CS918: LECTURE 10

Word Senses and Similarity

Arkaitz Zubiaga, 5th November, 2018

LECTURE 10: CONTENTS

- Word Senses: Concepts.
- Thesauri: Wordnet.
- Computing Word Similarity.
  - Thesaurus Methods.
  - Distributional Models of Similarity.
  - Evaluation.

WORD SENSES: CONCEPTS

HOMONYMY

- **Homonymy:** same word can have **different, unrelated meanings**:
  - I put my money in the **bank$_1$**.
  - We took the picture in the east **bank$_2$** of the Nile river.

- **bank$_1$**: financial institution.
- **bank$_2$**: sloping land.

POLYSEMY

- **Polysemy:** same word, **related meanings**:
  - The **bank**$_1$ was constructed in 1875 out of local red brick.
  - I withdrew the money from the **bank**$_2$.

- **bank**$_1$: the building belonging to a financial institution.
- **bank**$_2$: a financial institution.

SYSTEMATIC POLYSEMY

WARWICK

- **Polysemy** is often **systematic**:
  - **Building, people & organisation:**
    - I'm heading to the **university**. [location, campus]
    - The **university** has gone on strike. [its staff]
    - The **university** is ranked 10th. [the organisation]
  - **Author & work:**
    - **Shakespeare** wrote nice stuff. [himself]
    - I love reading **Shakespeare**. [his books]

SYNONYMY

- **Synonyms:** words with same meaning in some or all contexts.
  - filbert / hazelnut
  - couch / sofa
  - big / large

SYNONYMY

- But there are few (or no) examples of perfect synonymy.

- e.g. big/large:
  - My **big** sister… [= older sister]
  - My **large** sister… [≠ older sister]

ANTONYMY

WARWICK

- **Antonyms:** Senses that are **opposites with respect to one feature of meaning**, very similar otherwise:
  - dark/light.
  - short/long.
  - fast/slow.
  - hot/cold.
  - up/down.

HYPONYMY AND HYPERNYMY

- One sense is a **hyponym** of another if the first sense is **more specific**, denoting a **subclass** of the other:
  - car is a hyponym of vehicle
  - apple is a hyponym of fruit
- Conversely **hypernym**:
  - vehicle is a hypernym of car
  - fruit is a hypernym of apple

INSTANCES VS HYPONYMS

- An **instance** is an individual, a proper noun that is a unique entity:
  - London is an **instance** of city.

- But city is a **class**:
  - city is a **hyponym** of municipality or location.

THESAURI

THESAURI

- **Thesaurus** (plural **thesauri**) is a reference work that lists **words grouped together according to similarity of meaning**.

- Useful for different tasks (some of which we'll see in next lectures):
    - Information Extraction.
    - Information Retrieval.
    - Question Answering.
    - Machine Translation.

WORDNET 3.1

WARWICK

- Wordnet is a popular dataset (thesaurus + aspects of dictionary): https://wordnet.princeton.edu/
- It's integrated in NLTK: http://www.nltk.org/howto/wordnet.html
- Structured into 117,000 **synsets** (synonym sets).
  - Linked to other synsets through "**conceptual relations**".
  - Synsets have brief definition ("**gloss**") and example sentences.

SENSES OF "BASS" IN WORDNET

WARWICK

Noun

- S: (n) **bass** (the lowest part of the musical range)
- S: (n) **bass**, bass part (the lowest part in polyphonic music)
- S: (n) **bass**, basso (an adult male singer with the lowest voice)
- S: (n) sea bass, **bass** (the lean flesh of a saltwater fish of the family Serranidae)
- S: (n) freshwater bass, **bass** (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- S: (n) **bass**, bass voice, basso (the lowest adult male singing voice)
- S: (n) **bass** (the member with the lowest range of a family of musical instruments)
- S: (n) **bass** (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Adjective

- S: (adj) **bass**, deep (having or denoting a low vocal or instrumental range) *"a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"*

SYNSETS: SYNONYM SETS

- Example: **chump** as a noun with the **gloss**:
  - "a person who is gullible and easy to take advantage of"

- This **sense** of "**chump**" is **shared by 9 words**:
  - chump_1, fool_2, gull_1, mark_9, patsy_1, fall guy_1, sucker_1, soft touch_1, mug_2
  - Note: **only those senses of the words**, e.g. mug_1 is a "cup", belongs to a different synset.

WORDNET HYPERNYM HIERARCHY FOR "BASS"

- S: (n) **bass**, basso (an adult male singer with the lowest voice)
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) singer, vocalist, vocalizer, vocaliser (a person who sings)
      - S: (n) musician, instrumentalist, player (someone who plays a musical instrument (as a profession))
        - S: (n) performer, performing artist (an entertainer who performs a dramatic or musical work for an audience)
          - S: (n) entertainer (a person who tries to please or amuse)
            - S: (n) person, individual, someone, somebody, mortal, soul (a human being) *"there was too much for one person to do"*
              - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
                - S: (n) living thing, animate thing (a living (or once living) entity)
                  - S: (n) whole, unit (an assemblage of parts that is regarded as a

WORDNET NOUN RELATIONS

Relation	Also called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	$breakfast^1 \rightarrow meal^1$
Hyponym	Subordinate	From concepts to subtypes	$meal^1 \rightarrow lunch^1$
Member Meronym	Has-Member	From groups to their members	$faculty^2 \rightarrow professor^1$
Has-Instance		From concepts to instances of the concept	$composer^1 \rightarrow Bach^1$
Instance		From instances to their concepts	$Austen^1 \rightarrow author^1$
Member Holonym	Member-Of	From members to their groups	$copilot^1 \rightarrow crew^1$
Part Meronym	Has-Part	From wholes to parts	$table^2 \rightarrow leg^3$
Part Holonym	Part-Of	From parts to wholes	$course^7 \rightarrow meal^1$
Antonym		Opposites	$leader^1 \rightarrow follower^1$

THESAURUS METHODS

WORD SIMILARITY AND RELATEDNESS

- **Synonymy** is **binary**.
- **Similarity** is a looser metric, two words are similar when they **share features of meaning**.
  - **bank**$_1$ is similar to **fund**$_3$.
  - **bank**$_2$ is similar to **slope**$_5$.
- **Relatedness** measures possible **associations**.
  - **motorbike** and **bike** are **similar**.
  - **motorbike** and **fuel** are **related, not similar**.

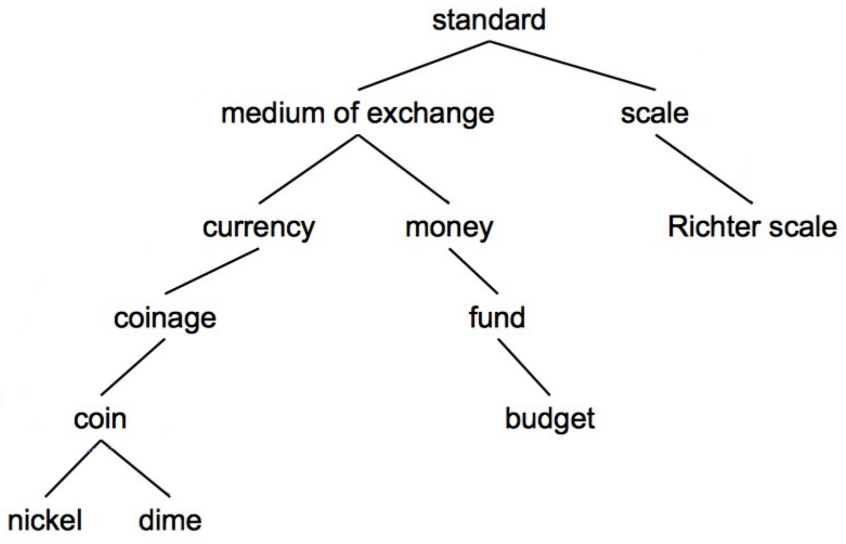
TWO TYPES OF SIMILARITY ALGORITHMS

- **Thesaurus-based algorithms:**
  - Are words “nearby” in hypernym hierarchy?
  - Do words have similar glosses (definitions)?

- **Distributional algorithms:**
  - Do words have similar distributional contexts?

PATH-BASED SIMILARITY

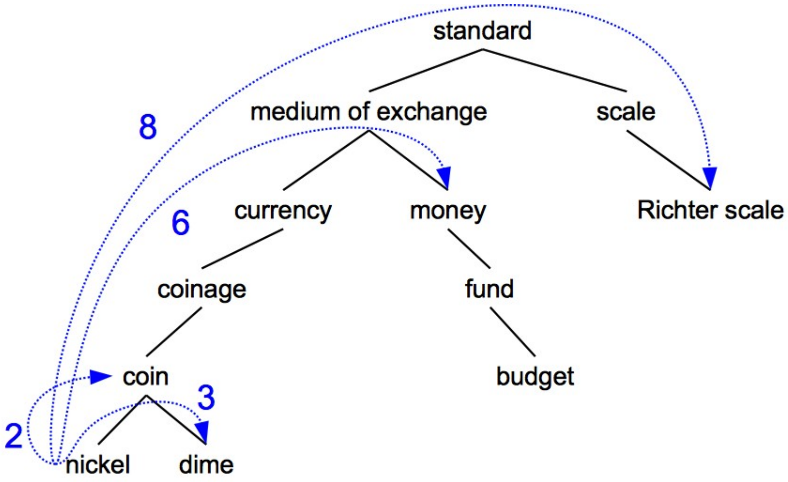
- **Two senses** are **similar if there is a short path** between them in the thesaurus hierarchy



PATH-BASED SIMILARITY

WARWICK

- pathlen(c1, c2) =
  1 + # of edges in shortest path
- simpath(c1,c2) = $\dfrac{1}{\text{pathlen}(c_1, c_2)}$
- simpath(nickel, coin) = 1/2 = .5
- simpath(nickel, Richter scale) = 1/8 = .125



PATH-BASED SIMILARITY

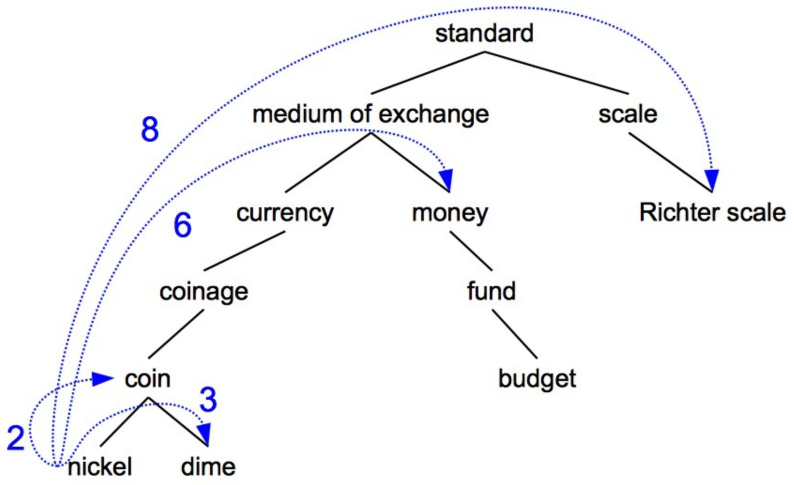
WARWICK

- **Two senses** are **similar if there is a short path** between them in the thesaurus hierarchy

- pathlen(c1,c2) = 1 + # of edges in shorte

- simpath(c1,c2) = $\frac{1}{\text{pathlen}(c_1, c_2)}$

- simpath(nickel,coin) = 1/2 = .5
- simpath(nickel,Richter scale) = 1/8 = .125



Problem: assumes uniform distance for all links.
simpath(nickel, money) = 1/7
simpath(nickel, standard) = 1/7

ALTERNATIVE THESAURUS-BASED SIMILARITY

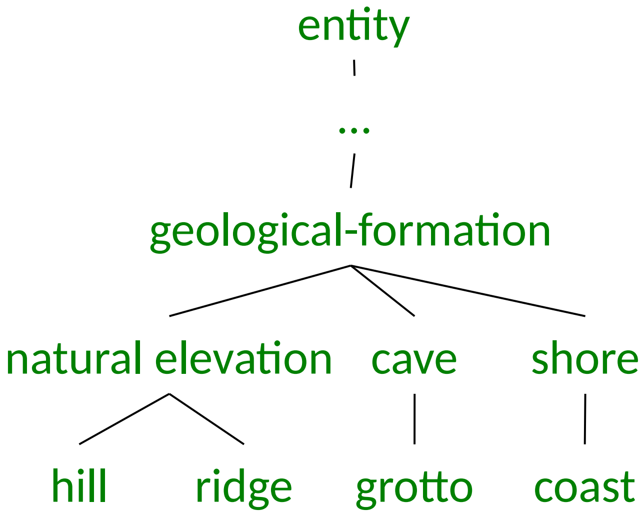
WARWICK

- Thesaurus-based similarity algorithms:
  - Information content for similarity measurement (Resnik).
  - Lin similarity function.
  - The Lesk algorithm.

INFORMATION CONTENT

- Train by **counting in a corpus**.
- Each **instance** of **hill counts towards frequency of its hypernyms**:
  natural elevation, geological formation, etc
- **words(c): set of words children of node c**

words(“geo-formation”) = {hill,ridge,grotto,coast,cave,shore,natural elevation}
words(“natural elevation”) = {hill, ridge}



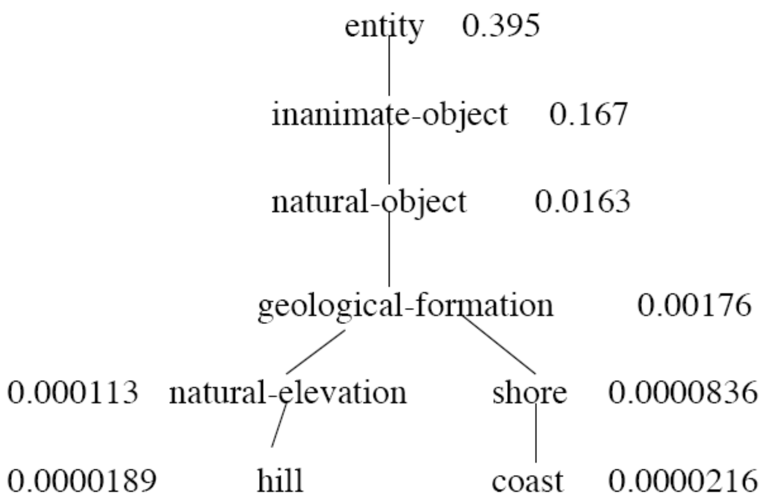
$$P(c) = \frac{\sum_{w \in words(c)} count(w)}{N}$$

→ Probability that a random word in the corpus is instance of c

(Resnik 1995)

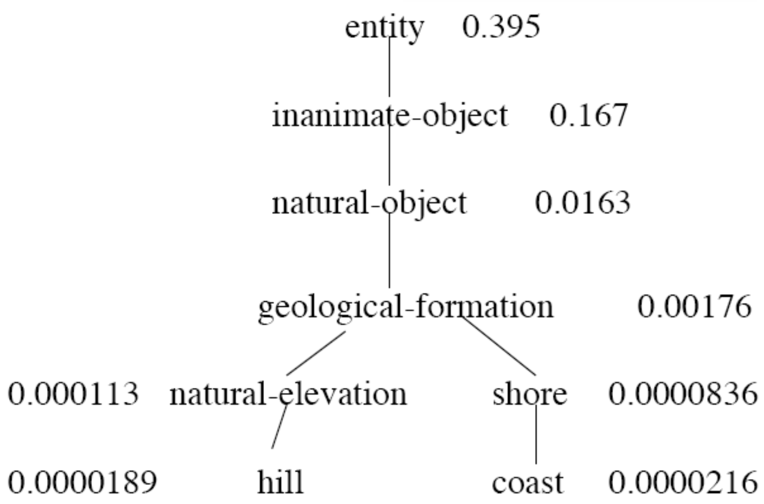
INFORMATION CONTENT

- **Information content**:
  - IC(c) = -log P(c)



LOWEST COMMON SUBSUMMER

- **Lowest common subsumer**:
  - LCS(c_1, c_2)



- The **most informative (lowest) node in the hierarchy subsuming both c_1 and c_2**

  e.g. LCS(hill, coast) = geological-formation

INFORMATION CONTENT FOR SIMILARITY

WARWICK

- **Intuition:** the more two words have in common, the more similar they are.

  i.e. the more infrequent their LCS is, the more similar they are.

  or… the lower their LCS is in the hierarchy, the more similar they are.

(Resnik 1995, 1999)

INFORMATION CONTENT FOR SIMILARITY

WARWICK

- Resnik's similarity:

similarity of c_1 and c_2

= information content of the lowest common subsumer (LCS) of the two nodes.

$$\text{sim}_{\text{resnik}}(c_1, c_2) = \text{IC}(\text{LCS}(c_1, c_2)) = -\log P(\text{LCS}(c_1, c_2))$$

(Resnik 1995, 1999)

DEKANG LIN METHOD

WARWICK

- **Intuition:** look at both commonalities and differences to measure similarity of words A and B.
  - **Commonality:** the more A and B have in common, the more similar they are.
    IC(common(A,B))
  - **Difference:** the more differences between A and B, the less similar.
    IC(description(A,B)-IC(common(A,B))

(Lin 1998)

DEKANG LIN SIMILARITY THEOREM

- **Similarity between A and B** is measured by the ratio between:
  - the **amount of information** needed to state the **commonality of A and B**.
  - the **information needed** to fully **describe what A and B** are.

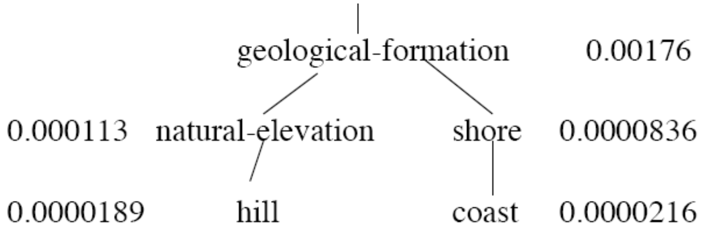
$$sim_{Lin}(A,B) \propto \frac{IC(common(A,B))}{IC(description(A,B))}$$

- Lin defines IC(common(A,B)) as 2 x information of the LCS.

$$sim_{Lin}(c_1,c_2) = \frac{2\log P(LCS(c_1,c_2))}{\log P(c_1) + \log P(c_2)}$$

EXAMPLE OF LIN SIMILARITY FUNCTION

$$sim_{Lin}(A,B)=\frac{2\log P(LCS(c_1,c_2))}{\log P(c_1)+\log P(c_2)}$$



$$sim_{Lin}(\text{hill},\text{coast})=\frac{2\log P(\text{geological-formation})}{\log P(\text{hill})+\log P(\text{coast})}$$

$$=\frac{2\log 0.00176}{\log 0.0000189+\log 0.0000216}$$

$$=.59$$

THE LESK ALGORITHM

WARWICK

- **Intuition:** A and B are similar if their **glosses** contain similar words.
  - **Drawing paper:** **paper** that is **specially prepared** for use in drafting.
  - **Decal:** the art of transferring designs from **specially prepared** **paper** to a wood or glass or metal surface.
- For each word phrase of length n that's in both glosses:
  - Add a score of n^2 (paper and specially prepared: $1^2 + 2^2 = 5$).
  - Compute overlap also for glosses of hypernyms and hyponyms.

(Lesk 1986)

DISTRIBUTIONAL MODELS OF SIMILARITY

WARWICK

THESAURUS-BASED APPROACHES: LIMITATIONS

WARWICK

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

DISTRIBUTIONAL MODELS OF MEANING

WARWICK

- Also called vector-space models of meaning.
- Offer much **higher recall than hand-built thesauri**.
  - Although they tend to have **lower precision**.

- **Intuition** of distributional models of meaning:
  - If **A and B have almost identical environments** we say that they are **synonyms**.
  - Remember lecture 6, word embeddings?

SYNONYMS IN SIMILAR ENVIRONMENTS

WARWICK

- **tackle** the task, **tackle** fake news.
  deal with the task, **deal with** fake news.

  tackle = deal with

- I like **chocolate**, **chocolate** is tasty, recipe for **chocolate**.
  I like **NLP**, **NLP** is hard, I’m in an **NLP** lecture.

  NLP ≠ chocolate

- How do we achieve this?

WORD-WORD CO-OCCURRENCE MATRIX

WARWICK

- Again, word-word co-occurrence matrix.

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

SAMPLE CONTEXTS: 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 study authorized in the first section of this

TERM-CONTEXT MATRIX FOR 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	

41

TERM-CONTEXT MATRIX FOR SIMILARITY

WARWICK

- NOTE: it's not the same to have:
  - high co-occurrence between w_1 and w_2.
  - w_1 and w_2 having similar vectors.

- "look" and "forward" will co-occur often.
  → "look forward" is common, but they're different in meaning.
- "automobile" and "car" will have similar vectors.
  → they'll co-occur with "drive", "crash", etc.
  → but not necessarily "automobile" and "car" together in many texts.

POINTWISE MUTUAL INFORMATION

- Instead of raw counts, **Pointwise Mutual Information (PMI) is often used**.
    - Do events ***x* and *y* co-occur more than if they were independent?**

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

POINTWISE MUTUAL INFORMATION

WARWICK

- PMI between two words: (Church & Hanks 1989)
  - Do **words *x* and *y* co-occur more than if they were independent?**

$$\text{PMI}(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}$$

- PPMI: Positive PMI between two words (Niwa & Nitta 1994)
  - Variant replacing all negative PMI values with zero

COMPUTING PMI

Count(w,context)

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

- P(w=information,c=data) = 6/19 = .32
- P(w=information) = 11/19 = .58
- P(c=data) = 7/19 = .37

p(w,context) / **p(w)**

	computer	data	pinch	result	sugar	p(w)
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	

COMPUTING PMI

	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	

$$pmi_{ij} = \log_2 \frac{p_{ij}}{p_{i*} p_{*j}}$$

PMI(information,data) = $\log_2(.32 / (.37 * .58)) = .57$

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

WARWICK

- PMI is **biased towards infrequent events**.

- Various weighting schemes help alleviate this (Turney and Pantel (2010)):
  - TF-IDF weighing scheme.
  - PPMI: replace negative scores with 0.

- Add-one smoothing can also help.

STATE-OF-THE-ART WORD SIMILARITY

- Currently, the state of the art approach for measuring word similarity is using **word embeddings**.
  - See Lecture 6!

- However for homonymy, hyponymy, etc. we still need more than embeddings.

RELATED TASK: WSD

- **Word Sense Disambiguation (WSD).**
  - Related task in which we aim to **identify which specific sense of a word is being used** in a particular instance in text.
    - I put my money in the **bank** → **bank**$_1$
    - We took the picture in the east **bank** of the Nile river → **bank**$_2$
  - As with computation of similarity, context can help WSD.

  https://www.cs.york.ac.uk/semeval-2013/task12/

EVALUATION

TWO TYPES OF EVALUATION

WARWICK

- Evaluation can be:
    - Intrinsic (in-vitro) evaluation.
    - Extrinsic (in-vivo) evaluation.

INTRINSIC EVALUATION

- Need a **corpus with human-annotated similarity scores**.
- Correlation between algorithm and human word similarity ratings.
- The WordSimilarity-353 Test Collection: http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/

EXTRINSIC EVALUATION

- A number of **tasks to test with**:
  - Spelling error detection.
  - Word sense disambiguation.
  - Taking multiple-choice vocabulary tests (TOEFL/Cambridge).

RESOURCES

- WordNet web interface: http://wordnetweb.princeton.edu/perl/webwn

- MeSH (Medical Subject Headings) thesaurus: https://www.ncbi.nlm.nih.gov/mesh

ASSOCIATED READING

- Jurafsky, Daniel, and James H. Martin. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 3rd edition. **Chapter 6.**