Word Meaning and Similarity

Word Senses and Word Relations

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**$_1$ can hold the investments in a custodial account…

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

Dan Jurafsky

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

Dan Jurafsky

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`)

Dan Jurafsky

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**:

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

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

Dan Jurafsky

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

  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

Synset

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

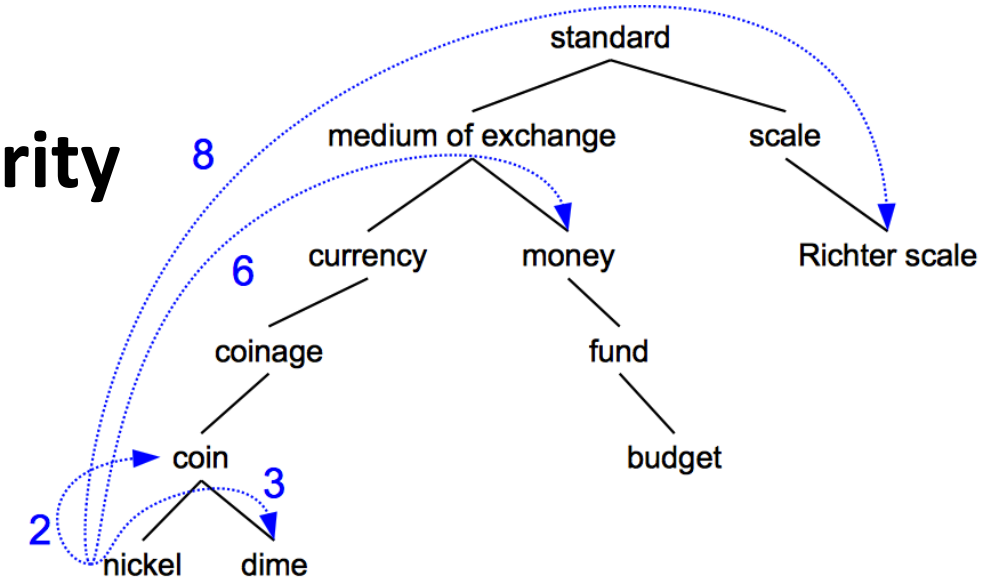
- We often distinguish **word similarity** from **word relatedness**
  - **Similar words**: near-synonyms
  - **Related words**: can be related any way
    - `car, bicycle`: **similar**
    - `car, gasoline`: **related**, not similar

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?

Dan Jurafsky

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)

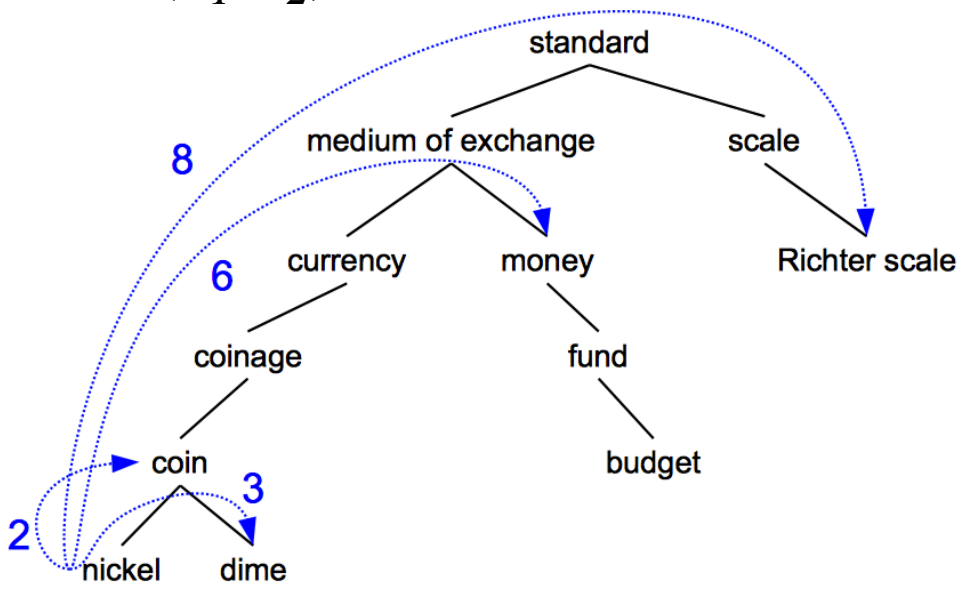
- $$\text{simpath}(c_1,c_2) = \frac{1}{\text{pathlen}(c_1,c_2)}$$

- $$\text{wordsim}(w_1,w_2) = \max_{c_1 \in \text{senses}(w_1), c_2 \in \text{senses}(w_2)} \text{sim}(c_1,c_2)$$

Dan Jurafsky

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

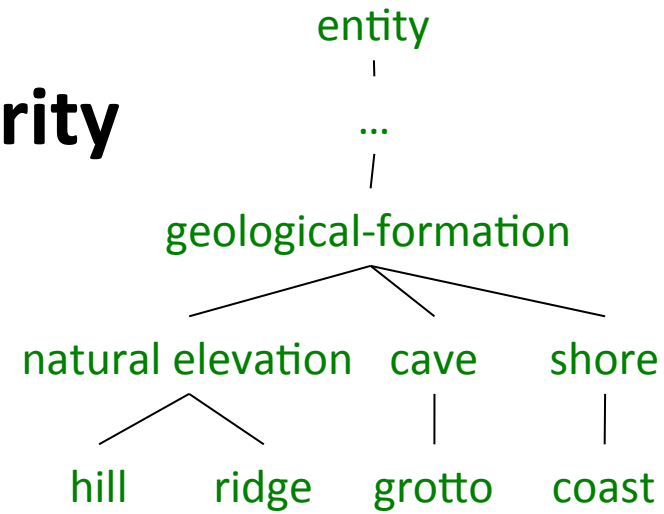
Information content similarity metrics

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)
    - $P(root)=1$
  - The lower a node in hierarchy, the lower its probability

Dan Jurafsky

Information content similarity



- Train by counting in a corpus
  - Each instance of `hill` counts toward frequency

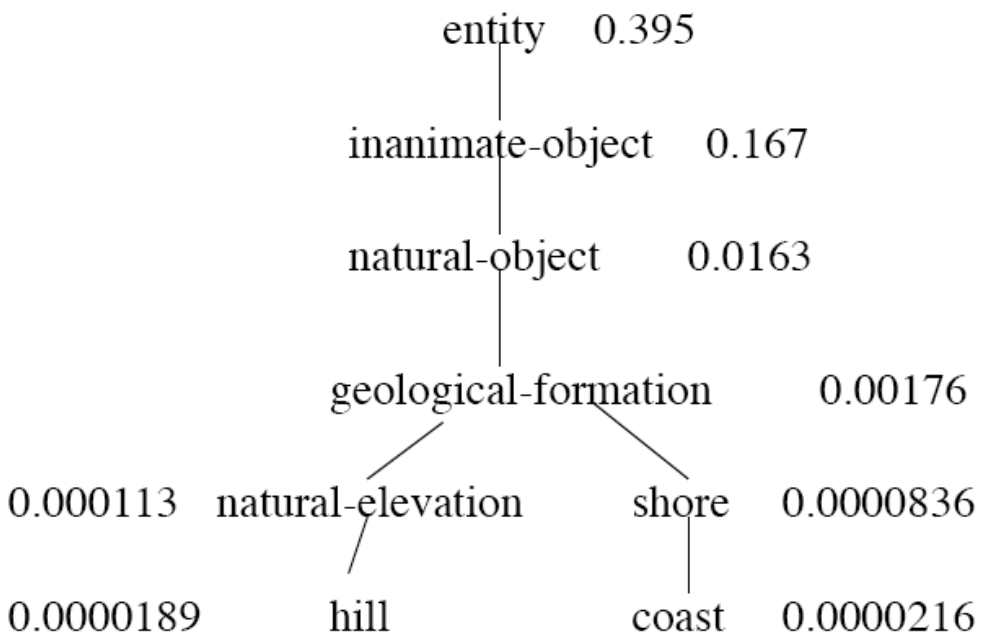
  of *natural elevation*, *geological formation*, *entity*, etc
  - Let $\text{words}(c)$ be the set of all words that are children of node c
    - words("geo-formation") = {hill,ridge,grotto,coast,cave,shore,natural elevation}
    - words("natural elevation") = {hill, ridge}

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

Information content similarity

- WordNet hierarchy augmented with probabilities P(c)

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



Information content: definitions

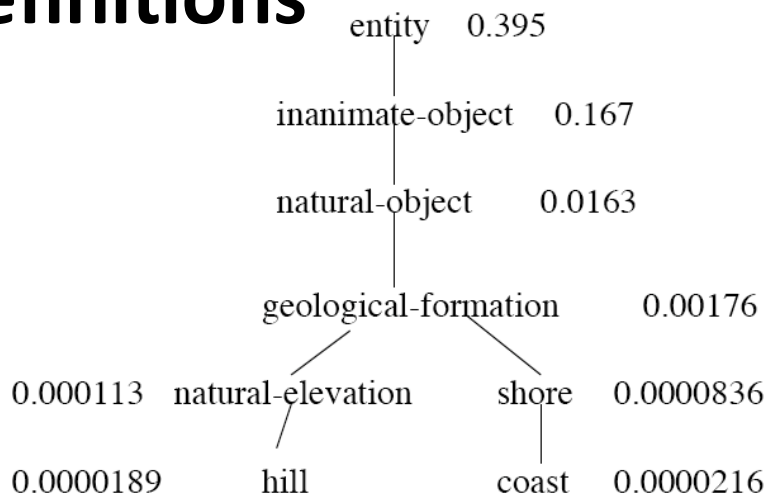
- Information content:

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

- Lowest common subsumer

  $LCS(c_1,c_2) =$

  The lowest node in the hierarchy that subsumes both c_1 and c_2



- How to use information content IC as a similarity metric?

Resnik method

- 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 lowest common subsumer of the two nodes
  - $\text{sim}_{\text{resnik}}(c_1,c_2) = -\log P(\text{LCS}(c_1,c_2))$

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

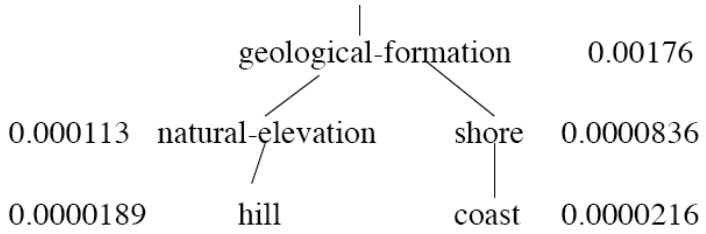
$$sim_{Lin}(A,B) \propto \frac{IC(common(A,B))}{IC(description(A,B))}$$

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

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

Dan Jurafsky

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

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

Levied is closest in meaning to:

imposed, believed, requested, correlated

Word Meaning and Similarity

Word Similarity: Thesaurus Methods