Lexical Semantics

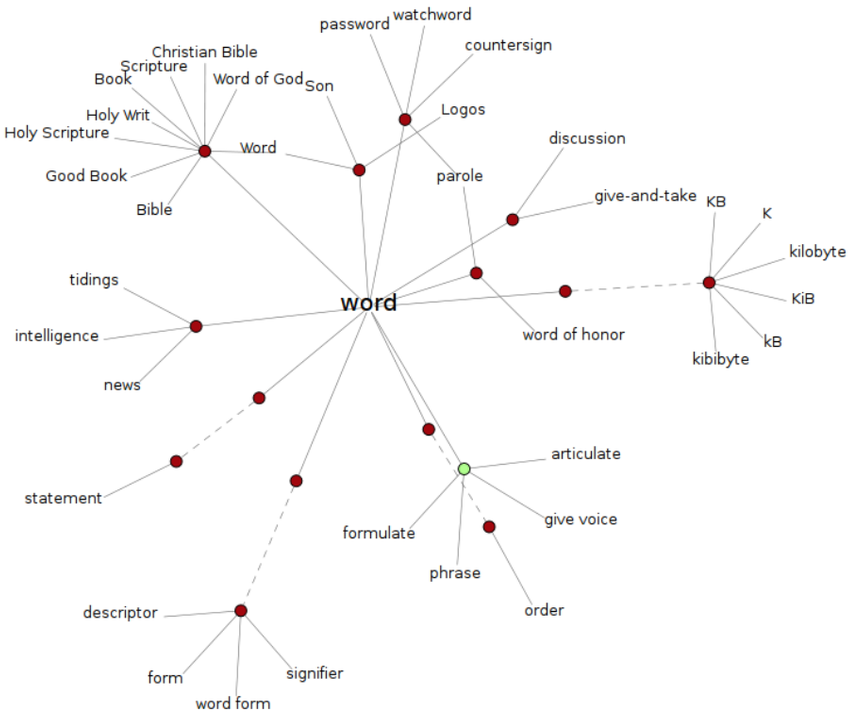
COMP90042

Natural Language Processing

Lecture 9

Semester 1 Week 5

Jey Han Lau



Sentiment Analysis

- Bag of words, kNN classifier. Training data:
  - “This is a good movie.” → ☺
  - “This is a great movie.” → ☺
  - “This is a terrible film.” → ☹
- “This is a wonderful film.” → ?
- Two problems:
  - The model does not know that “movie” and “film” are synonyms. Since “film” appears only in negative examples the model learns that it is a negative word.
  - “wonderful” is not in the vocabulary (OOV – Out-Of-Vocabulary).

Sentiment Analysis

- Comparing words directly will not work. How to make sure we compare word **meanings** instead?
- Solution: add this information explicitly through a **lexical database.**

Word Semantics

- Lexical semantics (this lecture)
  - How the meanings of words connect to one another.
  - Manually constructed resources: lexical database.
- Distributional semantics (next)
  - How words relate to each other in the text.
  - Automatically created resources from corpora.

Outline

- Lexical Database
- Word Similarity
- Word Sense Disambiguation

What Is Meaning?

- Their dictionary definition
  - But dictionary definitions are necessarily circular
  - Only useful if meaning is already understood

red *n.* the color of blood or a ruby.
blood *n.* the red liquid that circulates in the heart, arteries and veins of animals.

- Their relationships with other words
  - Also circular, but better for text analysis

Definitions

- A **word sense** describes one aspect of the meaning of a word

mouse[1] : .... a *mouse* controlling a computer system in 1968.
mouse[2] : .... a quiet animal like a *mouse*

Definitions

- A **word sense** describes one aspect of the meaning of a word
- If a word has multiple senses, it is **polysemous**

mouse^{1} : .... a *mouse* controlling a computer system in 1968.
mouse^{2} : .... a quiet animal like a *mouse*
bank^{1} : ...a *bank* can hold the investments in a custodial account ...
bank^{2} : ...as agriculture burgeons on the east *bank*, the river ...

Meaning Through Dictionary

- **Gloss**: textual definition of a sense, given by a dictionary
- *Bank*
  - financial institution that accepts deposits and channels the money into lending activities
  - sloping land (especially the slope beside a body of water)

Meaning Through Relations

- Another way to define meaning: by looking at how it relates to other words
- **Synonymy:** near identical meaning
  - *vomit* vs. *throw up*
  - *big* vs. *large*
- **Antonymy:** opposite meaning
  - *long* vs. *short*
  - *big* vs. *little*

Meaning Through Relations (2)

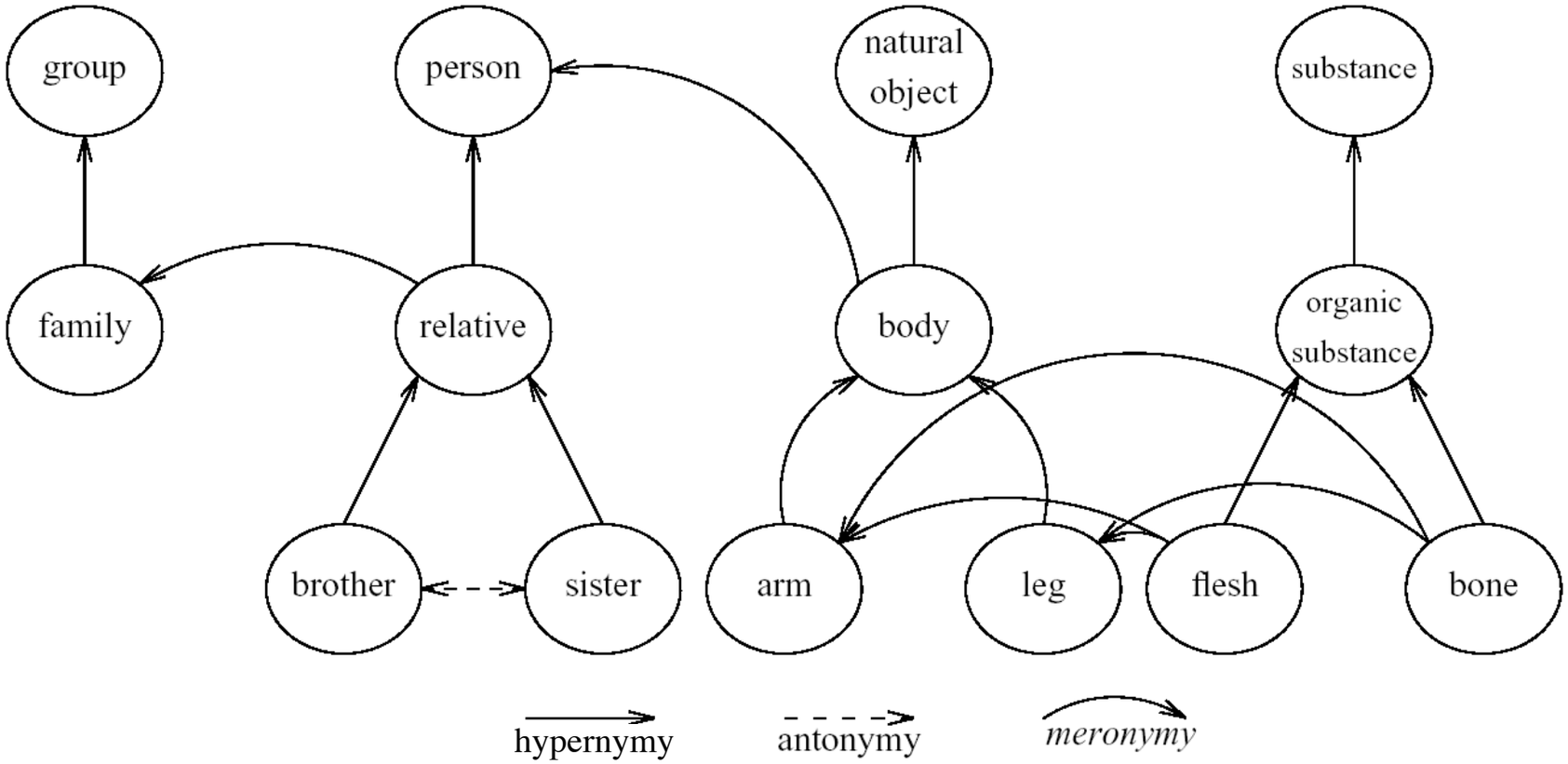
- **Hypernymy:** is-a relation
  - *cat* is an *animal*
  - *mango* is a *fruit*
- **Meronymy:** part-whole relation
  - *leg* is part of a *chair*
  - *wheel* is part of a *car*

What are the relations for these words?

- dragon and creature
- book and page
- comedy and tragedy

PollEv.com/jeyhanlau569

Meaning Through Relations (3)



WordNet

- A database of lexical relations
- English WordNet includes ~120,000 nouns, ~12,000 verbs, ~21,000 adjectives, ~4,000 adverbs
- On average: noun has 1.23 senses; verbs 2.16
- WordNets available in most major languages (www.globalwordnet.org, https://babelnet.org/)
- English version freely available (accessible via NLTK)

WordNet Example

The noun "bass" has 8 senses in WordNet.

1. bass[1] - (the lowest part of the musical range)
2. bass[2], bass part[1] - (the lowest part in polyphonic music)
3. bass[3], basso[1] - (an adult male singer with the lowest voice)
4. sea bass[1], bass[4] - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass[1], bass[5] - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass[6], bass voice[1], basso[2] - (the lowest adult male singing voice)
7. bass[7] - (the member with the lowest range of a family of musical instruments)
8. bass[8] - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Synsets

- Nodes of WordNet are not words or lemmas, but senses
- There are represented by sets of synonyms, or **synsets**
- *Bass* synsets:
  - {*bass*[1], *deep*[6]}
  - {*bass*[6], *bass voice*[1], *basso*[2]}
- Another synset:
  - {*chump*[1], *fool*[2], *gull*[1], *mark*[9], *patsy*[1], *fall guy*[1], *sucker*[1], *soft touch*[1], *mug*[2]}
  - Gloss: a person who is gullible and easy to take advantage of

Synsets (2)

>>> nltk.corpus.wordnet.synsets(‘bank’)

[Synset('bank.n.01'), Synset('depository_financial_institution.n.01'), Synset('bank.n.03'), Synset('bank.n.04'), Synset('bank.n.05'), Synset('bank.n.06'), Synset('bank.n.07'), Synset('savings_bank.n.02'), Synset('bank.n.09'), Synset('bank.n.10'), Synset('bank.v.01'), Synset('bank.v.02'), Synset('bank.v.03'), Synset('bank.v.04'), Synset('bank.v.05'), Synset('deposit.v.02'), Synset('bank.v.07'), Synset('trust.v.01')]

>>> nltk.corpus.wordnet.synsets(‘bank’)[0].definition()

u'sloping land (especially the slope beside a body of water)‘

>>> nltk.corpus.wordnet.synsets(‘bank’)[1].lemma_names()

[u'depository_financial_institution', u'bank', u'banking_concern', u'banking_company']

Hypernymy Chain

bass[3], basso (an adult male singer with the lowest voice)
=> singer, vocalist, vocalizer, vocaliser
=> musician, instrumentalist, player
=> performer, performing artist
=> entertainer
=> person, individual, someone...
=> organism, being
=> living thing, animate thing,
=> whole, unit
=> object, physical object
=> physical entity
=> entity

bass[7] (member with the lowest range of a family of instruments)
=> musical instrument, instrument
=> device
=> instrumentality, instrumentation
=> artifact, artefact
=> whole, unit
=> object, physical object
=> physical entity
=> entity

Word Similarity

Word Similarity

- Synonymy: *film* vs. *movie*
- What about *show* vs. *film*? *opera* vs. *film*?
- Unlike synonymy (which is a binary relation), word similarity is a spectrum
- We can use lexical database (e.g. WordNet) or thesaurus to estimate word similarity

Word Similarity with Paths

- Given WordNet, find similarity based on path length
- pathlen(c_1,c_2) = 1+ edge length in the shortest path between sense c_1 *and* c_2

Remember that a node in the Wordnet graph is a synset (sense), not a word!

- similarity between two senses (synsets)

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

- similarity between two words

$$\text{wordsim}(w_1, w_2) = \max_{c_1 \in \text{senses}(w_1), c2 \in \text{senses}(w_2)} \text{simpath}(c_1, c_2)$$

Examples

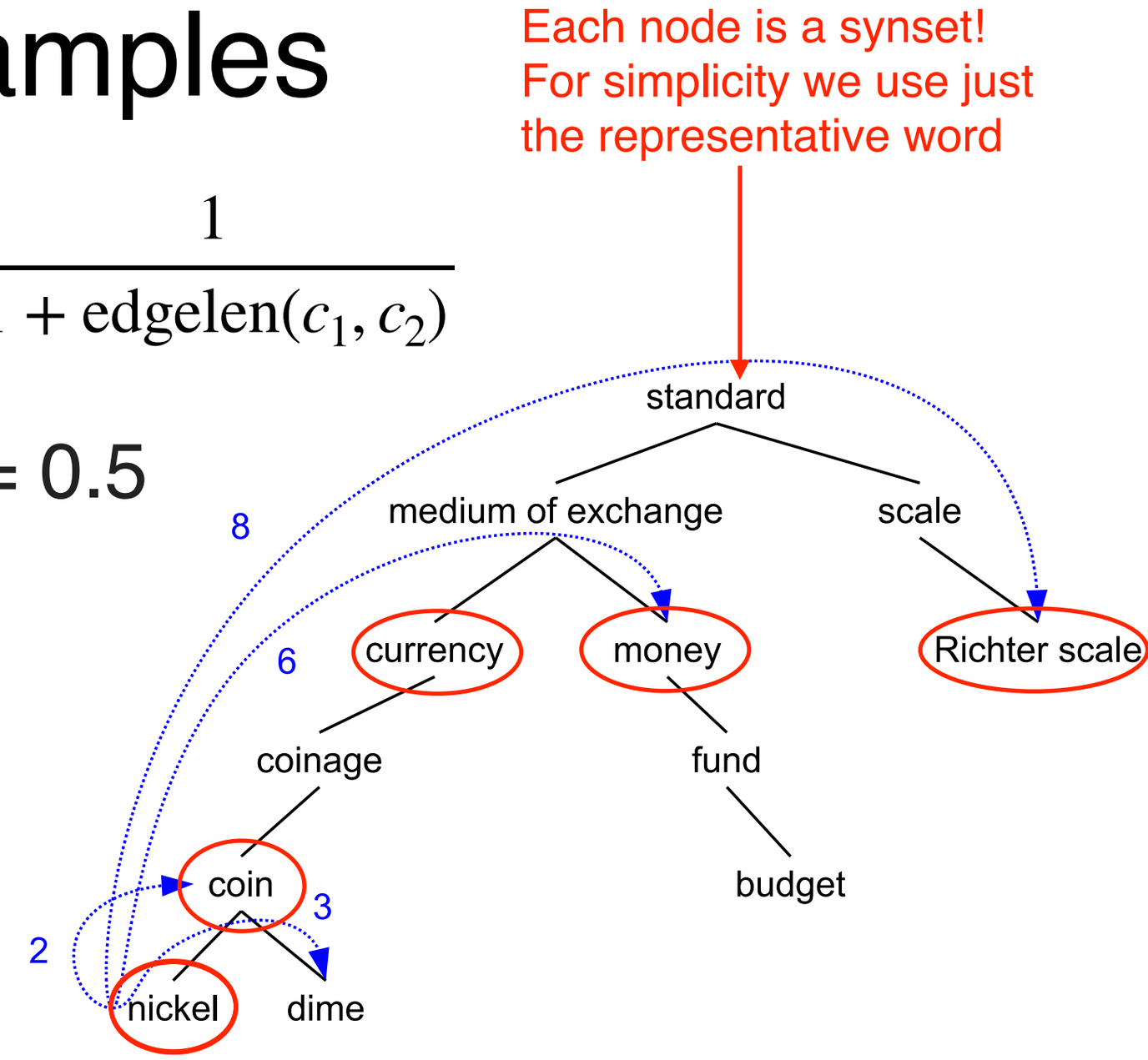
$$\text{simpath}(c_1, c_2) = \frac{1}{\text{pathlen}(c_1, c_2)} = \frac{1}{1 + \text{edgelen}(c_1, c_2)}$$

simpath(*nickel,coin*) = 1/2 = 0.5

simpath(*nickel,currency*)
= 1/4 = 0.25

simpath(*nickel,money*)
= 1/6 = 0.17

simpath(*nickel,Richter scale*)
= 1/8 = 0.13



Beyond Path Length

- simpath(nickel,money) = 0.17
- simpath(nickel,Richter scale) = 0.13
- Problem: edges vary widely in actual semantic distance
  - Much bigger jumps near top of hierarchy
- Solution 1: include depth information (Wu & Palmer)
  - Use path to find lowest common subsumer (LCS)
  - Compare using depths

$$\text{simwup}(c_1, c_2) = \frac{2 \times \text{depth}(\text{LCS}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)}$$

High simwup when:
- parent is deep
- senses are shallow

Examples

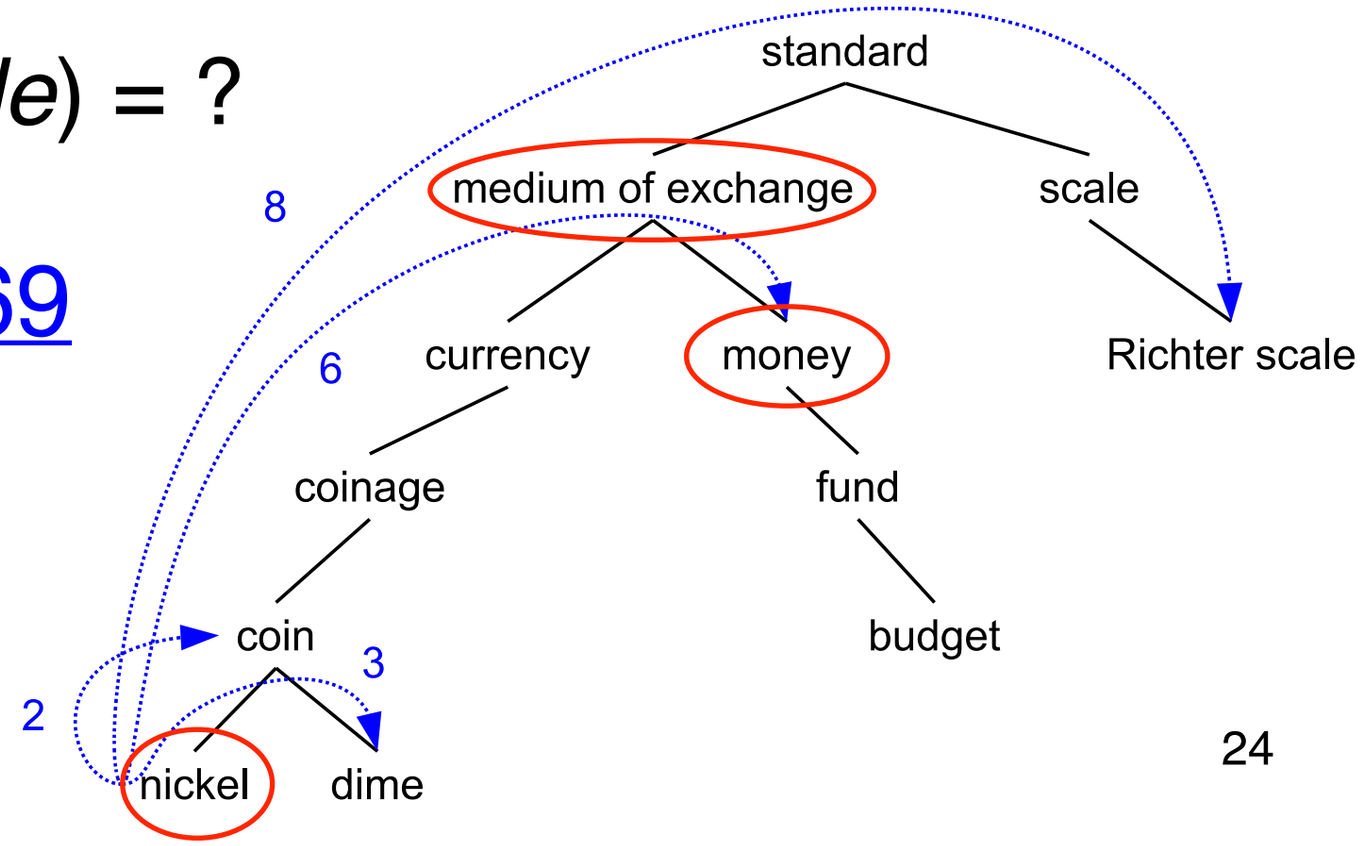
$$\text{simwup}(c_1, c_2) = \frac{2 \times \text{depth}(\text{LCS}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)}$$

simwup(*nickel,money*) =

2*2 / (6+3) = 0.44

simwup(*dime,Richter scale*) = ?

PollEv.com/jeyhanlau569



Examples

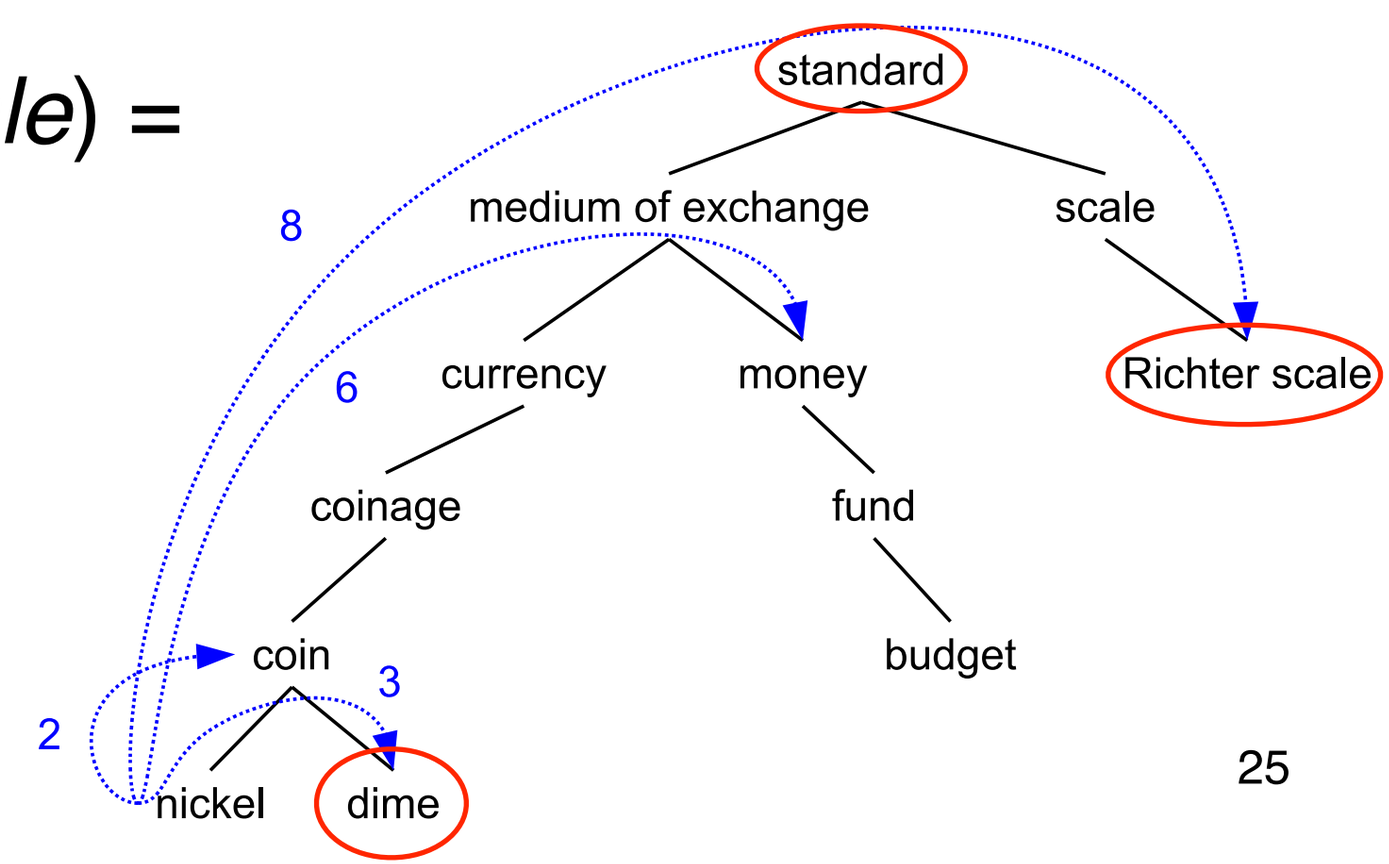
$$\text{simwup}(c_1, c_2) = \frac{2 \times \text{depth}(\text{LCS}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)}$$

simwup(*nickel,money*) =

2*2 / (6+3) = 0.44

simwup(*dime,Richter scale*) =

2*1 / (6+3) = 0.22



Abstract Nodes

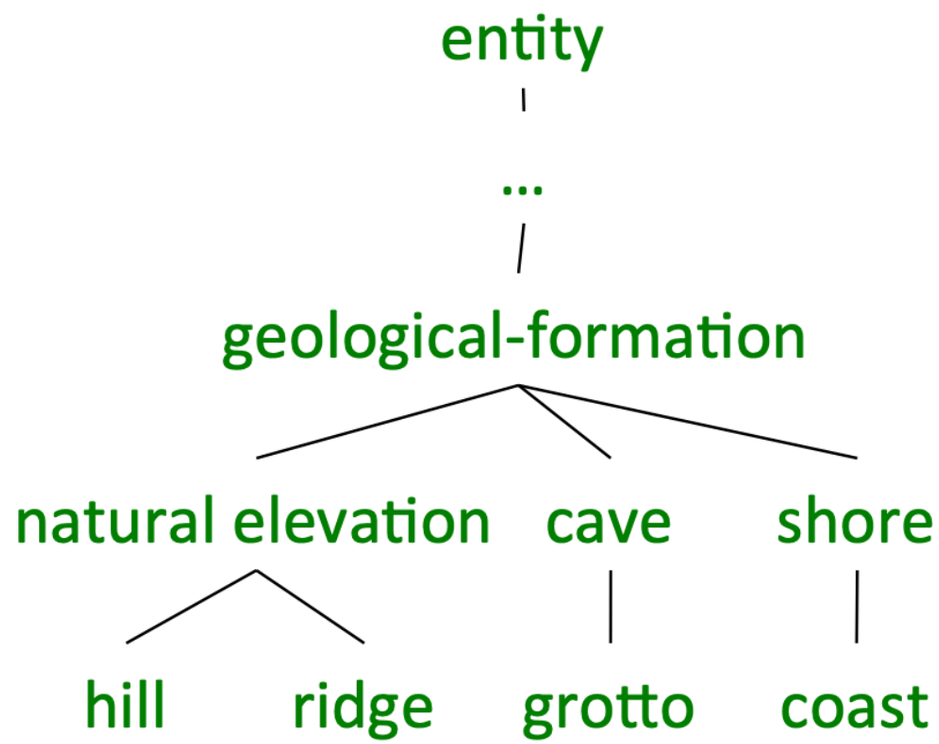
- But node depth is still poor semantic distance metric
  - simwup(*nickel,money*) = 0.44
  - simwup(*nickel,Richter scale*) = 0.22
- Nodes high in the hierarchy is very abstract or general
- How to better capture them?

Concept Probability Of A Node

- Intuition:
  general node → high concept probability (e.g. *object*)
  narrow node → low concept probability (e.g. *vocalist*)
- Find all the children ∈ node, and sum up their unigram probabilities!

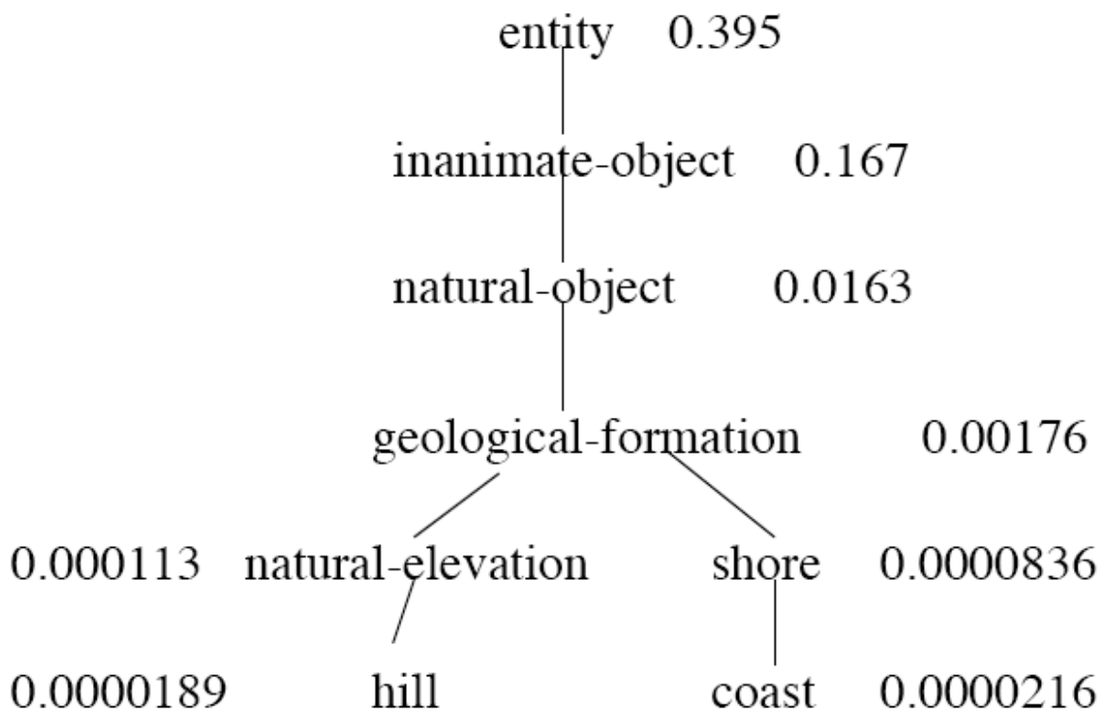
$$P(c) = \frac{\sum_{s \in \text{child}(c))} \text{count}(s)}{N}$$

- child(c): synsets that are children of *c*
- child(*geological-formation*) =
  {*hill*, *ridge*, *grotto*, *coast*,
  natural elevation, cave, shore}
- child(*natural elevation*) =
  {*hill*, *ridge*}



Example

- Abstract nodes higher in the hierarchy has a higher P(c)



Similarity with Information Content

use IC instead of depth (simwup)

general concept = small values
narrow concept = large values

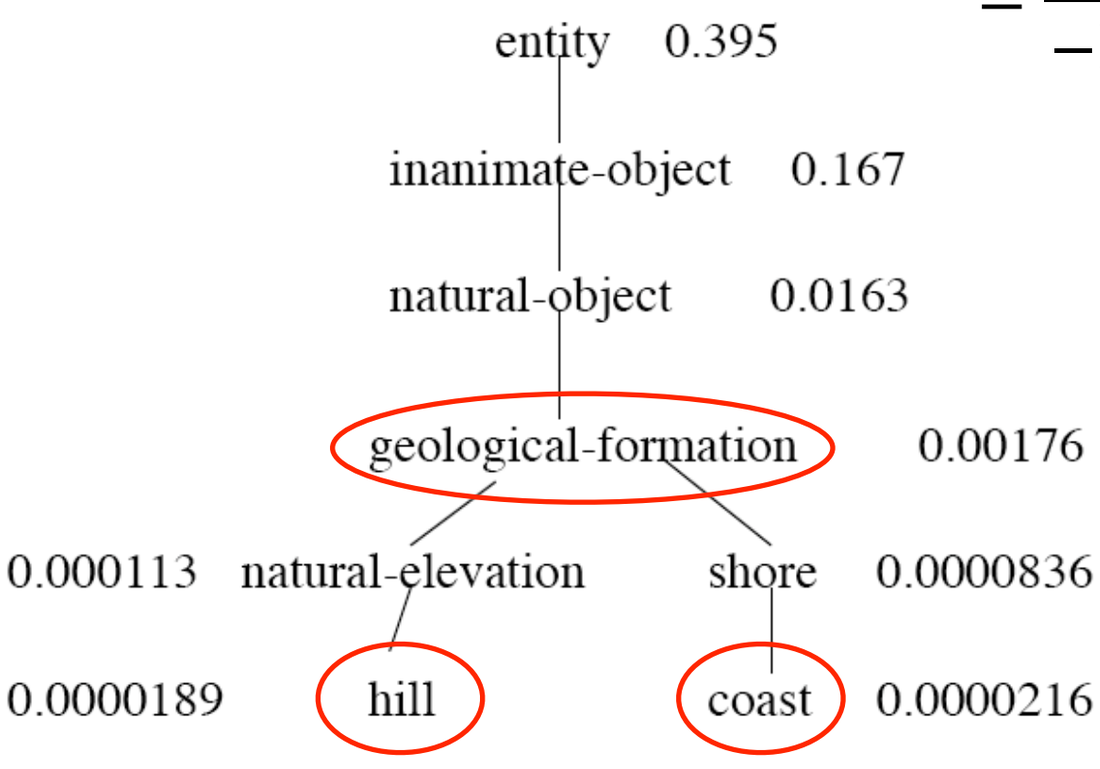
$$\text{IC} = -\log P(c)$$

$$\text{simlin}(c_1, c_2) = \frac{2 \times \text{IC}(\text{LCS}(c_1, c_2))}{\text{IC}(c_1) + \text{IC}(c_2)}$$

high simlin when:
- concept of parent is narrow
- concept of senses are general

$$\text{simlin}(\text{hill}, \text{coast}) = \frac{2 \times -\log P(\text{geological-formation})}{-\log P(\text{hill}) - \log P(\text{coast})}$$

$$= \frac{-2 \log 0.00176}{-\log 0.0000189 - \log 0.0000216} = 0.587$$



If LCS is *entity*, -2 log(0.395)!
simlin = 0.086

Word Sense Disambiguation

Word Sense Disambiguation

- Task: selects the correct sense for words in a sentence
- Baseline:
  - Assume the most popular sense
- Good WSD potentially useful for many tasks
  - Knowing which sense of *mouse* is used in a sentence is important!
  - Less popular nowadays; because sense information is implicitly captured by contextual representations (lecture 11)

Supervised WSD

- Apply standard machine classifiers
- Feature vectors typically words and syntax around target
  - But context is ambiguous too!
  - How big should context window be? (in practice small)
- Requires sense-tagged corpora
  - E.g. SENSEVAL, SEMCOR (available in NLTK)
  - Very time consuming to create!

Unsupervised: Lesk

- Lesk: Choose sense whose WordNet gloss overlaps most with the context
- *The **bank** can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.*
- ***bank*¹**: 2 overlapping non-stopwords, *deposits* and *mortgage*
- ***bank*²**: 0

bank[1]	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	"he cashed a check at the bank", "that bank holds the mortgage on my home"
bank[2]	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	"they pulled the canoe up on the bank", "he sat on the bank of the river and watched the currents"

Final Words

- Creation of lexical database involves expert curation (linguists)
- Modern methods attempt to derive semantic information directly from corpora, without human intervention
- Distributional semantics (next lecture!)

Reading

- JM3 Ch 23-23.4.1