Word Meaning and Similarity

Word Senses and Word Relations

Slides are adapted from Dan Jurafsky

Reminder: lemma and wordform

- A **lemma** or **citation form**
  - Same stem, part of speech, rough semantics
- A **wordform**
  - The "inflected" word as it appears in text

Wordform	Lemma
banks	bank
sung	sing
duermes	dormir

Lemmas have senses

- One lemma “bank” can have many meanings:
  - Sense 1: …a **bank** can hold the investments in a custodial account^1…
  - Sense 2: “…as agriculture burgeons on the east **bank**$_2$ the river will shrink even more”
- **Sense** (or **word sense**)
  - A discrete representation

    of an aspect of a word’s meaning.
- The lemma **bank** here has two senses

Homonymy

Homonyms: words that share a form but have unrelated, distinct meanings:

- bank$_1$: financial institution, bank$_2$: sloping land
- bat$_1$: club for hitting a ball, bat$_2$: nocturnal flying mammal

1. Homographs (bank/bank, bat/bat)
2. Homophones:
    1. Write and right
    2. Piece and peace

Homonymy causes problems for NLP applications

- Information retrieval
  - “bat care”
- Machine Translation
  - bat: murciélago (animal) or bate (for baseball)
- Text-to-Speech
  - bass (stringed instrument) vs. bass (fish)

Polysemy

- 1. The **bank** was constructed in 1875 out of local red brick.
- 2. 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: A systematic relationship between senses

- Lots of types of polysemy are systematic
  - `School, university, hospital`
  - All can mean the institution or the building.
- A systematic relationship:
  - Building ⟷ Organization
- 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`)

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

- The "zeugma" test: Two senses of `serve`?
  - `Which flights` **`serve`** `breakfast?`
  - `Does Lufthansa` **`serve`** `Philadelphia?`
  - ?Does Lufthansa serve breakfast and San Jose?
- Since this conjunction sounds weird,
  - we say that these are **two different senses of "serve"**

Synonyms

- Word that have the same meaning in some or all contexts.
  - filbert / hazelnut
  - couch / sofa
  - big / large
  - automobile / car
  - vomit / throw up
  - Water / H_20
- Two lexemes are synonyms
  - if they can be substituted for each other in all situations
  - If so they have the same **propositional meaning**

Synonyms

- But there are few (or 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.
- Example:
  - Water/H_20
  - 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 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

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*
  - *mango* is a hyponym of *fruit*
- Conversely **hypernym/superordinate** ("hyper is super")
  - *vehicle* is a **hypernym** of *car*
  - *fruit* is a hypernym of *mango*

Superordinate/hyper	vehicle	fruit	furniture
Subordinate/hyponym	car	mango	chair

Hyponymy more formally

- Extensional:
  - The class denoted by the superordinate extensionally includes the class denoted by the hyponym
- Entailment:
  - A sense A is a hyponym of sense B if *being an A* entails *being a B*
- Hyponymy is usually transitive
  - (A hypo B and B hypo C entails A hypo C)
- Another name: the **IS-A hierarchy**
  - A IS-A B (or A ISA B)
  - B **subsumes** A

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...`

Word Meaning and Similarity

Word Senses and Word Relations

Word Meaning and Similarity

WordNet and other Online Thesauri

Applications of Thesauri and Ontologies

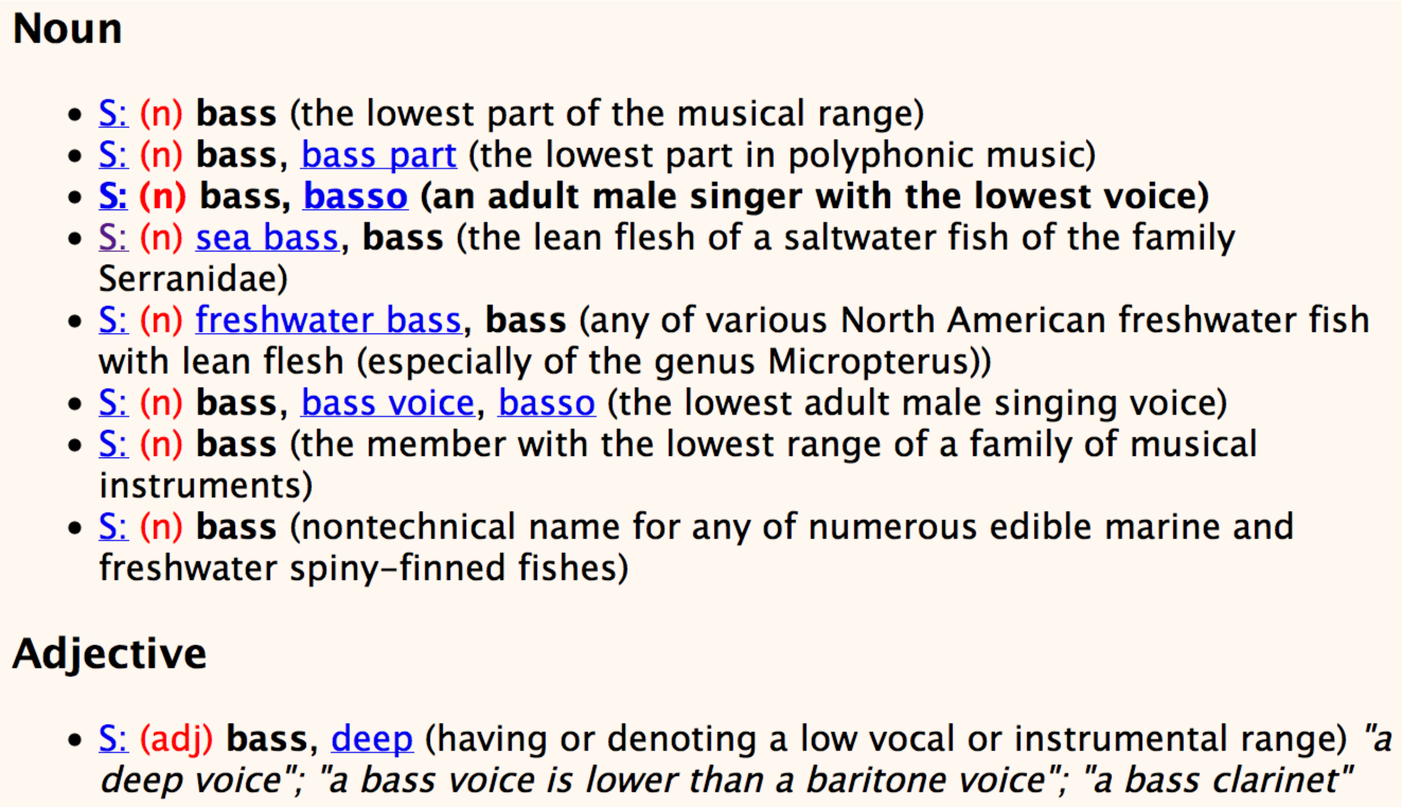
- Information Extraction
- Information Retrieval
- Question Answering
- Bioinformatics and Medical Informatics
- Machine Translation

WordNet 3.0

- A hierarchically organized lexical database
- On-line thesaurus + aspects of a dictionary
  - Some other languages available or under development
    - (Arabic, Finnish, German, Portuguese…)

Category	Unique Strings
Noun	117,798
Verb	11,529
Adjective	22,479
Adverb	4,481

Senses of "bass" in Wordnet



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"*

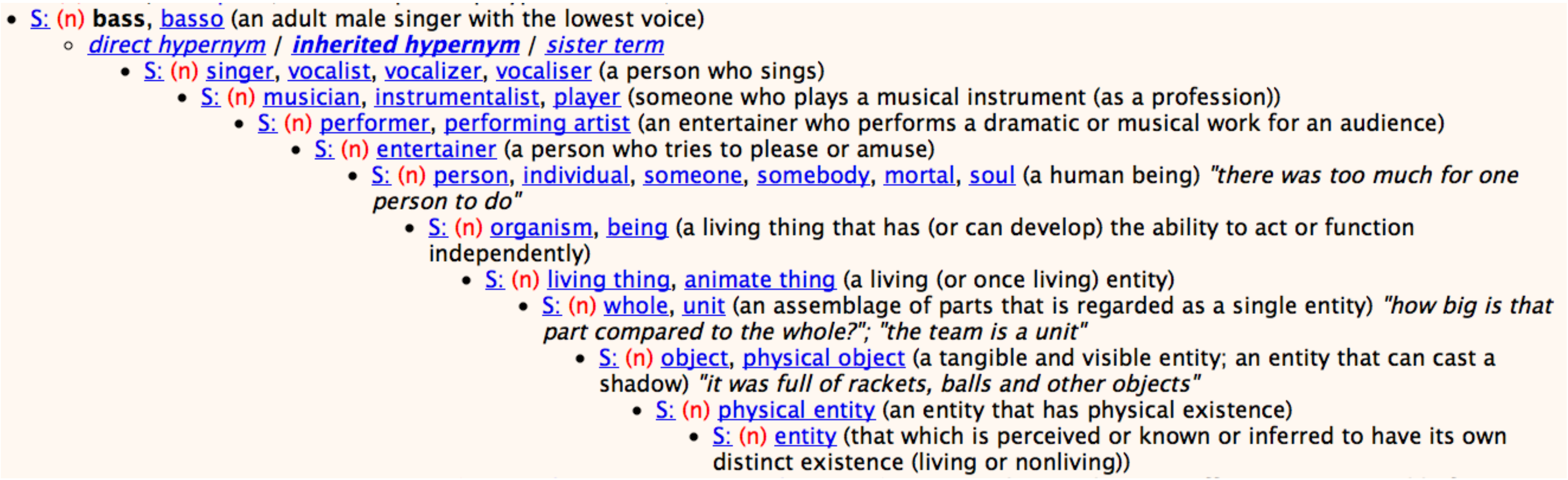
How is “sense” defined in WordNet?

- **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:

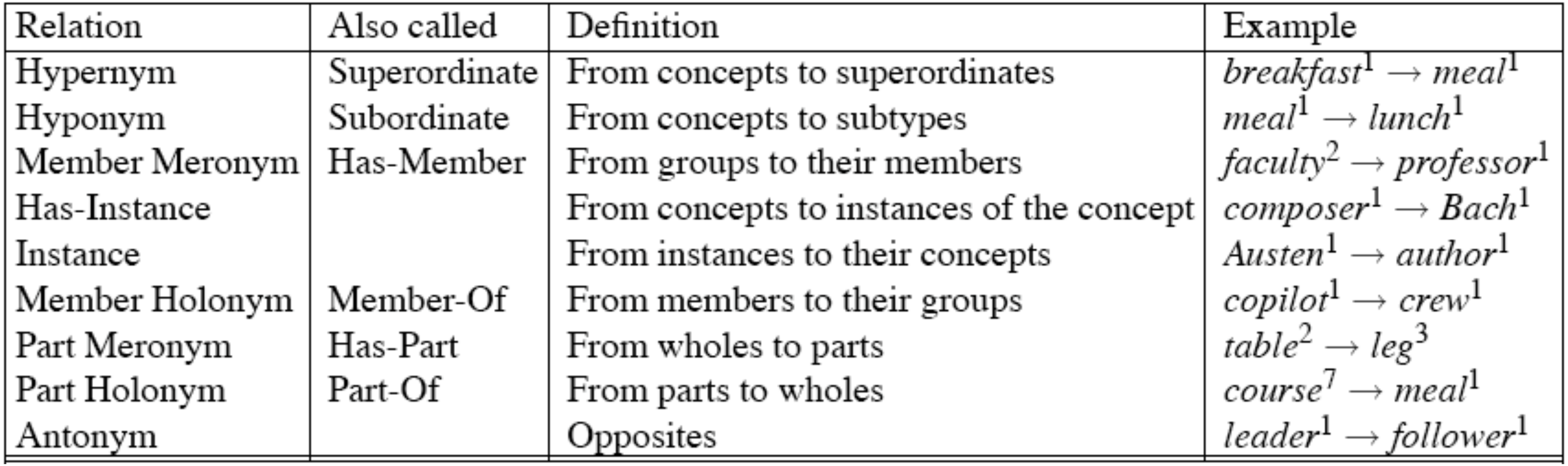
  `chump`[1], `fool`[2], `gull`[1], `mark`[9], `patsy`[1], `fall guy`[1], `sucker`[1], `soft touch`[1], `mug`[2]
- Each of **these** senses have this same gloss
  - (Not **every** sense; sense 2 of gull is the aquatic bird)

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 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))

WordNet Noun Relations



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

WordNet 3.0

- Where it is:
  - http://wordnetweb.princeton.edu/perl/webwn
- Libraries
  - Python: WordNet from NLTK
    - http://www.nltk.org/Home
  - Java:
    - JWNL, extJWNL on sourceforge

MeSH: Medical Subject Headings thesaurus from the National Library of Medicine

- **MeSH (Medical Subject Headings)**
  - 177,000 entry terms that correspond to 26,142 biomedical "headings"

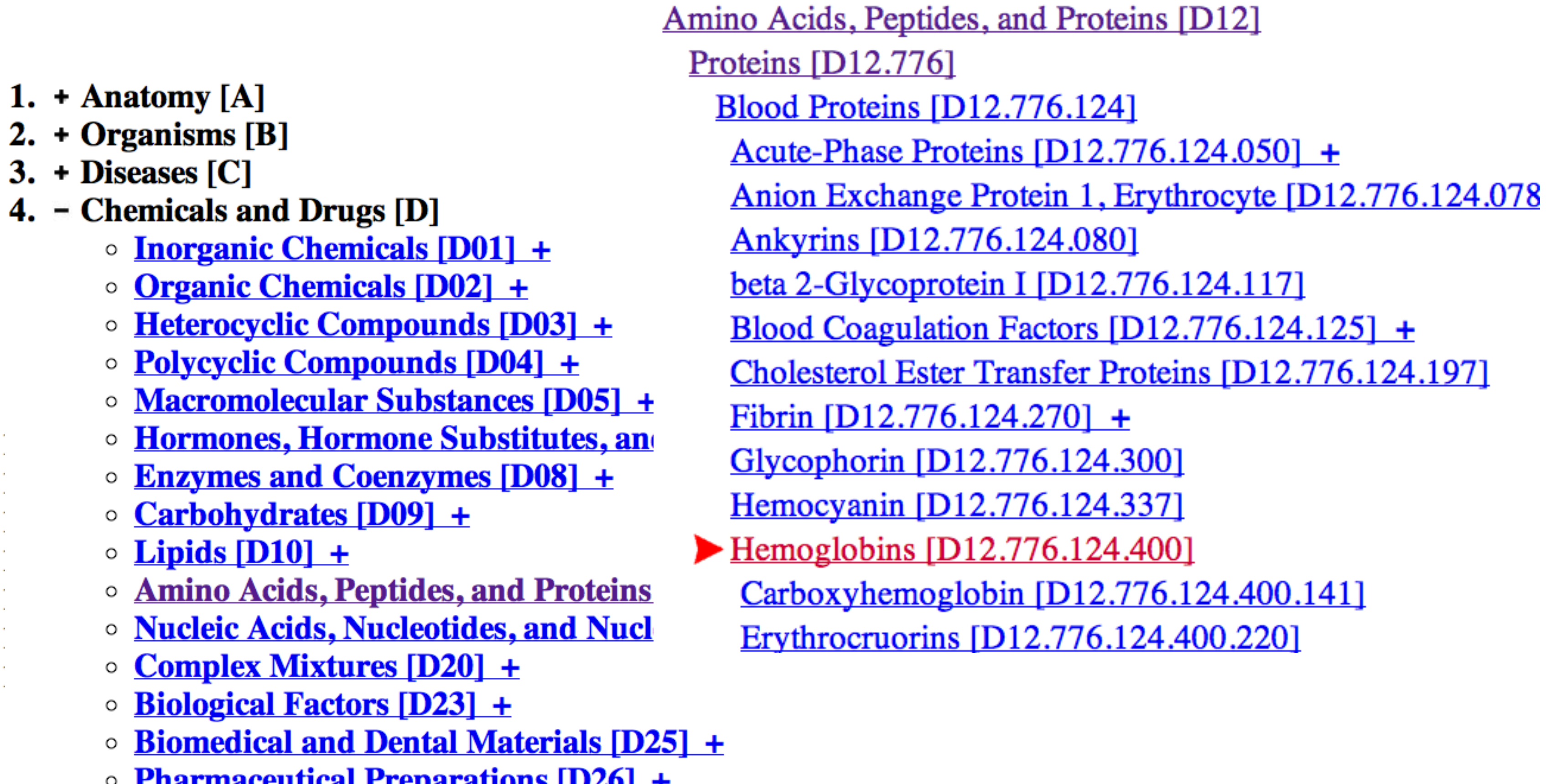
- **Hemoglobins**

  Entry Terms: Eryhem, Ferrous Hemoglobin, Hemoglobin

  Synset

  Definition: The oxygen-carrying proteins of ERYTHROCYTES. They are found in all vertebrates and some invertebrates. The number of globin subunits in the hemoglobin quaternary structure differs between species. Structures range from monomeric to a variety of multimeric arrangements

The MeSH Hierarchy



1. + Anatomy [A]
2. + Organisms [B]
3. + Diseases [C]
4. – Chemicals and Drugs [D]
   - Inorganic Chemicals [D01] +
   - Organic Chemicals [D02] +
   - Heterocyclic Compounds [D03] +
   - Polycyclic Compounds [D04] +
   - Macromolecular Substances [D05] +
   - Hormones, Hormone Substitutes, an
   - Enzymes and Coenzymes [D08] +
   - Carbohydrates [D09] +
   - Lipids [D10] +
   - Amino Acids, Peptides, and Proteins
   - Nucleic Acids, Nucleotides, and Nucl
   - Complex Mixtures [D20] +
   - Biological Factors [D23] +
   - Biomedical and Dental Materials [D25] +
   - Pharmaceutical Preparations [D26] +

Amino Acids, Peptides, and Proteins [D12]
Proteins [D12.776]
Blood Proteins [D12.776.124]
Acute-Phase Proteins [D12.776.124.050] +
Anion Exchange Protein 1, Erythrocyte [D12.776.124.078
Ankyrins [D12.776.124.080]
beta 2-Glycoprotein I [D12.776.124.117]
Blood Coagulation Factors [D12.776.124.125] +
Cholesterol Ester Transfer Proteins [D12.776.124.197]
Fibrin [D12.776.124.270] +
Glycophorin [D12.776.124.300]
Hemocyanin [D12.776.124.337]
➤ Hemoglobins [D12.776.124.400]
Carboxyhemoglobin [D12.776.124.400.141]
Erythrocruorins [D12.776.124.400.220]

Uses of the MeSH Ontology

- Provide synonyms (“entry terms”)
  - E.g., glucose and dextrose
- Provide hypernyms (from the hierarchy)
  - E.g., glucose ISA monosaccharide
- Indexing in MEDLINE/PubMED database
  - NLM’s bibliographic database:
    - 20 million journal articles
    - Each article hand-assigned 10-20 MeSH terms

Word Meaning and Similarity

WordNet and other Online Thesauri

Word Meaning and Similarity

Word Similarity: Thesaurus Methods

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

Why word similarity

- Information retrieval
- Question answering
- Machine translation
- Natural language generation
- Language modeling
- Automatic essay grading
- Plagiarism detection
- Document clustering

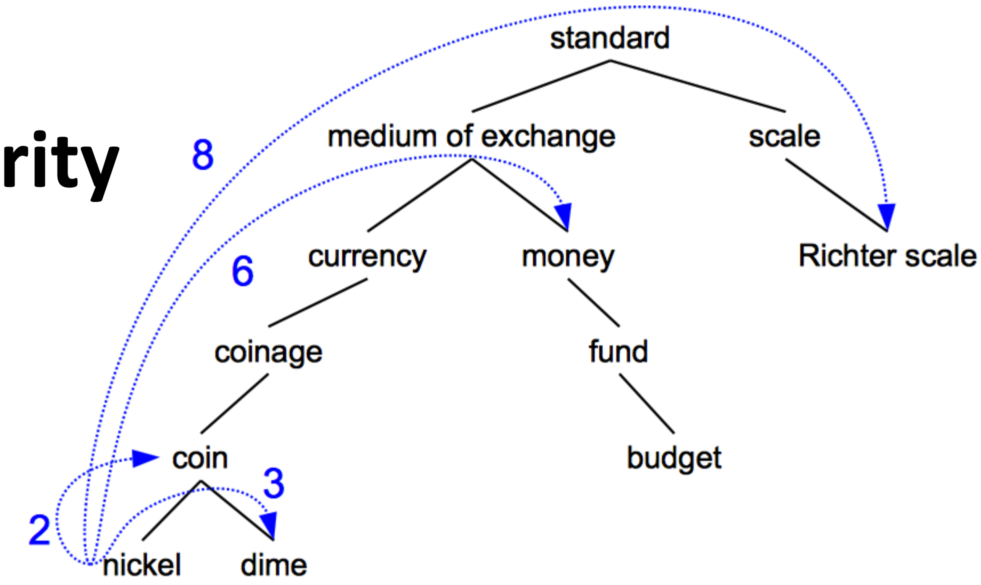
Word similarity and word relatedness

- 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



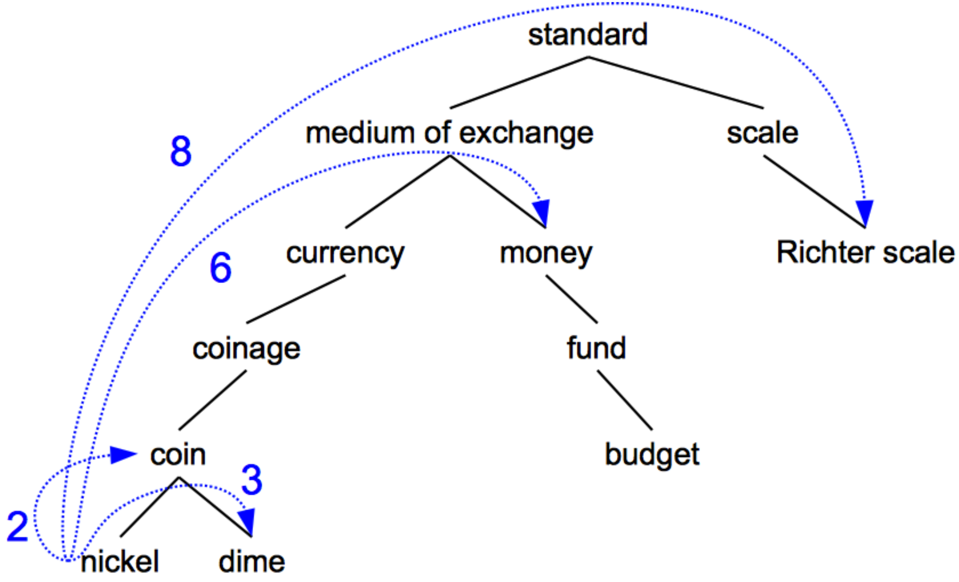
- Two concepts (senses/synsets) are similar if they are near each other in the thesaurus hierarchy
  - =have a short path between them
  - concepts have path 1 to themselves

Refinements to path-based similarity

- pathlen(c_1,c_2) = 1 + number of edges in the shortest path in the hypernym graph between sense nodes c_1 and c_2
- ranges from 0 to 1 (identity)
- $\text{simpath}(c_1,c_2) = \dfrac{1}{\text{pathlen}(c_1,c_2)}$
- $\text{wordsim}(w_1,w_2) = \max\limits_{c_1 \in \text{senses}(w_1), c_2 \in \text{senses}(w_2)} \text{sim}(c_1,c_2)$

Example: path-based similarity

$\text{simpath}(c_1,c_2) = 1/\text{pathlen}(c_1,c_2)$



simpath(*nickel,coin*) = $1/2 = .5$

simpath(*fund,budget*) = $1/2 = .5$

simpath(*nickel,currency*) = $1/4 = .25$

simpath(*nickel,money*) = $1/6 = .17$

simpath(*coinage,Richter scale*) = $1/6 = .17$

Problem with basic path-based similarity

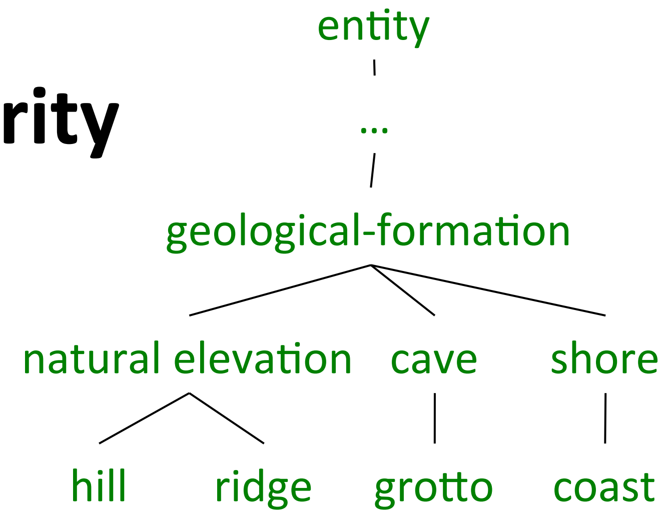
- Assumes each link represents a uniform distance
  - But *nickel* to *money* seems to us to be closer than *nickel* to *standard*
  - Nodes high in the hierarchy are very abstract
- We instead want a metric that
  - Represents the cost of each edge independently
  - Words connected only through abstract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

Information content similarity metrics

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

- Let's define $P(c)$ as:
  - The probability that a randomly selected word in a corpus is an instance of concept c
  - Formally: there is a distinct random variable, ranging over words, associated with each concept in the hierarchy
    - for a given concept, each observed noun is either
      - a member of that concept with probability $P(c)$
      - not a member of that concept with probability $1\text{-}P(c)$
  - All words are members of the root node (Entity)
    - $P(\text{root})=1$
  - The lower a node in hierarchy, the lower its probability

Information content similarity


entity
…
geological-formation
natural elevation
cave
shore
hill
ridge
grotto
coast

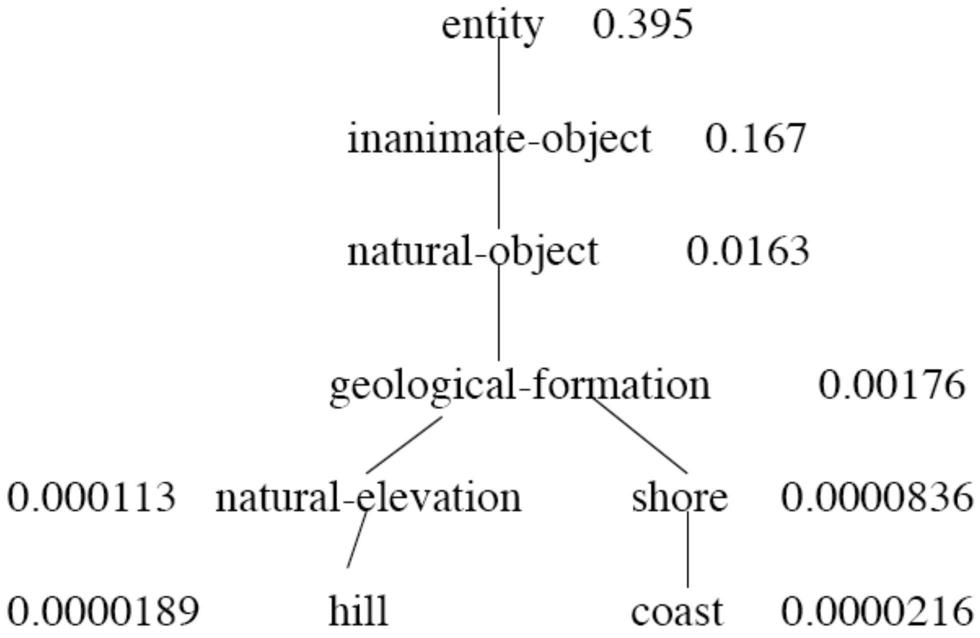
- Train by counting in a corpus
  - Each instance of `hill` counts toward frequency of *natural elevation*, *geological formation*, *entity*, etc
  - Let words(c) be the set of all words that are children of node c
    - words(“geo-formation”) = {hill,ridge,grotto,coast,cave,shore,natural elevation}
    - words(“natural elevation”) = {hill, ridge}

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

Information content similarity

- WordNet hierarchy augmented with probabilities P(c)

D. Lin. 1998. An Information-Theoretic Definition of Similarity. ICML 1998



Information content: definitions

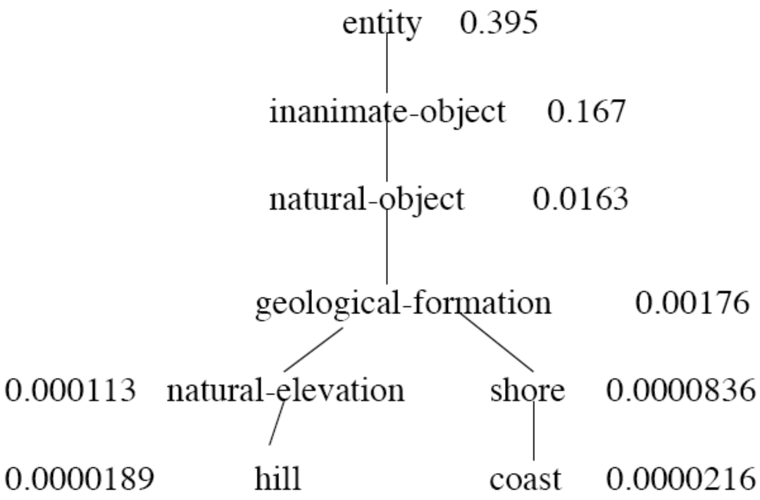
- Information content:

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

- Most informative subsumer (Lowest common subsumer)

  $LCS(c_1,c_2) =$

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



Using information content for similarity: the Resnik method

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 Application to Problems of Ambiguity in Natural Language. JAIR 11, 95-130.

- The similarity between two words is related to their common information
- The more two words have in common, the more similar they are
- Resnik: measure common information as:
  - The information content of the most informative (lowest) subsumer (MIS/LCS) of the two nodes
  - $\mathrm{sim}_{\mathrm{resnik}}(c_1,c_2) = -\log P(\ \mathrm{LCS}(c_1,c_2)\ )$

Dekang Lin method

Dekang Lin. 1998. An Information-Theoretic Definition of Similarity. ICML

- Intuition: Similarity between A and B is not just what they have in common
- The more **differences** between A and B, the less similar they are:
  - Commonality: the more A and B have in common, the more similar they are
  - Difference: the more differences between A and B, the less similar
- Commonality: IC(common(A,B))
- Difference: IC(description(A,B)-IC(common(A,B))

Dekang Lin similarity theorem

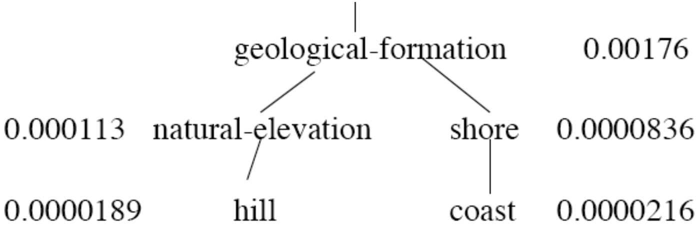
- The similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are

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

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

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

Lin similarity function



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

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

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

$$= .59$$

The (extended) Lesk Algorithm

- A thesaurus-based measure that looks at **glosses**
- Two concepts 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 *n*-word phrase that's in both glosses
  - Add a score of n^2
  - Paper and specially prepared for $1 + 2^2 = 5$
  - Compute overlap also for other relations
    - glosses of hypernyms and hyponyms

Summary: thesaurus-based similarity

$$\text{sim}_{\text{path}}(c_1,c_2) = \frac{1}{pathlen(c_1,c_2)}$$

$$\text{sim}_{\text{resnik}}(c_1,c_2) = -\log P(LCS(c_1,c_2)) \quad \text{sim}_{\text{lin}}(c_1,c_2) = \frac{2\log P(LCS(c_1,c_2))}{\log P(c_1) + \log P(c_2)}$$

$$\text{sim}_{\text{jiangconrath}}(c_1,c_2) = \frac{1}{\log P(c_1) + \log P(c_2) - 2\log P(LCS(c_1,c_2))}$$

$$\text{sim}_{eLesk}(c_1,c_2) = \sum_{r,q \in RELS} \text{overlap}(\text{gloss}(r(c_1)),\text{gloss}(q(c_2)))$$

Libraries for computing thesaurus-based similarity

- 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

- Extrinsic (task-based, end-to-end) Evaluation:
  - Question Answering
  - Spell Checking
  - Essay grading
- Intrinsic Evaluation:
  - Correlation between algorithm and human word similarity ratings
    - Wordsim353: 353 noun pairs rated 0-10. *sim(plane,car)=5.77*
  - Taking TOEFL multiple-choice vocabulary tests
    - `Levied is closest in meaning to:`

      `imposed, believed, requested, correlated`

Word Meaning and Similarity

Word Similarity: Thesaurus Methods

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 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*: $\mathrm{tf}_{t,d}$:
  - 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 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?

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

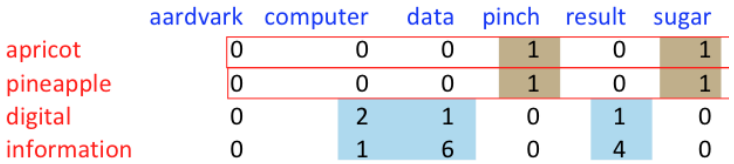
- **PMI between two words**: (Church & Hanks 1989)
  - Do words x and y co-occur more than if they were independent?

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

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

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

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

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

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

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

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

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(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

- 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]

	computer	data	pinch	result	sugar	p(w)
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)

Word Meaning and Similarity

Word Similarity: Distributional Similarity (II)

Using syntax to define a word’s context

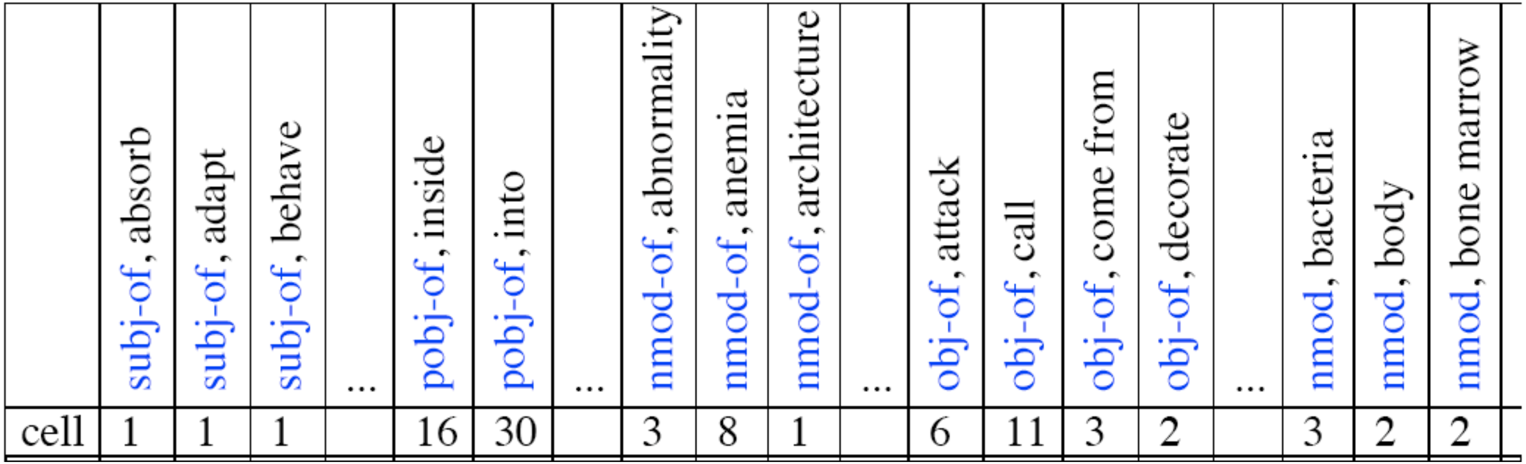
- Zellig Harris (1968)
  - “The meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities”
- Two words are similar if they have similar parse contexts
- **Duty** and **responsibility** (Chris Callison-Burch’s 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

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

- 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

PMI applied to dependency relations

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

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”

Reminder: cosine for computing similarity

Dot product   Unit vectors

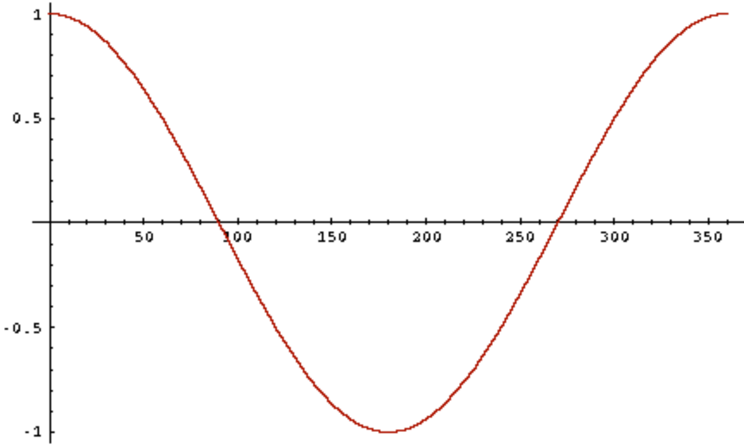
$$\cos(\vec{v},\vec{w}) = \frac{\vec{v}\bullet\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\vec{v}}{|\vec{v}|}\bullet\frac{\vec{w}}{|\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

v_i is the PPMI value for word *v* in context *i*
w_i is the PPMI value for word *w* in context *i*.

Cos($\vec{v}$,$\vec{w}$) is the cosine similarity of $\vec{v}$ and $\vec{w}$

Cosine as a similarity metric

- -1: vectors point in opposite directions
- +1: vectors point in same directions
- 0: vectors are orthogonal



- Raw frequency or PPMI are non-negative, so cosine range 0-1

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

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

Which pair of words is more similar?

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

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

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

Other possible similarity measures

$$\mathrm{sim}_{\mathrm{cosine}}(\vec{v},\vec{w}) = \frac{\vec{v}\cdot\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

$$\mathrm{sim}_{\mathrm{Jaccard}}(\vec{v},\vec{w}) = \frac{\sum_{i=1}^{N} \min(v_i,w_i)}{\sum_{i=1}^{N} \max(v_i,w_i)}$$

$$\mathrm{sim}_{\mathrm{Dice}}(\vec{v},\vec{w}) = \frac{2\times\sum_{i=1}^{N} \min(v_i,w_i)}{\sum_{i=1}^{N} (v_i+w_i)}$$

$$\mathrm{sim}_{\mathrm{JS}}(\vec{v}||\vec{w}) = D(\vec{v}|\frac{\vec{v}+\vec{w}}{2}) + D(\vec{w}|\frac{\vec{v}+\vec{w}}{2})$$

D: KL Divergence

Evaluating similarity
(the same as for thesaurus-based)

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

`Levied is closest in meaning to which of these:`

`imposed, believed, requested, correlated`

Word Meaning and Similarity

Word Similarity: Distributional Similarity (II)

Vector Semantics

Dense Vectors

Sparse versus dense vectors

- PPMI vectors are
  - **long** (length |V|= 20,000 to 50,000)
  - **sparse** (most elements are zero)
- Alternative: learn vectors which are
  - **short** (length 200-1000)
  - **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 represented as distinct dimensions; this fails to capture similarity between a word with *car* as a neighbor and a word with *automobile* as a neighbor

Three methods for getting short dense vectors

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

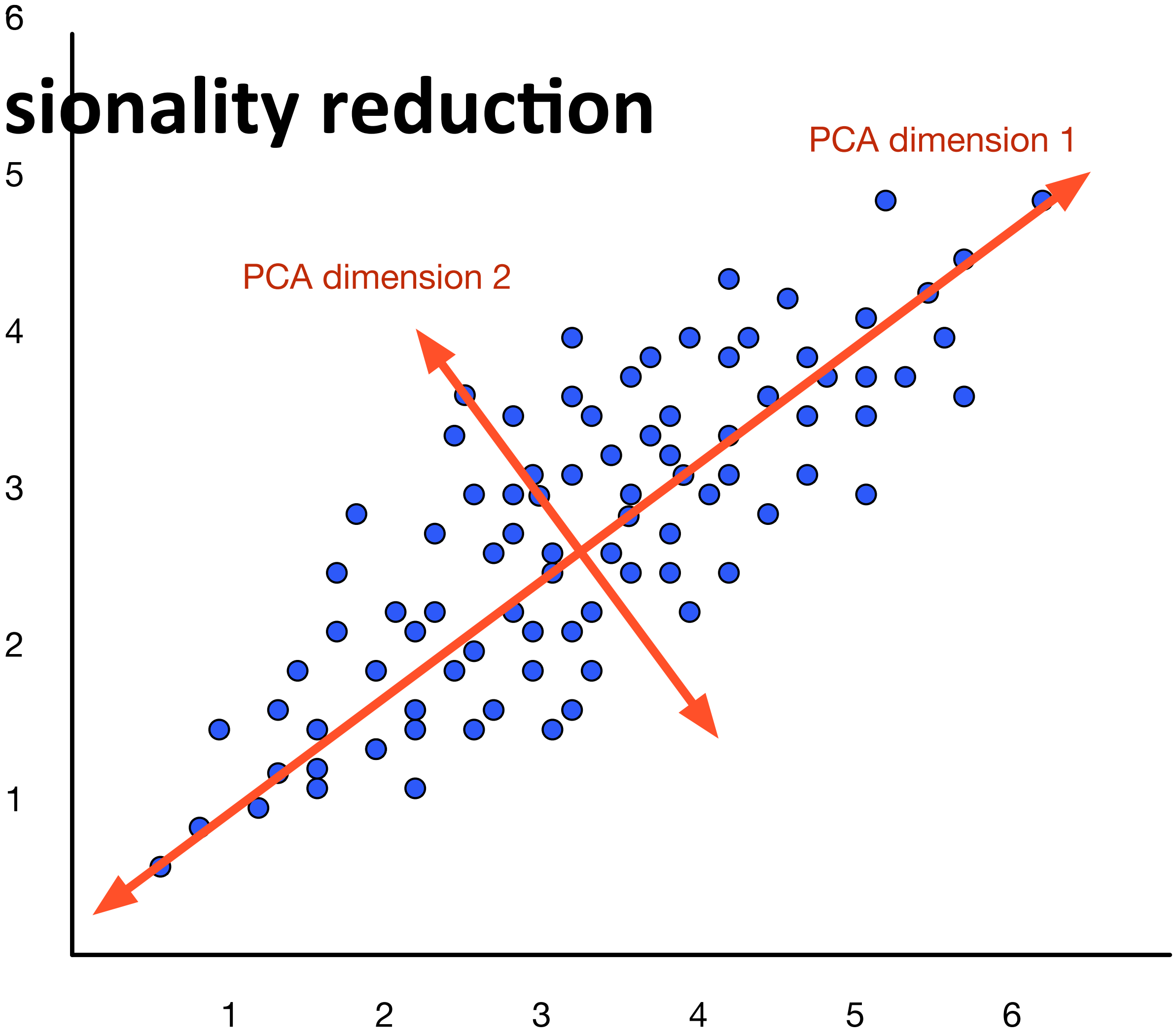
Vector Semantics

Dense Vectors via SVD

Intuition

- Approximate an N-dimensional dataset using fewer dimensions
- By first rotating the axes into a new space
- In which the highest order dimension captures the most variance in the original dataset
- And the next dimension captures the next most variance, etc.
- Many such (related) methods:
  - PCA – principle components analysis
  - Factor Analysis
  - SVD

Dimensionality reduction



Singular Value Decomposition

Any rectangular matrix X equals the product of 3 matrices:

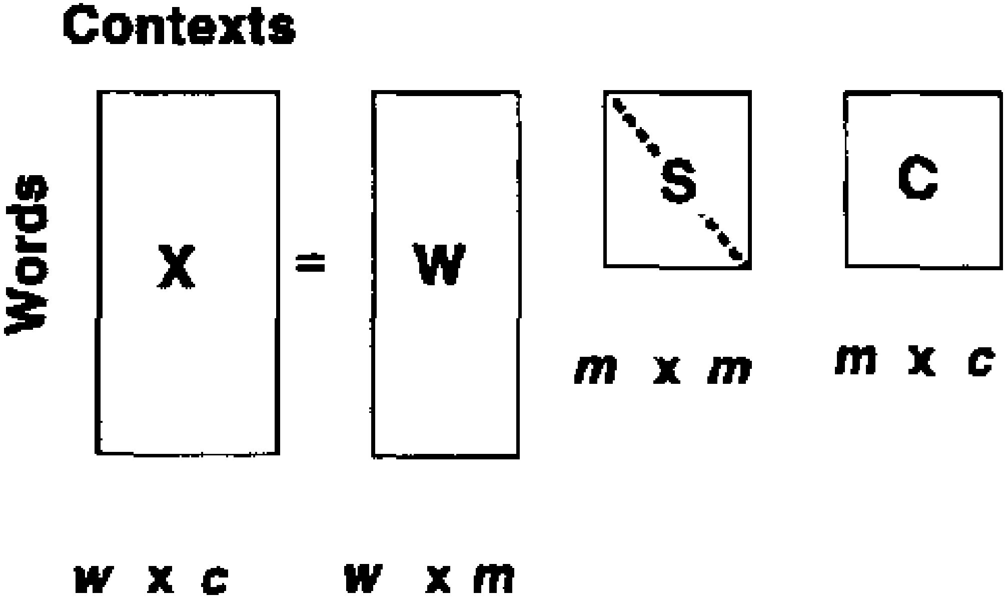
W: rows corresponding to original but m columns represents a dimension in a new latent space, such that

- M column vectors are orthogonal to each other
- Columns are ordered by the amount of variance in the dataset each new dimension accounts for

S: diagonal *m* x *m* matrix of **singular values** expressing the importance of each dimension.

C: columns corresponding to original but m rows corresponding to singular values

Singular Value Decomposition

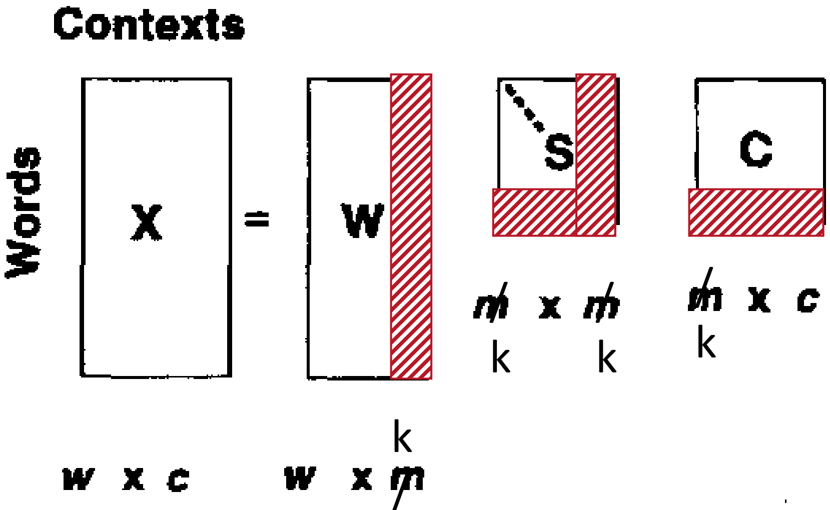


Landuaer and Dumais 1997

SVD applied to term-document matrix: Latent Semantic Analysis

Deerwester et al (1988)

- If instead of keeping all m dimensions, we just keep the top k singular values. Let's say 300.
- The result is a least-squares approximation to the original X
- But instead of multiplying, we'll just make use of W.
- Each row of W:
  - A k-dimensional vector
  - Representing word W



LSA more details

- 300 dimensions are commonly used
- The cells are commonly weighted by a product of two weights
  - Local weight: Log term frequency
  - Global weight: either idf or an entropy measure

Let’s return to PPMI word-word matrices

- Can we apply to SVD to them?

SVD applied to term-term matrix

$$\underset{|V| \times |V|}{\begin{bmatrix} & & \\ & & \\ & X & \\ & & \\ & & \end{bmatrix}} = \underset{|V| \times |V|}{\begin{bmatrix} & & \\ & & \\ & W & \\ & & \\ & & \end{bmatrix}} \underset{|V| \times |V|}{\begin{bmatrix} \sigma_1 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_V \end{bmatrix}} \underset{|V| \times |V|}{\begin{bmatrix} & & \\ & & \\ & C & \\ & & \\ & & \end{bmatrix}}$$

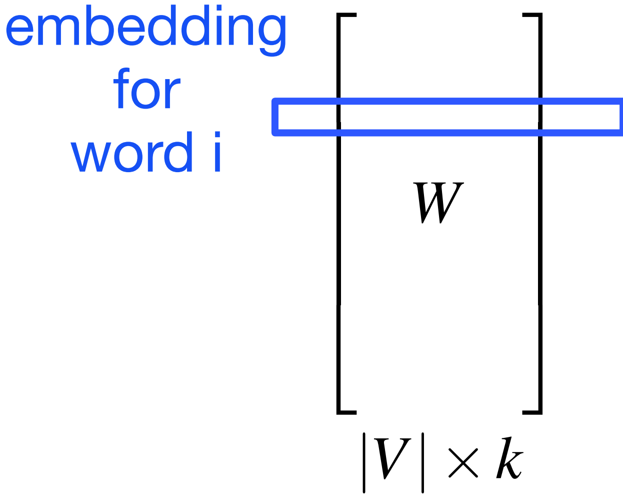
(I'm simplifying here by assuming the matrix has rank |V|)

Truncated SVD on term-term matrix

$$\underset{|V|\times|V|}{\begin{bmatrix} \\ \\ X \\ \\ \\ \end{bmatrix}} = \underset{|V|\times k}{\begin{bmatrix} \\ \\ W \\ \\ \\ \end{bmatrix}} \underset{k\times k}{\begin{bmatrix} \sigma_1 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_k \end{bmatrix}} \underset{k\times|V|}{\begin{bmatrix} & C & \end{bmatrix}}$$

Truncated SVD produces embeddings

- Each row of W matrix is a k-dimensional representation of each word *w*
- K might range from 50 to 1000
- Generally we keep the top k dimensions, but some experiments suggest that getting rid of the top 1 dimension or even the top 50 dimensions is helpful (Lapesa and Evert 2014).



Embeddings versus sparse vectors

- Dense SVD embeddings sometimes work better than sparse PPMI matrices at tasks like word similarity
  - Denoising: low-order dimensions may represent unimportant information
  - Truncation may help the models generalize better to unseen data.
  - Having a smaller number of dimensions may make it easier for classifiers to properly weigh the dimensions for the task.
  - Dense models may do better at capturing higher order co-occurrence.

Vector Semantics

Embeddings inspired by neural language models: skip-grams and CBOW

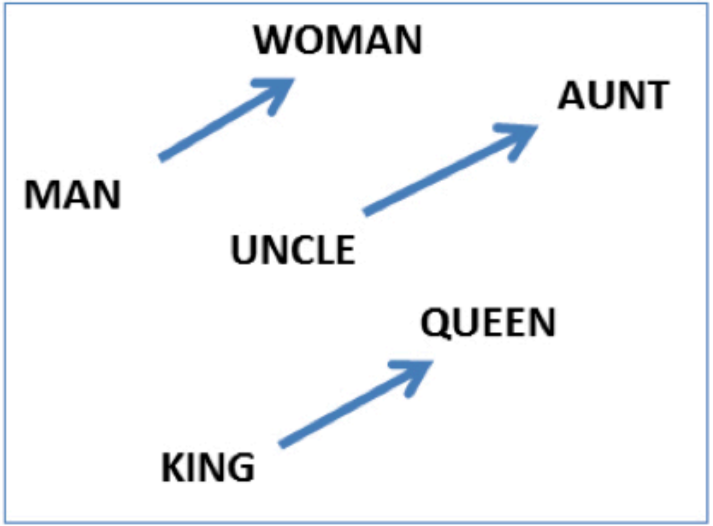
Prediction-based models: An alternative way to get dense vectors

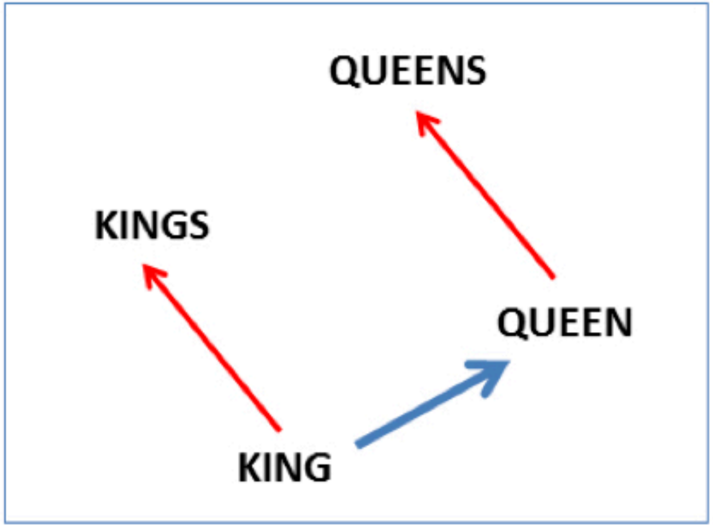
- **Skip-gram** (Mikolov et al. 2013a) **CBOW** (Mikolov et al. 2013b)
- Learn embeddings as part of the process of word prediction.
- Train a neural network to predict neighboring words
  - Inspired by **neural net language models**.
  - In so doing, learn dense embeddings for the words in the training corpus.
- Advantages:
  - Fast, easy to train (much faster than SVD)
  - Available online in the `word2vec` package
  - Including sets of pretrained embeddings!

Embeddings capture relational meaning!

vector(*‘king’*) - vector(*‘man’*) + vector(*‘woman’*) ≈ vector(‘queen’)

vector(*‘Paris’*) - vector(*‘France’*) + vector(*‘Italy’*) ≈ vector(‘Rome’)





Vector Semantics

Brown clustering

Brown clustering

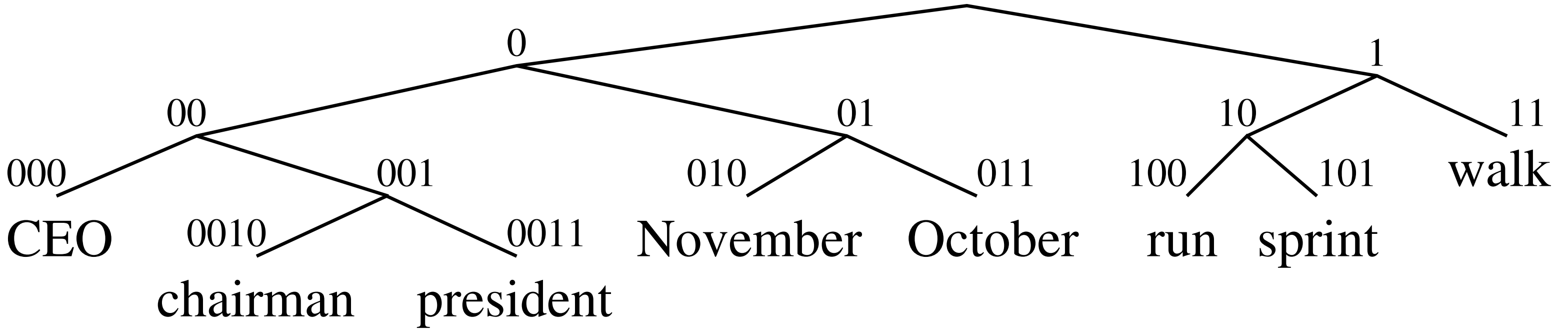
- An agglomerative clustering algorithm that clusters words based on which words precede or follow them
- These word clusters can be turned into a kind of vector
- We'll give a very brief sketch here.

Brown clustering algorithm

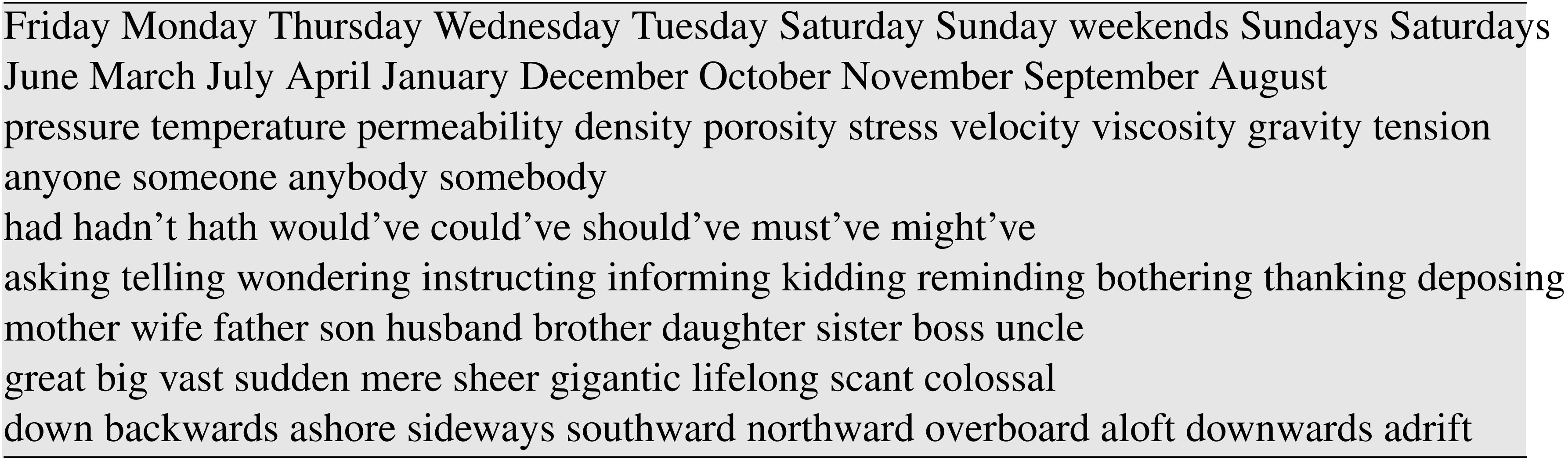
- Each word is initially assigned to its own cluster.
- We now consider merging each pair of clusters. Highest quality merge is chosen.
    - Quality = merges two words that have similar probabilities of preceding and following words
    - (More technically quality = smallest decrease in the likelihood of the corpus according to a class-based language model)
- Clustering proceeds until all words are in one big cluster.

Brown Clusters as vectors

- By tracing the order in which clusters are merged, the model builds a binary tree from bottom to top.
- Each word represented by binary string = path from root to leaf
- Each intermediate node is a cluster
- Chairman is 0010, “months” = 01, and verbs = 1



Brown cluster examples



Friday Monday Thursday Wednesday Tuesday Saturday Sunday weekends Sundays Saturdays

June March July April January December October November September August

pressure temperature permeability density porosity stress velocity viscosity gravity tension

anyone someone anybody somebody

had hadn’t hath would’ve could’ve should’ve must’ve might’ve

asking telling wondering instructing informing kidding reminding bothering thanking deposing

mother wife father son husband brother daughter sister boss uncle

great big vast sudden mere sheer gigantic lifelong scant colossal

down backwards ashore sideways southward northward overboard aloft downwards adrift

Class-based language model

- Suppose each word was in some class $c_{i:}$

$$P(w_i|w_{i-1}) = P(c_i|c_{i-1})P(w_i|c_i)$$

$$P(\text{corpus}|C) = \prod_{i-1}^{n} P(c_i|c_{i-1})P(w_i|c_i)$$

Vector Semantics

Evaluating similarity

Evaluating similarity

- Extrinsic (task-based, end-to-end) Evaluation:
  - Question Answering
  - Spell Checking
  - Essay grading
- Intrinsic Evaluation:
  - Correlation between algorithm and human word similarity ratings
    - Wordsim353: 353 noun pairs rated 0-10. *sim(plane,car)=5.77*
  - Taking TOEFL multiple-choice vocabulary tests
    - `Levied is closest in meaning to:`
      `imposed, believed, requested, correlated`

Summary

- Distributional (vector) models of meaning
  - **Sparse** (PPMI-weighted word-word co-occurrence matrices)
  - **Dense**:
    - Word-word SVD 50-2000 dimensions
    - Skip-grams and CBOW
    - Brown clusters 5-20 binary dimensions.