Word meaning, Embeddings, and Word2vec
What do words mean?

First thought: look in a dictionary

http://www.oed.com/
pepper, n.

Pronunciation: /ˈpepər/, U.S. /ˈpepər/

Forms: OE peopor (rare), OE pipcer (transmission error), OE pipur, OE piper (rare).

Frequency (in current use):

Etymology: A borrowing from Latin. Etymology: Latin piper. < classical Latin piper, a loanword < Indo-Aryan (as is ancient Greek πιπερ); compare Sanskrit pipur.

I. The spice or the plant.

1. A hot pungent spice derived from the prepared fruits (peppercorns) of the pepper plant, Piper nigrum (see sense 2), used from early times to season food, either whole or ground to powder (often in association with salt). Also (locally, chiefly with distinguishing word): a similar spice derived from the fruits of certain other species of the genus Piper; the fruits themselves.

The ground spice from Piper nigrum comes in two forms, the more pungent black pepper, produced from black peppercorns, and the milder white pepper, produced from white peppercorns: see black adj. and n. Special uses 5a, peppers on n. 1a, and white adj. and n. Special uses 7b(a).

2. The plant Piper nigrum (family Piperaceae), a climbing shrub indigenous to South Asia and also cultivated elsewhere in the tropics, which has alternate stalked entire leaves, with pendulous spikes of small green flowers opposite the leaves, succeeded by small berries turning red when ripe. Also more widely: any plant of the genus Piper or the family Piperaceae.

3. Usu. with distinguishing word: any of numerous plants of other families having hot pungent fruits or leaves which resemble pepper (1a) in taste and in some cases are used as a substitute for it.

b. The California pepper tree, Schinus molle. Cf. pepper tree n.


3. Any of various forms of capsicum, esp. Capsicum annuum var. annuum. Originally (chiefly with distinguishing word): any variety of the C. annuum Longum group, with elongated fruits having a hot, pungent taste, the source of cayenne, chilli powder, paprika, etc., or of the perennial C. frutescens, the source of Tabasco sauce. Now frequently (more fully sweet pepper): any variety of the C. annuum Grossum group, with large, bell-shaped or apple-shaped, mild-flavoured fruits, usually ripening to red, orange, or yellow and eaten raw in salads or cooked as a vegetable. Also: the fruit of any of these capsicums.

Sweet peppers are often used in their green immature state (more fully green pepper), but some new varieties remain green when ripe.
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.
There are relations between senses
Relation: Synonymy

Synonyms have the same meaning in some or all contexts.
- filbert / hazelnut
- couch / sofa
- big / large
- automobile / car
- vomit / throw up
- water / H₂O
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.

The Linguistic Principle of Contrast:
- Difference in form -> difference in meaning
Relation: Synonymy?
water/H₂O
big/large
brave/courageous
Relation: Antonymy

Senses that are opposites with respect to 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: Similarity

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

car, bicycle
cow, horse
Ask humans how similar 2 words are

<table>
<thead>
<tr>
<th>word1</th>
<th>word2</th>
<th>similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>vanish</td>
<td>disappear</td>
<td>9.8</td>
</tr>
<tr>
<td>behave</td>
<td>obey</td>
<td>7.3</td>
</tr>
<tr>
<td>belief</td>
<td>impression</td>
<td>5.95</td>
</tr>
<tr>
<td>muscle</td>
<td>bone</td>
<td>3.65</td>
</tr>
<tr>
<td>modest</td>
<td>flexible</td>
<td>0.98</td>
</tr>
<tr>
<td>hole</td>
<td>agreement</td>
<td>0.3</td>
</tr>
</tbody>
</table>

SimLex-999 dataset (Hill et al., 2015)
Relation: Word relatedness

Also called "word association"

Words 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
\textit{surgeon, scalpel, nurse, anaesthetic, hospital}

restaurants
\textit{waiter, menu, plate, food, menu, chef}

houses
\textit{door, roof, kitchen, family, bed}
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 subordinate of *mango*

<table>
<thead>
<tr>
<th><strong>Superordinate</strong></th>
<th>vehicle</th>
<th>fruit</th>
<th>furniture</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subordinate</strong></td>
<td>car</td>
<td>mango</td>
<td>chair</td>
</tr>
</tbody>
</table>
These levels are not symmetric

One level of category is distinguished from the others

The "basic level"
Name these items
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

Words have **affective** meanings

positive connotations (*happy*)

negative connotations (*sad*)

positive evaluation (*great, love*)

negative evaluation (*terrible, hate*).
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
- Connotation
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 and 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?

[Graph showing percentage against relative diameter for cups and bowls.]

1. 
2. 
3. 
4.
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 $r_b = \alpha_1 + \alpha_2 + \ldots \alpha_0$ and $\alpha_1$ 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”
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 on "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

<table>
<thead>
<tr>
<th>Game 1</th>
<th>Game 2</th>
<th>Game 3</th>
<th>Game 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>BCD</td>
<td>ACD</td>
<td>ABD</td>
</tr>
</tbody>
</table>

“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
How about a radically different approach?
Ludwig Wittgenstein

Pl #43:
"The meaning of a word is its use in the language"
Let's define words by their usages

In particular, 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 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
Ong choi: *Ipomoea aquatica*
"Water Spinach"
We'll build a new model of meaning focusing on 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

Fine-grained model of meaning for similarity

- NLP tasks like sentiment analysis
  - With words, requires **same** word to be in training and test
  - With embeddings: ok if **similar** words occurred!!!
2 kinds of embeddings

**tf-idf**
- The workhorse of Information Retrieval!
- 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 distinguish nearby and far-away words
Review: words, vectors, and co-occurrence matrices
Term-document matrix

Each document is represented by a vector of words

<table>
<thead>
<tr>
<th></th>
<th>As You Like It</th>
<th>Twelfth Night</th>
<th>Julius Caesar</th>
<th>Henry V</th>
</tr>
</thead>
<tbody>
<tr>
<td>battle</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>good</td>
<td>114</td>
<td>80</td>
<td>62</td>
<td>89</td>
</tr>
<tr>
<td>fool</td>
<td>36</td>
<td>58</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>wit</td>
<td>20</td>
<td>15</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Visualizing document vectors

Henry V [4,13]
Julius Caesar [1,7]
As You Like It [36,1]
Twelfth Night [58,0]
Vectors are the basis of information retrieval

<table>
<thead>
<tr>
<th></th>
<th>As You Like It</th>
<th>Twelfth Night</th>
<th>Julius Caesar</th>
<th>Henry V</th>
</tr>
</thead>
<tbody>
<tr>
<td>battle</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>good</td>
<td>114</td>
<td>80</td>
<td>62</td>
<td>89</td>
</tr>
<tr>
<td>fool</td>
<td>36</td>
<td>58</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>wit</td>
<td>20</td>
<td>15</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Vectors are similar for the two comedies
Different than the history

Comedies have more *fools* and *wit* and fewer *battles*.
New idea for word meaning: Words can be vectors too!!!

### Table: Term-Document Matrix

<table>
<thead>
<tr>
<th></th>
<th>As You Like It</th>
<th>Twelfth Night</th>
<th>Julius Caesar</th>
<th>Henry V</th>
</tr>
</thead>
<tbody>
<tr>
<td>battle</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>good</td>
<td>114</td>
<td>80</td>
<td>62</td>
<td>89</td>
</tr>
<tr>
<td>fool</td>
<td>36</td>
<td>58</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>wit</td>
<td>20</td>
<td>15</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

_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

<table>
<thead>
<tr>
<th></th>
<th>aardvark</th>
<th>computer</th>
<th>data</th>
<th>pinch</th>
<th>result</th>
<th>sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>pineapple</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>digital</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>information</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

sugar, a sliced lemon, a tablespoon of oyment. Cautiously she sampled her first vell suited to programming on the digital for the purpose of gathering data and apricot jam, a pinch ea and another fru In finding the o necessary for tl
Cosine for computing word similarity

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

\(v_i\) is the count for word \(v\) in context \(i\)

\(w_i\) is the count for word \(w\) in context \(i\).

\(\vec{a} \cdot \vec{b} = |\vec{a}| |\vec{b}| \cos \theta\)

\(\frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|} = \cos \theta\)

\(\text{Cos}(v,w)\) is the cosine similarity of \(v\) and \(w\)
Cosine as a similarity metric

-1: vectors point in opposite directions
+1: vectors point in same directions
0: vectors are orthogonal
\[
\cos(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{\|\vec{v}\| \|\vec{w}\|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2} \sqrt{\sum_{i=1}^{N} w_i^2}}
\]

Which pair of words is more similar?

\[
\text{cosine}(\text{apricot}, \text{information}) = \frac{1 + 0 + 0}{\sqrt{1 + 0 + 0} \sqrt{1 + 36 + 1}} = \frac{1}{\sqrt{38}} = .16
\]

\[
\text{cosine}(\text{digital}, \text{information}) = \frac{0 + 6 + 2}{\sqrt{0 + 1 + 4} \sqrt{1 + 36 + 1}} = \frac{8}{\sqrt{38} \sqrt{5}} = .58
\]

\[
\text{cosine}(\text{apricot}, \text{digital}) = \frac{0 + 0 + 0}{\sqrt{1 + 0 + 0} \sqrt{0 + 1 + 4}} = 0
\]

<table>
<thead>
<tr>
<th></th>
<th>large</th>
<th>data</th>
<th>computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>digital</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>information</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>
Visualizing cosines (well, angles)
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!
Many functions can help in reweighting the counts

**tf-idf:**

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

$$w_{t,d} = tf_{t,d} \times idf_t$$

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

**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.
Vector Semantics

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 (less weights to tune)
- Dense vectors may **generalize** better than storing 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**
Three methods for getting short dense vectors

- “Neural Language Model”-inspired models
  - Word2vec (skip-grams, CBOW), Glove
- Singular Value Decomposition (SVD)
  - A special case of this is called LSA – Latent Semantic Analysis
- Brown clustering
Vector Semantics

Embeddings inspired by neural language models: word2vec
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
Word2Vec: Skip-Gram Task

Word2vec provides a variety of options. Let's do

"skip-gram with negative sampling" (SGNS)
Approach: predict if candidate word $c$ is a "neighbor"

- Treat the target word $t$ and a true neighboring context word $c$ as **positive examples**.
- Randomly sample other words in the lexicon to get negative examples.
- Use logistic regression to train a classifier to distinguish those two cases.
- Use the weights as the embeddings.
Skip-Gram Training Data

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

- I would like to **go** someplace nearby for lunch

[target]

The training data:

- input/output pairs (centering on **go**)


Skip-Gram Training data

I would like to go someplace nearby for lunch

Positive

{target, context}

{go, like}
{go, to}
{go, someplace}
{go, nearby}
Skip-Gram Training data

I would like to go someplace nearby for lunch

Positive

{target, context}

{go, like}
{go, to}
{go, someplace}
{go, nearby}

That’s fine for positive data. But for training a binary classifier we need negative examples.

Let's sample other words from the lexicon (that don’t occur with the target word in this context).
I would like to go someplace nearby for lunch

Positive
{+ target, context}
{go, like}
{go, to}
{go, someplace}
{go, nearby}

Negative
{- target, context}
{go, aardvark}
{go, incubate}
{go, twelve}
{go, therefore}
Setup

Let's represent words as vectors of some length (say 300), randomly initialized.

So we start with $300 \times V$ random parameters.

Over the entire training set, we'd like to adjust those word vectors such that we

- Maximize the similarity of the target word, context word pairs $(t,c)$ drawn from the positive data.
- Minimize the similarity of the $(t,c)$ pairs drawn from the negative data.
The classifier's goal

- Compute the probability that \( c \) is a real context word and not a fake noise word
- \( P(\text{real} \mid t,c) \)
- \( P(\text{fake} \mid t,c) \)
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:
  - $\text{Similarity}(t,c) \propto t \cdot c$
- We’ll need to normalize to get a probability
  - (cosine isn't a probability either)
Turning dot products into probabilities

- \( \text{Sim}(t, c) = t \cdot c \)
- To turn this into a probability.
- We'll use the sigmoid from logistic regression:

\[
P(+|t, c) = \frac{1}{1 + e^{-\text{sim}(t, c)}}
\]
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.
increase similarity(banana, peel)

decrease similarity(banana, Tolstoy)
Objective Criteria

We want to maximize...

\[ \sum_{(t,c) \in +} \log P(+) \mid t, c) + \sum_{(t,c) \in -} \log P(- \mid t, c) \]

Maximize

- the + label for the pairs from the positive training data
- the – label for the pairs sample from the negative data.

Train using stochastic gradient descent
Summary: How to learn word2vec (skip-gram) embeddings

• Start with V random 300-dimensional vectors as initial embeddings
• Use logistic regression:
  • 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.
Or in other words

- Start with some initial embeddings (e.g., random)
- iteratively make the embeddings for a word
  - more like the embeddings of its neighbors
  - less like the embeddings of other words.
Properties of embeddings: Word similarity/relatedness!

Nearest words to some embeddings (Mikolov et al. 2013)

<table>
<thead>
<tr>
<th>target:</th>
<th>Redmond</th>
<th>Havel</th>
<th>ninjutsu</th>
<th>graffiti</th>
<th>capitate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Redmond Wash.</td>
<td>Vaclav Havel</td>
<td>ninja</td>
<td>spray paint</td>
<td>capitulation</td>
</tr>
<tr>
<td></td>
<td>Redmond Washington</td>
<td>president Vaclav Havel</td>
<td>martial arts</td>
<td>grafitti</td>
<td>capitulated</td>
</tr>
<tr>
<td></td>
<td>Microsoft</td>
<td>Velvet Revolution</td>
<td>swordsmanship</td>
<td>taggers</td>
<td>capitulating</td>
</tr>
</tbody>
</table>

As with skip-grams, the projection vector \( h \) is multiplied by the output matrix \( W_0 \). The result \( o = W_0 h \) is a \( 1 \times |V| \) dimensional output vector giving a score for each of the \( |V| \) words. In doing so, the element \( o_k \) was computed by multiplying \( h \) by the output embedding for word \( w_k \):

\[
o_k = v_0^k h
\]

Finally we normalize this score vector, turning the score for each element \( o_k \) into a probability by using the soft-max function.

19.6 Compositionality in Vector Models of Meaning

To be written.
Analogies! Embeddings capture relational meaning!

$$\text{vector('king') - vector('man') + vector('woman') \approx vector('queen')}$$

$$\text{vector('Paris') - vector('France') + vector('Italy') \approx vector('Rome')}$$
Embeddings reflect cultural bias!


Ask “Paris : France :: Tokyo : x”
° x = Japan

Ask “father : doctor :: mother : x”
° x = nurse

Ask “man : computer programmer :: woman : x”
° x = homemaker
Two things we can do about this cultural bias problem

1. Find ways to debias word embeddings
2. Use the embeddings to study cultural bias!
First: embeddings as a window on history


Will Hamilton    Jure Leskovec
Train embeddings on old books to study changes in word meaning
Visualizing changes

Project 300 dimensions down into 2

~30 million books, 1850-1990, Google Books data
The evolution of sentiment words

Negative words change faster than positive words
Historical embedding: a tool to investigate history of cultural biases

Now use our historical embeddings as a tool to investigate cultural biases

- From the historical embeddings
- Compute historical biases of words:
  - **Gender bias**: how much closer a word is to "woman" synonyms than "man" synonyms.
  - **Ethnic bias**: how much closer a word is to last names of a given ethnicity than to names of Anglo ethnicity
- Correlate with occupational data from historical census
- Look at how all these change over time
Embedding bias correlates with actual occupation data

Is "nurse" closer to "man" than "woman"?
Embeddings reflect gender bias in occupations across time (1910-1990)
Embeddings also reflect **framings** of women over time.

Embeddings for **competence** adjectives are biased toward men.

- *Smart, wise, brilliant, intelligent, resourceful, thoughtful, logical, etc.*

This bias is slowly decreasing 1960-1990.
Embeddings reflect ethnic stereotypes over time

- "Princeton trilogy" experiments
- Attitudes toward ethnic groups (1933, 1951, 1969) scores for adjectives
  - industrious, superstitious, nationalistic, etc
- Embedding association with Chinese ethnicity correlates with adjective scores and with the change 1933-1979
- In other words: we can run social psychology experiments in the past!
Changes in Framing: The most biased Asian (vs. White) adjectives over time

<table>
<thead>
<tr>
<th>1910</th>
<th>1950</th>
<th>1990</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irresponsible</td>
<td>Disorganized</td>
<td>Inhibited</td>
</tr>
<tr>
<td>Envious</td>
<td>Outrageous</td>
<td>Passive</td>
</tr>
<tr>
<td>Barbaric</td>
<td>Pompous</td>
<td>Dissolute</td>
</tr>
<tr>
<td>Aggressive</td>
<td>Unstable</td>
<td>Haughty</td>
</tr>
<tr>
<td>Transparent</td>
<td>Effeminate</td>
<td>Complacent</td>
</tr>
<tr>
<td>Monstrous</td>
<td>Unprincipled</td>
<td>Forceful</td>
</tr>
<tr>
<td>Hateful</td>
<td>Venomous</td>
<td>Fixed</td>
</tr>
<tr>
<td>Cruel</td>
<td>Disobedient</td>
<td>Active</td>
</tr>
<tr>
<td>Greedy</td>
<td>Predatory</td>
<td>Sensitive</td>
</tr>
<tr>
<td>Bizarre</td>
<td>Boisterous</td>
<td>Hearty</td>
</tr>
</tbody>
</table>
Conclusion

• **Concepts** or word senses
  • Have a complex many-to-many association with **words** (homonymy, multiple senses)
  • Have relations with each other
    • Synonymy, Antonymy, Superordinate
  • But are hard to define formally (necessary & sufficient conditions)

• **Embeddings** = vector models of meaning
  • More fine-grained than just a string or index
  • Especially good at modeling similarity/analogy
    • Just download them and use cosines!!
  • Can use sparse count models (tf-idf/PPMI) or dense predict models (word2vec, GLoVE)
  • Useful in practice but also encode cultural stereotypes
    • Can **debias** embeddings, and use them to **study bias**
Embeddings in classes

- CS224U: Next Quarter
- CS224N: Winter 2019