

Word Meaning and Similarity

Word Similarity: Distributional Similarity (I)



Problems with thesaurus-based meaning

- We don't have a thesaurus for every language
- Even if we do, they have problems with **recall**
 - Many words are missing
 - Most (if not all) phrases are missing
 - Some connections between senses are missing
 - Thesauri work less well for verbs, adjectives
 - Adjectives and verbs have less structured hyponymy relations



Distributional models of meaning

- Also called vector-space models of meaning
- Offer much higher recall than hand-built thesauri
 - Although they tend to have lower precision
- Zellig Harris (1954): “**oculist** and **eye-doctor** ... occur in almost the same environments....
If A and B have almost identical environments we say that they are synonyms.
- Firth (1957): “You shall know a word by the
3 company it keeps!”



Intuition of distributional word similarity

- Nida example:

A bottle of *tesgüino* is on the table
Everybody likes *tesgüino*
Tesgüino makes you drunk
We make *tesgüino* out of corn.

- From context words humans can guess *tesgüino* means
 - an alcoholic beverage like **beer**
- Intuition for algorithm:
 - Two words are similar if they have similar word contexts.



Reminder: Term-document matrix

- Each cell: count of term t in a document d : $tf_{t,d}$:
 - Each document is a **count vector** in \mathbb{N}^v : a column below

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	6	117	0	0



Reminder: Term-document matrix

- Two documents are similar if their vectors are similar

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	6	117	0	0



The words in a term-document matrix

- Each word is a **count vector** in \mathbb{N}^D : a row below

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	6	117	0	0



The words in a term-document matrix

- Two **words** are similar if their vectors are similar

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	6	117	0	0



The Term-Context matrix

- Instead of using entire documents, use smaller contexts
 - Paragraph
 - Window of 10 words
- A word is now defined by a vector over counts of context words



Sample contexts: 20 words (Brown corpus)

- equal amount of sugar, a sliced lemon, a tablespoonful of **apricot** preserve or jam, a pinch each of clove and nutmeg,
 - on board for their enjoyment. Cautiously she sampled her first **pineapple** and another fruit whose taste she likened to that of
 - of a recursive type well suited to programming on the **digital** computer. In finding the optimal R-stage policy from that of
 - substantially affect commerce, for the purpose of gathering data and **information** necessary for the
- 10 study authorized in the first section of this



Term-context matrix for word similarity

- Two **words** are similar in meaning if their context vectors are similar

	aardvark	computer	data	pinch	result	sugar	...
apricot	0	0	0	1	0	1	
pineapple	0	0	0	1	0	1	
digital	0	2	1	0	1	0	
information	0	1	6	0	4	0	



Should we use raw counts?

- For the term-document matrix
 - We used **tf-idf** instead of raw term counts
- For the term-context matrix
 - **Positive Pointwise Mutual Information (PPMI)** is common



Pointwise Mutual Information

- **Pointwise mutual information:**

- Do events x and y co-occur more than if they were independent?

$$\text{PMI}(X, Y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

- **PMI between two words:** (Church & Hanks 1989)

- Do words x and y co-occur more than if they were independent?

$$\text{PMI}(\text{word}_1, \text{word}_2) = \log_2 \frac{P(\text{word}_1, \text{word}_2)}{P(\text{word}_1)P(\text{word}_2)}$$

- **Positive PMI between two words** (Niwa & Nitta 1994)

- Replace all PMI values less than 0 with zero



Computing PPMI on a term-context matrix

- Matrix F with W rows (words) and C columns (contexts)
- f_{ij} is # of times w_i occurs in context c_j

	aardvark	computer	data	pinch	result	sugar
apricot	0	0	0	1	0	1
pineapple	0	0	0	1	0	1
digital	0	2	1	0	1	0
information	0	1	6	0	4	0

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^W \sum_{j=1}^C f_{ij}} \quad p_{i*} = \frac{\sum_{j=1}^C f_{ij}}{\sum_{i=1}^W \sum_{j=1}^C f_{ij}} \quad p_{*j} = \frac{\sum_{i=1}^W f_{ij}}{\sum_{i=1}^W \sum_{j=1}^C f_{ij}}$$

$$pmi_{ij} = \log_2 \frac{p_{ij}}{p_{i*} p_{*j}} \quad ppmi_{ij} = \begin{cases} pmi_{ij} & \text{if } pmi_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$



$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^W \sum_{j=1}^C f_{ij}}$$

apricot
pineapple
digital
information

Count(w,context)

computer	data	pinch	result	sugar
0	0	1	0	1
0	0	1	0	1
2	1	0	1	0
1	6	0	4	0

$$p(w=\text{information}, c=\text{data}) = 6/19 = .32$$

$$p(w=\text{information}) = 11/19 = .58$$

$$p(c=\text{data}) = 7/19 = .37$$

$$p(w_i) = \frac{\sum_{j=1}^C f_{ij}}{N}$$

$$p(c_j) = \frac{\sum_{i=1}^W f_{ij}}{N}$$

p(w,context)

p(w)

	computer	data	pinch	result	sugar	
apricot	0.00	0.00	0.05	0.00	0.05	0.11
pineapple	0.00	0.00	0.05	0.00	0.05	0.11
digital	0.11	0.05	0.00	0.05	0.00	0.21
information	0.05	0.32	0.00	0.21	0.00	0.58
p(context)	0.16	0.37	0.11	0.26	0.11	



$$pmi_{ij} = \log_2 \frac{p_{ij}}{p_i * p_j}$$

	p(w,context)					p(w)
	computer	data	pinch	result	sugar	
apricot	0.00	0.00	0.05	0.00	0.05	0.11
pineapple	0.00	0.00	0.05	0.00	0.05	0.11
digital	0.11	0.05	0.00	0.05	0.00	0.21
information	0.05	0.32	0.00	0.21	0.00	0.58
p(context)	0.16	0.37	0.11	0.26	0.11	

- $pmi(\text{information}, \text{data}) = \log_2 (.32 / (.37 * .58)) = .58$

(.57 using full precision)

	PPMI(w,context)				
	computer	data	pinch	result	sugar
apricot	-	-	2.25	-	2.25
pineapple	-	-	2.25	-	2.25
digital	1.66	0.00	-	0.00	-
information	0.00	0.57	-	0.47	-



Weighing PMI

- PMI is biased toward infrequent events
- Various weighting schemes help alleviate this
 - See Turney and Pantel (2010)
- Add-one smoothing can also help



	Add-2 Smoothed Count(w,context)				
	computer	data	pinch	result	sugar
apricot	2	2	3	2	3
pineapple	2	2	3	2	3
digital	4	3	2	3	2
information	3	8	2	6	2

	p(w,context) [add-2]					p(w)
	computer	data	pinch	result	sugar	
apricot	0.03	0.03	0.05	0.03	0.05	0.20
pineapple	0.03	0.03	0.05	0.03	0.05	0.20
digital	0.07	0.05	0.03	0.05	0.03	0.24
information	0.05	0.14	0.03	0.10	0.03	0.36
p(context)	0.19	0.25	0.17	0.22	0.17	



PPMI(w,context)

	computer	data	pinch	result	sugar
apricot	-	-	2.25	-	2.25
pineapple	-	-	2.25	-	2.25
digital	1.66	0.00	-	0.00	-
information	0.00	0.57	-	0.47	-

PPMI(w,context) [add-2]

	computer	data	pinch	result	sugar
apricot	0.00	0.00	0.56	0.00	0.56
pineapple	0.00	0.00	0.56	0.00	0.56
digital	0.62	0.00	0.00	0.00	0.00
information	0.00	0.58	0.00	0.37	0.00



Word Meaning and Similarity

Word Similarity: Distributional Similarity (I)