Semantics

Natural Language Processing

Master in Business Analytics and Big Data

acastellanos@faculty.ie.edu

Word Senses

- One lemma “bank” can have many meanings:

- **Sense** (or **word sense**)
  - Every aspect of a word’s meaning.
- The lemma **bank** here has **two senses**

Homonymy

- **Homonyms**: same lemma, distinct meanings

  bank$_1$: financial institution, bank$_2$: sloping land

  bat$_1$: club for hitting a ball, bat$_2$: nocturnal flying mammal

  - **Homographs**

    bank/bank

    bat/bat

  - **Homophones:**

    Write/right

    Piece/peace

Polysemy

Bank1: The bank was constructed in 1875 out of local red brick.

Bank2: I withdrew the money from the bank

- Are those the same sense?
  - Sense 2: “A financial institution”
  - Sense 1: “The building belonging to a financial institution”
- A **polysemous** word has **related** meanings
  - Most non-rare words have multiple meanings

Metonymy or Systematic Polysemy

- Lots of types of polysemy are **systematic**
  - **IE University**

Organization

Building

- Other such kinds of systematic polysemy:

Author (**Jane Austen** wrote Emma) ↔ Works of Author (I love **Jane Austen**)

Tree (**Plums** have beautiful blossoms) ↔ Fruit (I ate a preserved **plum**)

The “zeugma” test

- **How do we know when a word has more than one sense?**

- Two senses of `serve`?
    - `Which flights` **`serve`** `breakfast?`
    - `Does Lufthansa` **`serve`** `Philadelphia?`

- Does Lufthansa serve breakfast and Philadelphia?
    - **two different senses of “serve”**

Synonyms

- Word that have the **same meaning** in some or all contexts.
  - `big / large`
  - `automobile / car`

- **Two lexemes are synonyms iff:**
  - Can be substituted for each other in all situations
  - Have the same **propositional meaning**

Synonyms

- **There is not perfect synonyms**
  - Many aspects of meaning are identical
  - Notions of politeness, slang, register, genre, etc.
- Example:
  - Water/H_2O
  - Big/large
  - Brave/courageous

Synonymy is a relation between senses rather than words

- Consider the words *big* and *large*
- Are they synonyms?
  - How **big** is that plane?
  - Would I be flying on a **large** or small plane?
- How about here:
  - Miss Nelson became a kind of **big** sister to Benjamin.
  - Miss Nelson became a kind of **large** sister to Benjamin.
- Why?
  - *big* has a sense that means being older, or grown up
  - *large* lacks this sense

Antonyms

- Senses that are **opposites with respect to one meaning**
- Otherwise, **they are very similar!**

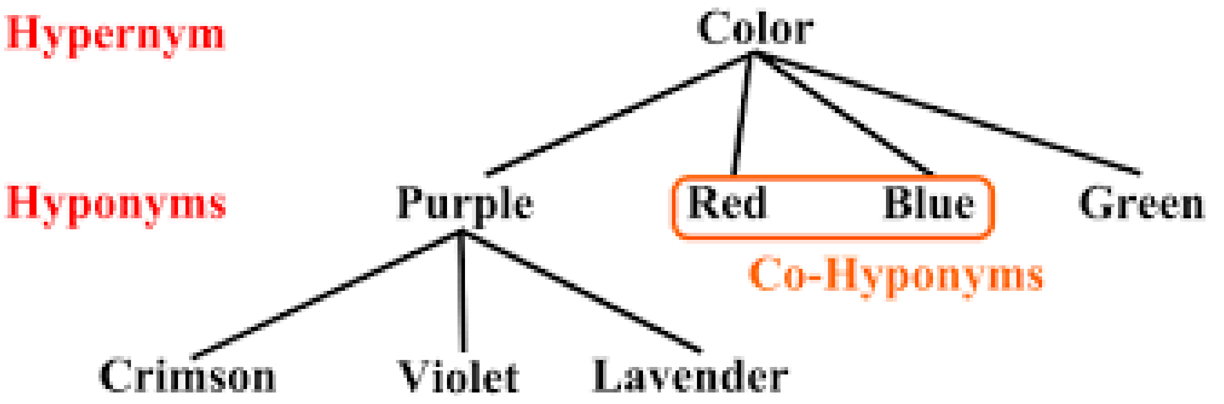
  dark/light short/long fast/slow rise/fall

  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

Hyponymy and Hypernymy

- One sense is a **hyponym** of another if the first sense is more specific, denoting a **subclass** of the other
- Conversely **hypernym/superordinate** ("hyper is super")

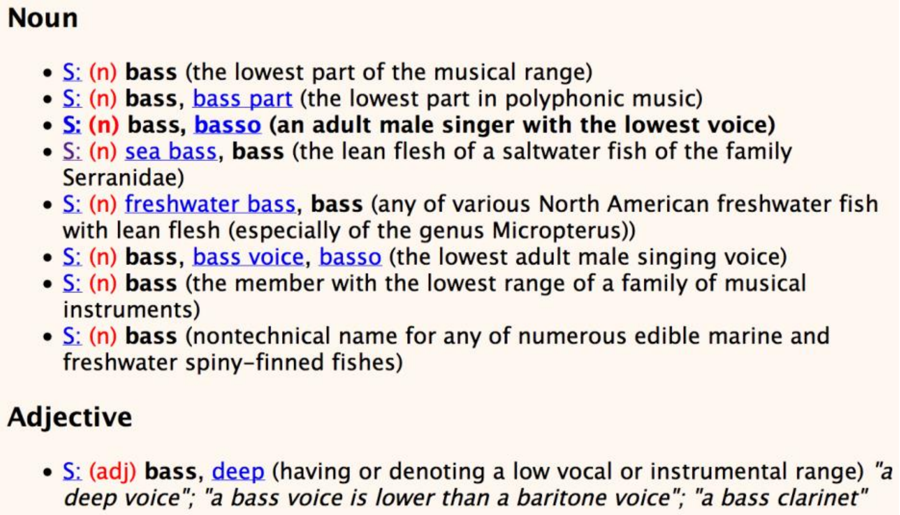


Hyponyms and Instances

- WordNet has both **classes** and **instances**.
- An **instance** is an individual, a proper noun that is a unique entity
  - `San Francisco` is an **instance** of `city`
- But `city` is a class
  - `city` is a **hyponym** of `municipality... location...`

WordNet 3.0

- A hierarchically organized lexical database
  - On-line thesaurus + aspects of a dictionary

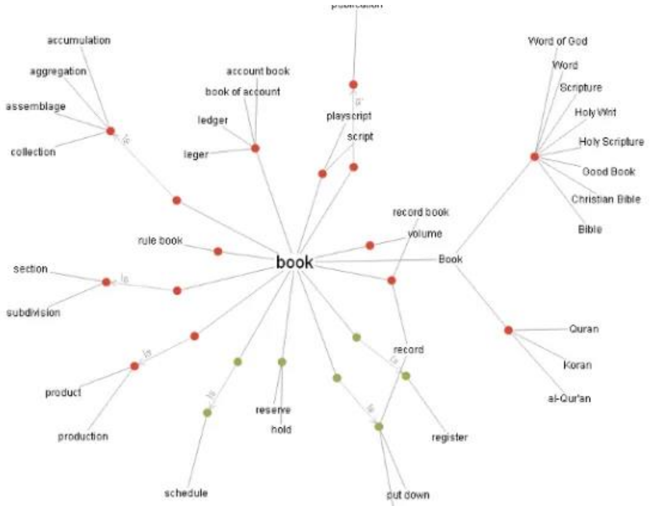


Noun

- 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"*



Senses in WordNet?

- **The synset (synonym set),** the set of near-synonyms, instantiates a sense or concept, with a gloss

- Example: chump as a noun with the gloss:

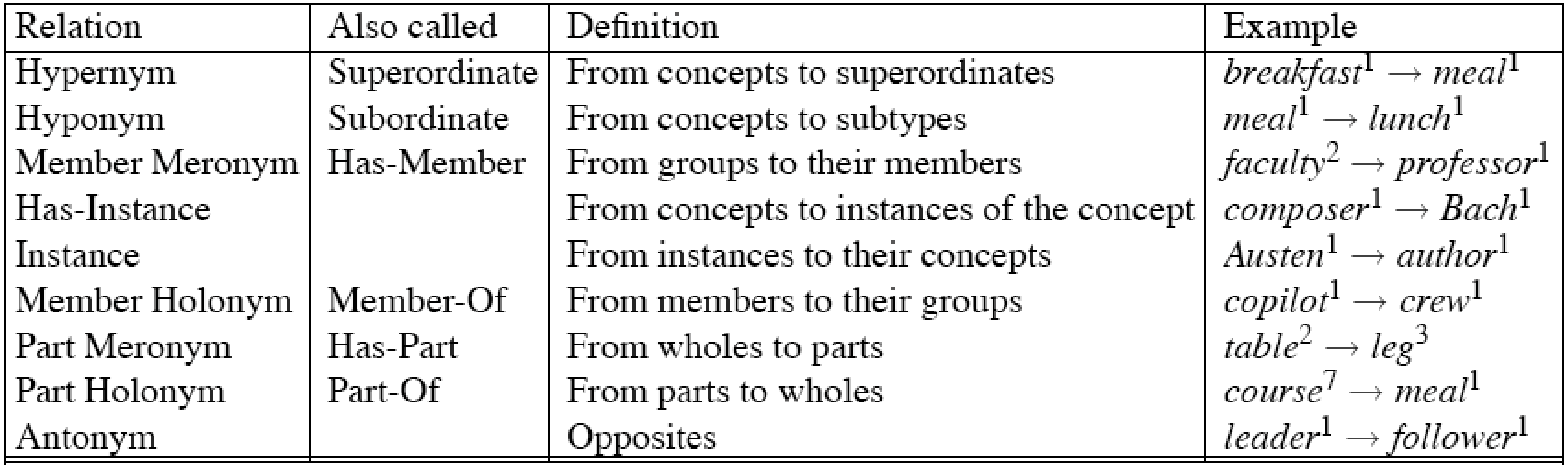
  "a person who is gullible and easy to take advantage of"

- 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}

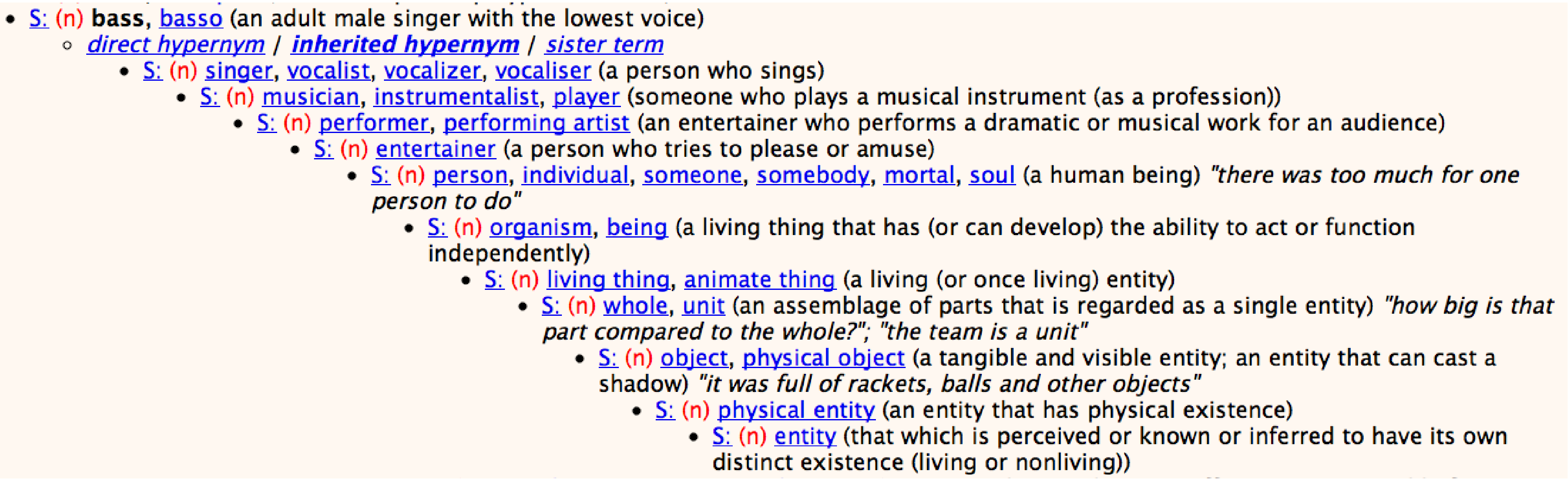
- Each of **these** senses have this same gloss

WordNet Relations



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

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 single entity) *"how big is that part compared to the whole?"; "the team is a unit"*
                    - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*
                      - S: (n) physical entity (an entity that has physical existence)
                        - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Word Similarity

- **Synonymy**: a binary relation
  - Two words are either synonymous or not
- **Similarity** (or **distance**): a looser metric
  - Two words are more similar if they share more features of meaning
- Similarity is properly a relation between **senses**
  - The word "`bank`" is not similar to the word "`slope`"
  - Bank[1] is similar to fund[3]
  - Bank[2] is similar to slope[5]
- But we'll compute similarity over both words and senses

Word similarity and word relatedness

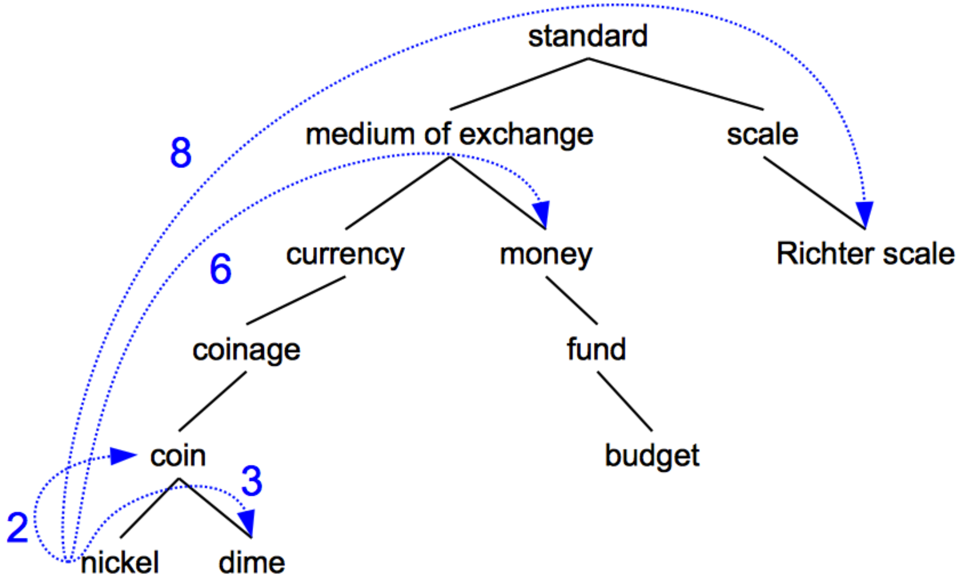
- We often distinguish **word similarity** from **word relatedness**
  - **Similar words**: near-synonyms
  - **Related words**: can be related any way
    - `car, bicycle:` **similar**
    - `car, gasoline:` **related**, not similar

Two classes 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 concepts (senses/synsets) are similar if they are **near each other in the thesaurus hierarchy**

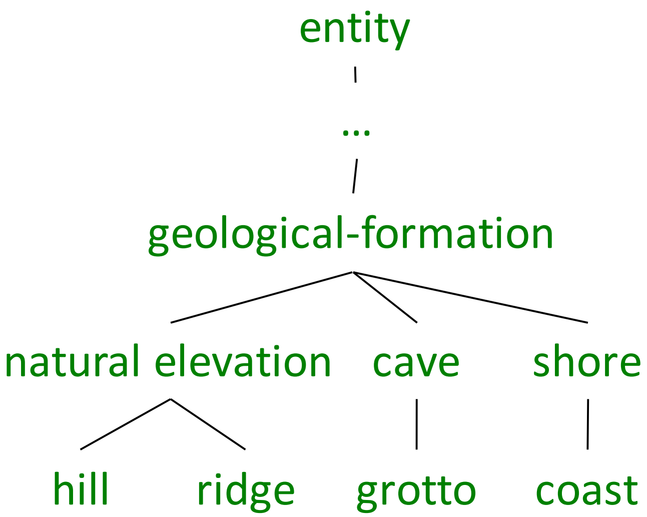


Problem with basic path-based similarity

- Assumes **each link represents a uniform distance**
  - But *nickel* to *money* seems to us to be closer than *nickel* to *standard*
  - **Nodes high in the hierarchy are very abstract**
- We instead want a **metric** that
  - Represents the cost of each edge independently
  - **Words connected only through abstract nodes are less similar**

Information Content Similarity

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



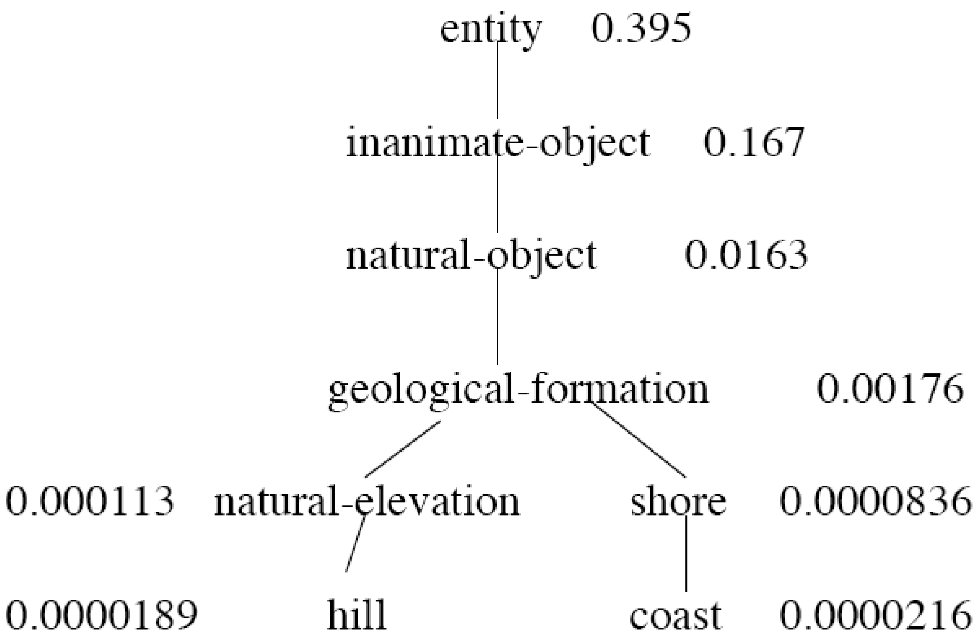
words(“geo-formation”) = {hill,ridge,grotto,coast,cave,shore,natural elevation}

words(“natural elevation”) = {hill, ridge}

Resnik 1995. Using information content to evaluate semantic similarity in a taxonomy. IJCAI

Information Content Similarity

- WordNet hierarchy augmented with probabilities P(c)



Information Content Similarity

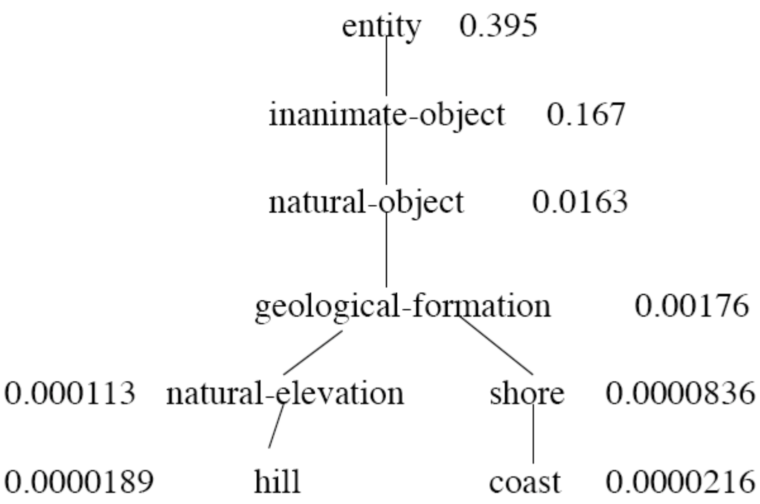
- Information content:

$$IC(c) = -\log P(c)$$

- Most informative subsume

$$LCS(c_1, c_2)$$

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



Using IC for similarity

- The similarity between two words is related to their common information
- Common information:
  - The information content of the most informative (lowest) subsumer (MIS/LCS) of the two nodes

$$sim_{resnik}(c_1, c_2) = -\log P(\ LCS(c_1, c_2)\ )$$

Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. IJCAI 1995
Philip Resnik. 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Applicatio
to Problems of Ambiguity in Natural Language. JAIR 11, 95-130

Dekang Lin similarity theorem

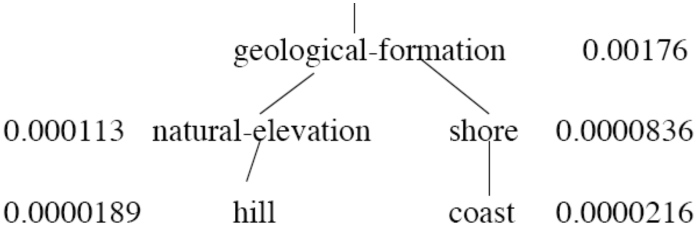
- Ratio between the amount of **information needed to state the commonality of A and B** and the **information needed to fully describe A and B**

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

- Lin defines IC as **2 x information of the LCS**

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

Dekang Lin Similarity



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

$$sim_{Lin}(\text{hill},\text{coast}) = \frac{2\log P(\text{geological-formation})}{\log P(\text{hill})+\log P(\text{coast})} = \frac{2\ln 0.00176}{\ln 0.0000189+\ln 0.0000216} = .59$$

Libraries

- NLTK

  http://nltk.github.com/api/nltk.corpus.reader.html?highlight=similarity - nltk.corpus.reader.WordNetCorpusReader.res_similarity

- WordNet::Similarity

  http://wn-similarity.sourceforge.net/

  - Web-based interface:

    http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi

Evaluating similarity

- Intrinsic Evaluation:
  - Correlation between algorithm and human word similarity ratings
- Extrinsic (task-based, end-to-end) Evaluation:
  - Malapropism (spelling error) detection
  - WSD
  - Taking TOEFL multiple-choice vocabulary tests

    Levied is closest in meaning to:

    imposed, believed, requested, correlated

Two classes 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?

Problems with thesaurus-based meaning

- **We don’t have a thesaurus for every language**
  - http://globalwordnet.org/resources/wordnets-in-the-world/
- **Low-resource settings**: problems with **recall**
  - Missing words
  - Missing connections between senses
  - 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):**
  - A and B have almost identical environments -> synonyms
- **Firth (1957):**
  - “You shall know a word by the 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**

Reminder: Term-document matrix

- Each cell: count of term *t* in a document *d*: $\mathrm{tf}_{t,d}$:
  - Each document is a **count vector**: 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 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 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?

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

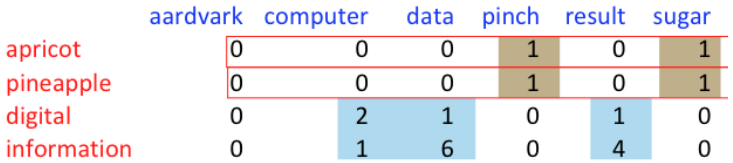
- **PMI between two words**: (Church & Hanks 1989)
  - Do words x and y co-occur more than if they were independent?

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

- **Positive PMI between two words** (Niwa & Nitta 1994)
  - Replace all PMI values less than 0 with zero

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

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

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

PPMI on a term-context matrix

$$p(w = \text{information}, c = \text{data}) \ = \ 6/19 \ = \ 0.32$$

$$p(w = \text{information}) \ = \ 11 \,/\, 19 \ = \ 0.58$$

$$p(c = \text{data}) \ = \ 7/19 \ = \ 0.37$$

$$p(w = \text{information}, c = \text{data}) \ = \ 6/19 \ = \ 0.32$$

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

$$p(w = \text{information}) \ = \ 11 \,/\, 19 \ = \ 0.58$$

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

$$p(c = \text{data}) \ = \ 7/19 \ = \ 0.37$$

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

PPMI on a term-context matrix

	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	

PPMI on a term-context matrix

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

	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(information,data) = $\log_2$ (0.32 / (0.37 * 0.58) = 0.57

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

- **PMI is biased toward infrequent events**
- Various weighting schemes help alleviate this
  - From Frequency to Meaning: Vector Space Models of Semantics (https://www.microsoft.com/en-us/research/wp-content/uploads/2017/07/jair10.pdf)
- Add-one smoothing can also help

Weighing PMI

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	

Weighing PMI

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

Using syntax to define a word’s context

- Zellig Harris (1968)

  “The meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities”

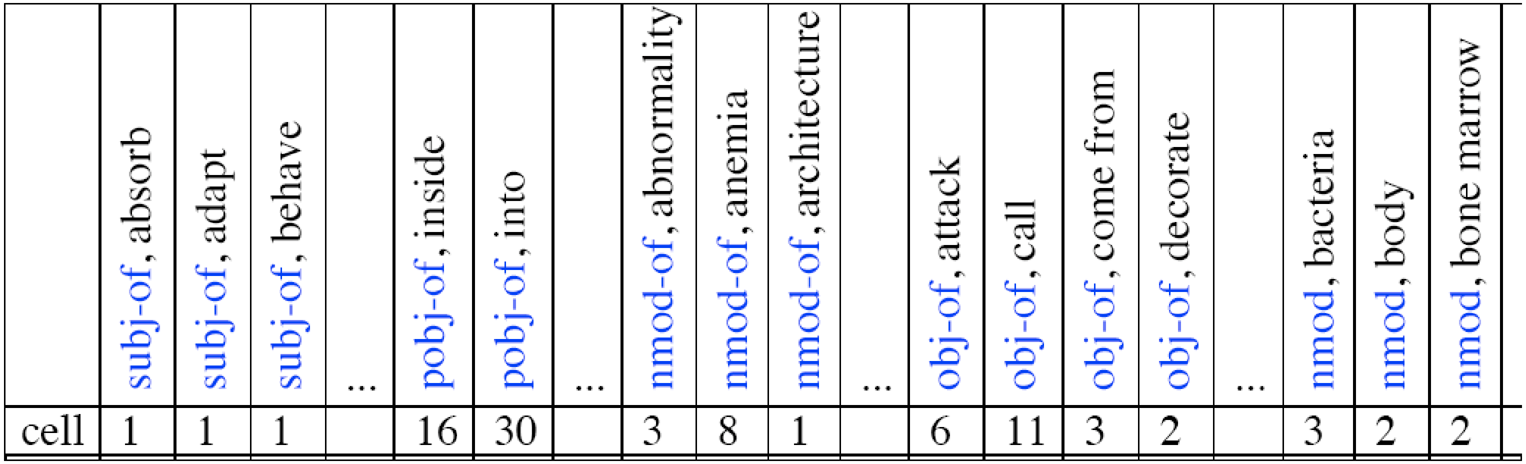
- **Two words are similar if they have similar parse contexts**

- **Duty** and **responsibility** (Chris Callison-Burch’s example)

Modified by adjectives	additional, administrative, assumed, collective, congressional, constitutional …
Objects of verbs	assert, assign, assume, attend to, avoid, become, breach …

Co-occurrence vectors based on syntactic dependencies

- The contexts C are different dependency relations
  - Subject-of- “absorb”
  - Prepositional-object of “inside”
- Counts for the word cell:



	subj-of, absorb	subj-of, adapt	subj-of, behave	...	pobj-of, inside	pobj-of, into	...	nmod-of, abnormality	nmod-of, anemia	nmod-of, architecture	...	obj-of, attack	obj-of, call	obj-of, come from	obj-of, decorate	...	nmod, bacteria	nmod, body	nmod, bone marrow
cell	1	1	1		16	30		3	8	1		6	11	3	2		3	2	2

Dekang Lin, 1998 “Automatic Retrieval and Clustering of Similar Words”

PMI applied to dependency relations

Object of “drink”	Count	PMI
tea	2	11.8
liquid	2	10.5
wine	2	9.3
anything	3	5.2
it	3	1.3

- “Drink it” more common than “drink wine”
- But “wine” is a better “drinkable” thing than “it”

Hindle, Don. 1990. Noun Classification from Predicate-Argument Structure. ACL

PMI applied to dependency relations

$$\cos(\vec{v},\vec{w}) = \frac{\vec{v}\bullet\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\vec{v}}{|\vec{v}|}\bullet\frac{\vec{w}}{|\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

	large	data	computer
apricot	1	0	0
digital	0	1	2
information	1	6	1

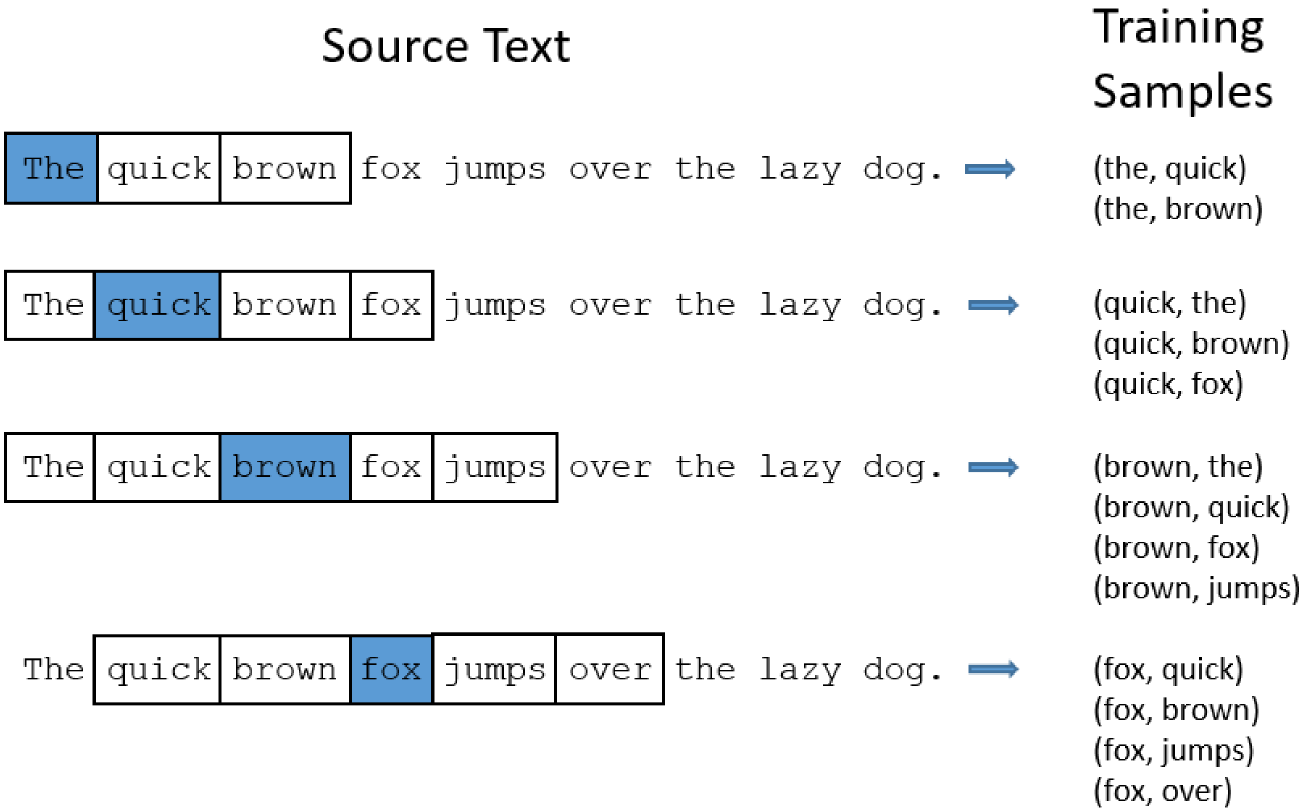
Which pair of words is more similar?

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

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

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

Word2vec



Source Text

Training Samples

The quick brown fox jumps over the lazy dog. → (the, quick) (the, brown)

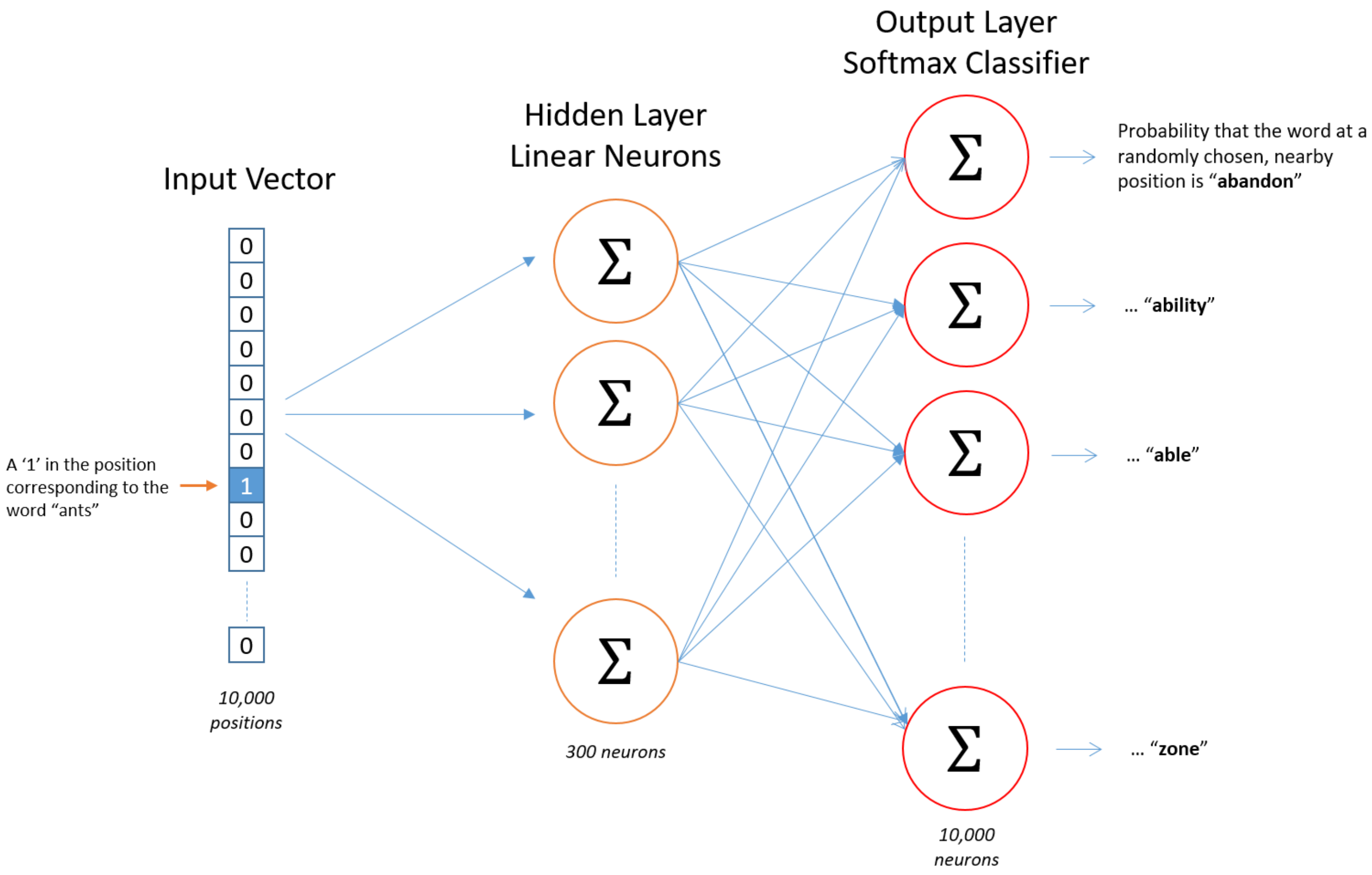
The quick brown fox jumps over the lazy dog. → (quick, the) (quick, brown) (quick, fox)

The quick brown fox jumps over the lazy dog. → (brown, the) (brown, quick) (brown, fox) (brown, jumps)

The quick brown fox jumps over the lazy dog. → (fox, quick) (fox, brown) (fox, jumps) (fox, over)

http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

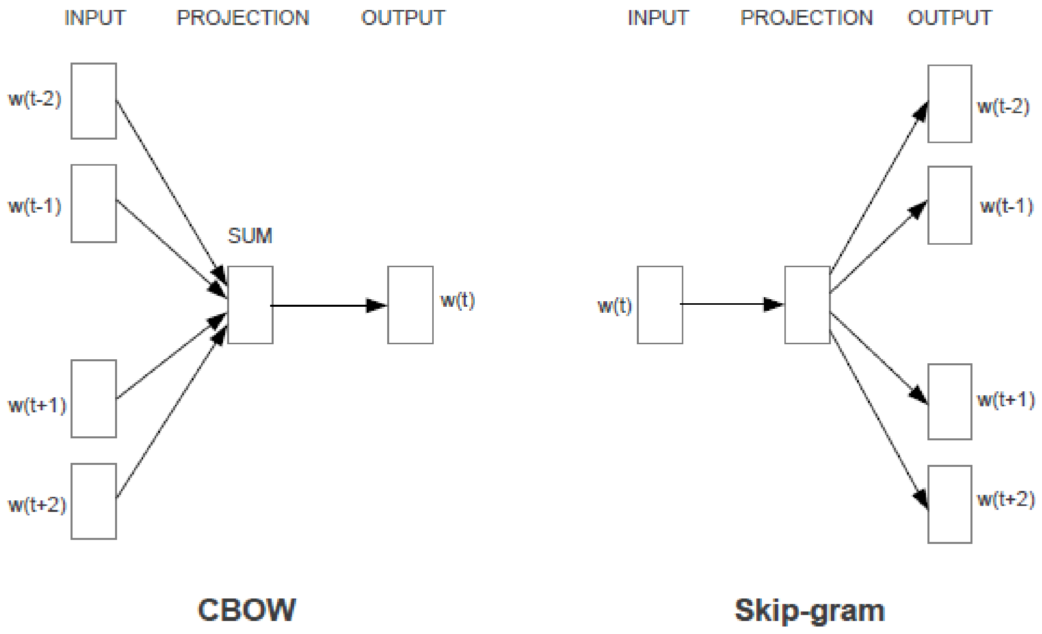
Word2vec



http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

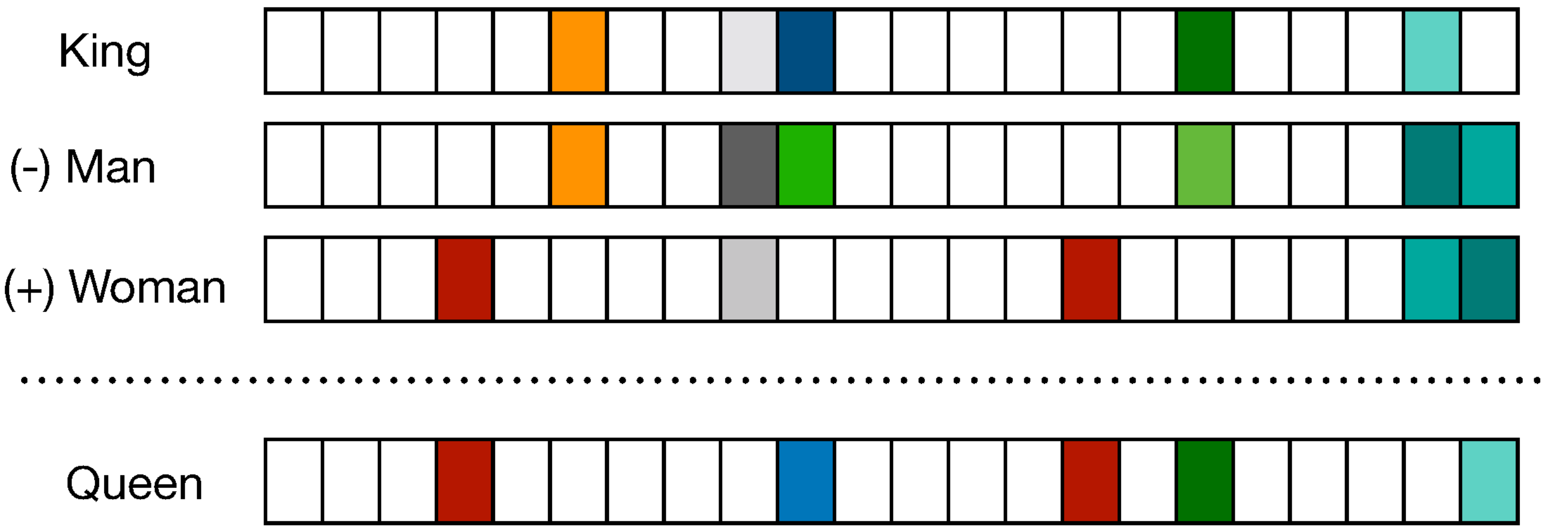
Word2vec

- Skip-grams vs CBOW:
  - Skip-grams: p(time -> It is ? to finish)
  - CBOW: P(it is ? to finish -> time)



Word2vec

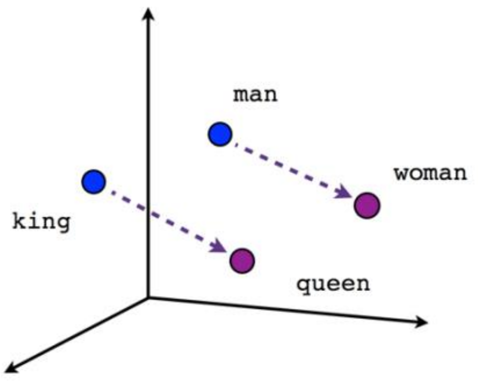
- The **result is a dense vector for each word**, good at predicting other words appearing in its **context** (also represented by vectors)



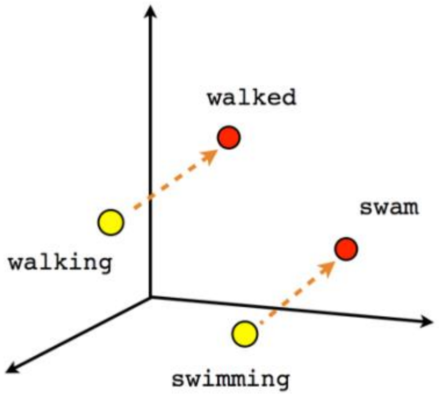
Source: http://veredshwartz.blogspot.sg.

Word2vec

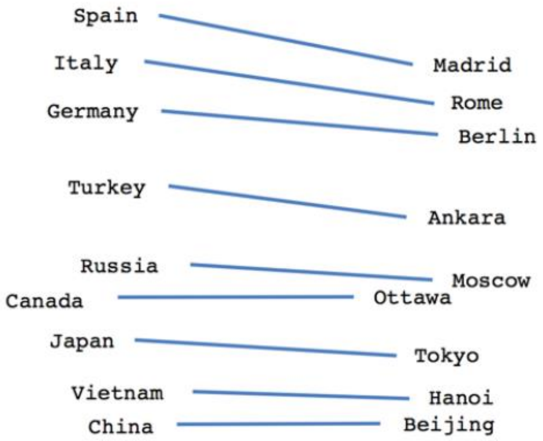
- The **result is a dense vector for each word**, good at predicting other words appearing in its **context** (also represented by vectors)



Male-Female



Verb tense



Country-Capital

https://www.tensorflow.org/tutorials/word2vec

Anything2Vec

- Med2vec: embeddings for medical codes
  - https://arxiv.org/abs/1602.05568
- Author2vec: embeddings of authors based on contents and authorships
  - https://researchweb.iiit.ac.in/~soumyajit.ganguly/papers/A2v_1.pdf
- Citation2vec: embedding of papers based on the citations
  - https://arxiv.org/pdf/1703.06587.pdf
- Doc2Vec: embeddings of whole documents
  - https://cs.stanford.edu/~quocle/paragraph_vector.pdf
- Many More:
  - http://nlp.town/blog/anything2vec/

Resources

- Christopher Olah's post on Word Embeddings
  - http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/
- Tensorflow tutorial on Word2Vec (with Code):
  - https://www.tensorflow.org/tutorials/word2vec
- GloVe: Global Vectors for Word Representation
  - https://nlp.stanford.edu/projects/glove/
- Word embeddings vs. other distributional semantic models
  - http://blog.aylien.com/overview-word-embeddings-history-word2vec-cbow-glove/