts and river contexts
  - Then how can we use this to study polysemy, or anything to do with polysemous words?
  - But: big step forward with contextualized embeddings
- **Word vectors cannot tell antonyms apart:**
  “good” and “evil” tend to appear in the same contexts
  - But: how well do contextualized embeddings do on this? I don’t know
- **Similarity doesn’t clearly distinguish between semantic relations:**
  - synonymy, hypernymy, co-hyponymy, “context-onymy”
  - But we can distinguish “context-onymy” from taxonomic similarity to some degree, more on this later

How usable are vector space representations as models of meaning?

- Distributional models are very fine-grained representations
- Distributional models are learned from corpus data,
  and can be used to explore word usages in a particular corpus
- Distributional models are empirical: I don't just have to trust my own intuitions on word meaning, I can challenge them with data

A quick overview over types of distributional models

A quick overview of types of distributional models

- **Count-based models:**
  count context words around the target. One vector per word.
- **Bayesian models: Topic modeling**
  context words as coming from latent clusters to be inferred.
  Typically used to characterize documents, not words (
  documents characterized by words appearing in them)
- **Prediction-based models at the word type level:**
  use machine learning to fit vectors by predicting target-context co-occurrence.
  One vector per word.
- **Prediction-based models at the word token level:**
  use machine learning to fit vectors by predicting target-context co-occurrence
  in a particular sentence context. One vector per word occurrence.

A quick overview of types of distributional models

Count-based models: mix and match

- Pick a context window size
- Transforming counts to weights: pick a method
- Pick whether to do dimensionality reduction (clustering of dimensions)
- No clear best choice

One particular set of choices:

- LSA, Latent Semantic Analysis: longtime the method of choice in psychology