

Ling/CSE 472: Introduction to Computational Linguistics

5/16: Vector Semantics

Overview

- Meaning of words
 - Relations between meanings of words
- Vector representations
- tf-idf & PPMI (Slides borrowed from Sara Ng)
- Reading questions

Word meanings

- Synonymy
- Similarity
- Semantic field
- Semantic frames & roles
- Connotation



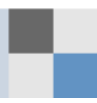

W Pairs of synonyms



Total Results: 0

Powered by  **Poll Everywhere**

Start the presentation to see live content. For screen share software, share the entire screen. Get help at pollev.com/app





Pairs of non-synonymous related words

Total Results: 0

Powered by  **Poll Everywhere**

Start the presentation to see live content. For screen share software, share the entire screen. Get help at pollev.com/app



Semantic field: College

Total Results: 0

Powered by  **Poll Everywhere**

Start the presentation to see live content. For screen share software, share the entire screen. Get help at pollev.com/app



Word meanings

- Synonymy: WordNet - <http://globalwordnet.org/resources/wordnets-in-the-world/>
- Similarity
- Semantic field
- Semantic frames & roles: FrameNet - <https://www.globalframenet.org/>
- Connotation

Word meanings: What else is missing?

Word meanings: What else is missing?

- Qualia structure (Pustejovsky & Jezek 2016, p7)
 - *Formal*: encoding taxonomic information about the lexical item (the `is-a` relation);
 - *Constitutive*: encoding information on the parts and constitution of an object (`part-of` or `made-of` relation);
 - *Telic*: encoding information on purpose and function (the *used-for* or `functions-as` relation);
 - *Agentive*: encoding information about the origin of the object (the `created-by` relation).

Word meanings: What else is missing?

- Qualia structure (Pustejovsky & Jezek 2016, p9)

- (12) a. He owns a two-story house. (house as artifact (F))
b. Lock your house when you leave. (part of house, door (C))
c. We bought a comfortable house. (purpose of house (T))
d. The house is finally finished. (origin of house (A))

$$(13) \left[\begin{array}{l} \textit{house} \\ \text{QUALIA} = \left[\begin{array}{l} \text{F} = \mathbf{\text{building}} \\ \text{C} = \{\mathbf{\text{door, rooms, ...}}\} \\ \text{T} = \mathbf{\text{live_in}} \\ \text{A} = \mathbf{\text{build}} \end{array} \right] \end{array} \right]$$

Word meanings: What else is missing?

- Lexical entailments, such as factivity
 - Kim knows that it is hot outside
 - Kim believes that it is hot outside
- More detailed approach: CommitmentBank (de Marneffe et al 2019)

Word meanings: What else is missing?

- Register/formality: What does word usage say about the current situation?
- Index of social address: What does word usage say about the speaker?
- Are these the same thing as connotation?

Overview

- Meaning of words
 - Relations between meanings of words
- Vector representations
- tf-idf & PPMI (Slides borrowed from Sara Ng)
- Reading questions

Vector representations

- Model words in terms of:
 - What documents they occur in
 - What other words they occur with
- What properties of words does this model approximate?
- What properties does it miss?
- Why/when are they useful (anyway)?

Simple counts: term-document, term-term

	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

Figure 6.2 The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document.

	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	...

Figure 6.6 Co-occurrence vectors for four words in the Wikipedia corpus, showing six of the dimensions (hand-picked for pedagogical purposes). The vector for *digital* is outlined in red. Note that a real vector would have vastly more dimensions and thus be much sparser.

(J&M Ch 6)

Normalized counts: tf-idf & PPMI

TF-IDF

Term Frequency – Inverse Document Frequency, or **TF-IDF** for short, is a common baseline model for embedding words. Each cell in a TF-IDF matrix is calculated as:

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

- ⊗ In TF-IDF, words (a.k.a. *terms*) are represented by a simple function of the *counts* of nearby words, given a corpus of documents
- ⊗ TF-IDF is a staple in information retrieval (IR)

Term frequency (tf)

The **term frequency** for word t is the number of times t appears in document d :

$$tf_{t,d} = \text{count}(t, d)$$

- ⊙ We often log weight term frequencies to squash raw frequencies

$$tf_{t,d} = \log_{10}(\text{count}(t, d) + 1)$$

- Intuition: If a word appears 100 times in a document, this doesn't mean that word is 100× more relevant to the document
- (We add 1 to each count since it's not possible to calculate the log of 0.)

Document frequency (df)

The **document frequency** of a word t is the number of documents in which t occurs.

- ⊕ Note that **term frequency** \neq **collection frequency**
- ⊕ A word's **collection frequency** is the total number of times a word appears across an entire corpus

○ E.g., the

	Collection Frequency	Document Frequency
Romeo	113	1
action	113	31

e

Shakespeare play:

Inverse document frequency (idf)

The importance of a word for a given document is emphasized by taking the **inverse** of the its document frequency:

$$\text{idf}_t = \log_{10} \left(\frac{N}{\text{df}_t} \right)$$

where N is the total number of documents in the corpus.

Word	df	idf
Romeo	1	1.57
salad	2	1.27
Falstaff	4	0.967
forest	12	0.489
battle	21	0.246
wit	34	0.037
fool	36	0.012
good	37	0
sweet	37	0

- ⊕ idf_t is *higher* when t appears in *fewer* documents

TF-IDF: The big idea

- ⊙ Vanilla term-document matrices consist purely of $tf_{t,d}$ values
- ⊙ TF-IDF *weights* a word's raw frequency in a document ($tf_{t,d}$) by its inverse document frequency (idf_t)

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


- ⊙ In summary, TF-IDF plays up a word's importance to a document when that word appears in relatively fewer documents overall

	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

Figure 6.2 The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.074	0	0.22	0.28
good	0	0	0	0
fool	0.019	0.021	0.0036	0.0083
wit	0.049	0.044	0.018	0.022

Figure 6.9 A tf-idf weighted term-document matrix for four words in four Shakespeare plays, using the counts in Fig. 6.2. For example the 0.049 value for *wit* in *As You Like It* is the product of $tf = \log_{10}(20 + 1) = 1.322$ and $idf = .037$. Note that the idf weighting has eliminated the importance of the ubiquitous word *good* and vastly reduced the impact of the almost-ubiquitous word *fool*.

From raw counts to TF-IDF

PMI

Pointwise mutual information (PMI) is used to measure to what extent two words, w_1 and w_2 , are more likely to co-occur than by chance.

$$\text{PMI}(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

- ⊗ PMI is a measure of **association** from information theory
- ⊗ If w_1 and w_2 are independent, then $P(w_1, w_2) = P(w_1)P(w_2)$
- ⊗ PMI values range from $-\infty$ to ∞

Computing PMI

Assume we have a term-context matrix F with W rows (i.e., words) and C columns (i.e., contexts), where $f_{i,j}$ gives the number of times word w_i co-occurs with context word c_j .

	computer	data	result	pie	sugar	count(w)
cherry	2	8	9	442	25	486
strawberry	0	0	1	60	19	80
digital	1670	1683	85	5	4	3447
information	3325	3982	378	5	13	7703
count(context)	4997	5673	473	512	61	11716

Computing PMI (cont'd)

$$\text{PMI}(w_i, c_j) = \log_2 \frac{p(w_i, c_j)}{p(w_i)p(c_j)}$$

$$p(w_i, c_j) = \frac{f_{ij}}{\sum_{i=1}^W \sum_{j=1}^C f_{ij}} = \frac{\text{count}(w_i, c_j)}{\text{total co-occurrences}}$$

$$p(w_i) = \frac{\sum_{j=1}^C f_{ij}}{\sum_{i=1}^W \sum_{j=1}^C f_{ij}} = \frac{\text{\# of times } w_i \text{ co-occurs with a context word}}{\text{total co-occurrences}}$$

$$p(c_j) = \frac{\sum_{i=1}^W f_{ij}}{\sum_{i=1}^W \sum_{j=1}^C f_{ij}} = \frac{\text{\# of times } c_j \text{ co-occurs with a term word}}{\text{total co-occurrences}}$$

Computing PMI (cont'd)

$$\text{PMI}(w_i, c_j) = \log_2 \frac{p(w_i, c_j)}{p(w_i)p(c_j)}$$

$$p(w_i, c_j) = \frac{f_{ij}}{\sum_{i=1}^W \sum_{j=1}^C f_{ij}} = \frac{\text{count}(w_i, c_j)}{\text{total co-occurrences}}$$

$$p(w_i) = \frac{\sum_{j=1}^C f_{ij}}{\sum_{i=1}^W \sum_{j=1}^C f_{ij}} = \frac{\text{\# of times } w_i \text{ co-occurs with a context word}}{\text{total co-occurrences}}$$

$$p(c_j) = \frac{\sum_{i=1}^W f_{ij}}{\sum_{i=1}^W \sum_{j=1}^C f_{ij}} = \frac{\text{\# of times } c_j \text{ co-occurs with a term word}}{\text{total co-occurrences}}$$

Computing PMI (cont'd)

$$\text{PMI}(w_i, c_j) = \log_2 \frac{p(w_i, c_j)}{p(w_i)p(c_j)}$$

$$p(w_i, c_j) = \frac{f_{ij}}{\sum_{i=1}^W \sum_{j=1}^C f_{ij}} = \frac{\text{count}(w_i, c_j)}{\text{total co-occurrences}}$$

$$p(w_i) = \frac{\sum_{j=1}^C f_{ij}}{\sum_{i=1}^W \sum_{j=1}^C f_{ij}} = \frac{\text{\# of times } w_i \text{ co-occurs with a context word}}{\text{total co-occurrences}}$$

$$p(c_j) = \frac{\sum_{i=1}^W f_{ij}}{\sum_{i=1}^W \sum_{j=1}^C f_{ij}} = \frac{\text{\# of times } c_j \text{ co-occurs with a term word}}{\text{total co-occurrences}}$$

Computing PMI (cont'd)

$$\text{PMI}(w_i, c_j) = \log_2 \frac{p(w_i, c_j)}{p(w_i)p(c_j)}$$

$$p(w_i, c_j) = \frac{f_{ij}}{\sum_{i=1}^W \sum_{j=1}^C f_{ij}} = \frac{\text{count}(w_i, c_j)}{\text{total co-occurrences}}$$

$$p(w_i) = \frac{\sum_{j=1}^C f_{ij}}{\sum_{i=1}^W \sum_{j=1}^C f_{ij}} = \frac{\text{\# of times } w_i \text{ co-occurs with a context word}}{\text{total co-occurrences}}$$

$$p(c_j) = \frac{\sum_{i=1}^W f_{ij}}{\sum_{i=1}^W \sum_{j=1}^C f_{ij}} = \frac{\text{\# of times } c_j \text{ co-occurs with a term word}}{\text{total co-occurrences}}$$

PMI example

$$p(\text{information, data}) = \frac{3,982}{11,716} = .3399$$

$$p(\text{information}) = \frac{7,703}{11,716} = .6575$$

$$p(\text{data}) = \frac{5,673}{11,716} = .4842$$

	computer	data	result	pie	sugar	count(w)
cherry	2	8	9	442	25	486
strawberry	0	0	1	60	19	80
digital	1670	1683	85	5	4	3447
information	3325	3982	378	5	13	7703
count(context)	4997	5673	473	512	61	11716

$$\text{PMI}(\text{information, data}) = \log_2 \frac{.3399}{(.6575)(.4842)} = 0.0944$$

PMI drawbacks

Recall that PMI values range from **negative** infinity to **positive** infinity.

Negative PMI values imply that w_1 and w_2 co-occur *less often* than if by chance, but they are problematic:

- ⊕ It's not clear humans are good at judging "unrelatedness"
- ⊕ Negative PMI values are only reliable with *enormous* corpora
 - Imagine w_1 and w_2 whose probability is each 10^{-6}
 - The probability of them co-occurring by chance is 10^{-12}
 - We need *lots* of data to be sure $p(w_1, w_2)$ is significantly different than 10^{-12}

Solution: PPMI

Positive PMI (PPMI) simply replaces negative PMI values with 0:

$$\text{PPMI}(w, c) = \max \left(\log_2 \frac{P(w, c)}{P(w)P(c)}, 0 \right)$$

	computer	data	result	pie	sugar	count(w)
cherry	2	8	9	442	25	486
strawberry	0	0	1	60	19	80
digital	1670	1683	85	5	4	3447
information	3325	3982	378	5	13	7703
count(context)	4997	5673	473	512	61	11716

Figure 6.10 Co-occurrence counts for four words in 5 contexts in the Wikipedia corpus, together with the marginals, pretending for the purpose of this calculation that no other words/contexts matter.

	computer	data	result	pie	sugar
cherry	0	0	0	4.38	3.30
strawberry	0	0	0	4.10	5.51
digital	0.18	0.01	0	0	0
information	0.02	0.09	0.28	0	0

Figure 6.12 The PPMI matrix showing the association between words and context words, computed from the counts in Fig. 6.11. Note that most of the 0 PPMI values are ones that had a negative PMI; for example $PMI(cherry, computer) = -6.7$, meaning that *cherry* and *computer* co-occur on Wikipedia less often than we would expect by chance, and with PPMI we replace negative values by zero.

From raw counts to PPMI

Overview

- Meaning of words
 - Relations between meanings of words
- Vector representations
- tf-idf & PPMI (Slides borrowed from Sara Ng)
- Reading questions

Reading questions

- I feel like even though yes, a lot of words could probably be 'similar' to each other just because of their placement in comparisons with other words, this may not apply to other words farther than just, say, their part-of-speech. Like the words 'bird' and 'car' could probably show up in a lot of the same places (Ex: this is a bird/car), but they basically have no relation to each other aside from the fact that they both happened to be subjects in these cases (which to me doesn't mean they are 'similar' - not in meaning anyway).

Reading questions

- Question: It makes a lot of sense why synonyms might show up in similar contexts, but I imagine that antonyms probably also show up in a similar context, albeit with an added negation word (e.g. "Antarctica is cold. Antarctica is not hot.) Can this result in these kinds of words showing up in a similar location in the vector space?

Reading questions

- We discussed previously how antonyms aren't really opposite words of each other (hot and cold are both temperature measures). Could we use the vector space representation to find words that are true 'opposites' of each other, in the sense that the two words have the furthest possible embedded meaning relationships from each other? Not sure what that might be useful for, just curious.

Reading questions

- How exactly do vector semantics deal with negation? In the reading we can see that it correctly designates "not good" as negative. Is the process the same or different for any other sentiment analysis?
- Can this kind of lexical semantics keep up with things like sarcasm? I would imagine not but just curious.

Reading questions

- Would vector semantics work differently at all with words that aren't nouns/verbs/adjectives? Some words, like “as,” and “than” don't really carry a lot of meaning on their own.
- I noticed the cluster of “to,” “now,” “that,” “you,” “is,” etc. in Figure 6.1 (two dimensional projection). Are these words clustered together in the image simply because they are neither positive nor negative? Where would other words that are neither positive or negative like “surprising” and “shocking” fit in on this projection?



Figure 6.1 A two-dimensional (t-SNE) projection of embeddings for some words and phrases, showing that words with similar meanings are nearby in space. The original 60-dimensional embeddings were trained for sentiment analysis. Simplified from [Li et al. \(2015\)](#) with colors added for explanation.

(J&M Ch 6)

Reading questions

- How does vector semantics deal with lemmas where one wordform has multiple very different senses? i.e. mouse (the rodent) and mouse (for computers); since they occur in different contexts with different surrounding words, would there be two separate labels for mouse? Or one label which is spread thin by the various meanings?

Reading questions

- The reading on word embedding clarified a lot on a package I was reading about. However, that package produce a word embedding in the form of m by n matrix full of floats per word. As I understand from the book, the word embedding is usually a vector and the vector in around the same area in the space will have similarities between them. How does this transfer to a matrix word embedding or this is impossible and I read something wrong?

Reading questions

- Vector semantics are introduced in the context of the distributional hypothesis. It is not true that the human brain is computing probabilities and taking in statistics of natural language everyday? Are our grammaticality judgments not based purely on the frequency of the structures of sentences we hear around us everyday, the most frequent forms being the acceptable ones vs the least frequent/infrequent forms being the unacceptable ones? Are our understandings of words not purely based on the contexts they appear in, just like in word embeddings? Where does the human level of understanding of words' meanings and grammaticality come from?
- If computers use human-crafted language and make statistical judgements and create probabilities based on this data to replicate natural language, what are we doing? We do that and more, right? What is the more?
- => Baria & Cross 2021 <https://arxiv.org/abs/2107.14042>

Reading questions

- Is there ever a time when having more information could be a bad thing? I'm not sure that I can think of a way that it would be, but for example, the reading associated the word information with digital rather than cherry, as you would expect! However, could some data ever be tainted or biased to an extent that word associations produce unfavorable results in terms of how accurate to reality the associations it makes are?

NLP/Compling in the news

- <https://www.cnbc.com/2023/05/16/openai-ceo-woos-lawmakers-ahead-of-first-testimony-before-congress.html>
- <https://www.scientificamerican.com/article/how-ai-knows-things-no-one-told-it/>