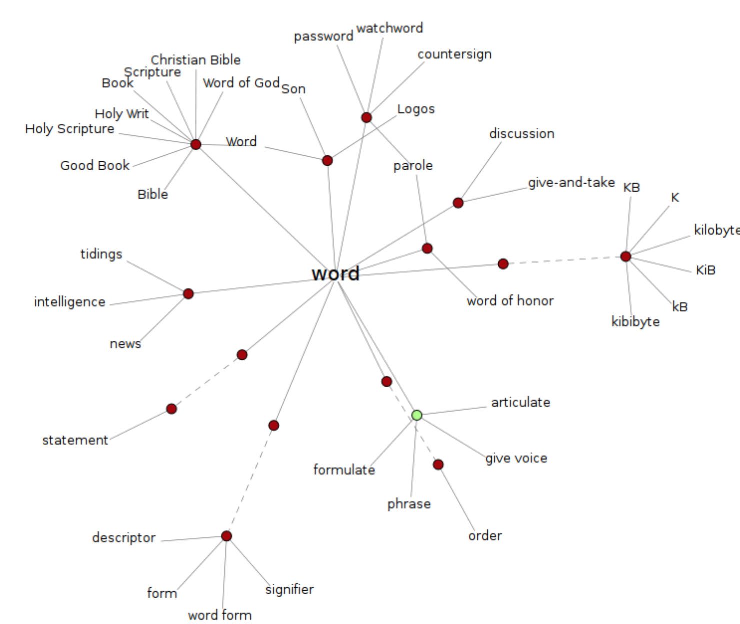
Lexical Semantics

COMP90042

Natural Language Processing

Lecture 9



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: lexicons, thesauri, ontologies, etc.
- Distributional semantics (next)
  - How words relate to each other in the text.
  - Automatically created resources from corpora.

What Do Words Mean?

- Referents in the physical or social world
  - But not usually useful in text analysis
- 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 more practical

Word Senses

- 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*
bank[1] : ...a *bank* can hold the investments in a custodial account ...
bank[2] : ...as agriculture burgeons on the east *bank*, the river ...

Word Glosses

- 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)
- If a word has multiple senses, it is **polysemous**

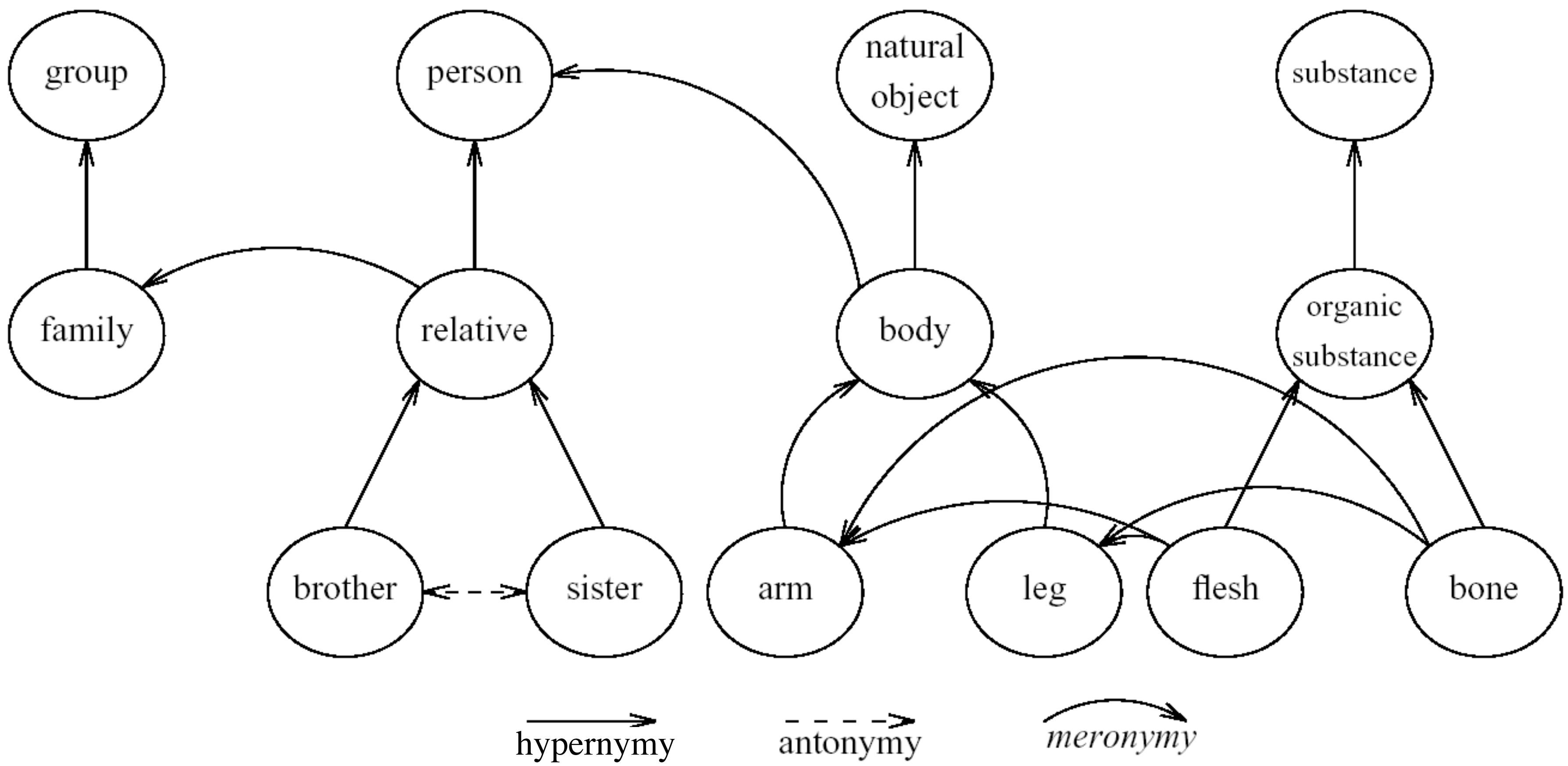
Meaning Through Relations

- 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*

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)

J&M3 Ch. 19

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']
```

Lexical Relations in WordNet

Relation	Also Called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	*breakfast*^{1} $\rightarrow$ *meal*^{1}
Hyponym	Subordinate	From concepts to subtypes	*meal*^{1} $\rightarrow$ *lunch*^{1}
Instance Hypernym	Instance	From instances to their concepts	*Austen*^{1} $\rightarrow$ *author*^{1}
Instance Hyponym	Has-Instance	From concepts to their instances	*composer*^{1} $\rightarrow$ *Bach*^{1}
Part Meronym	Has-Part	From wholes to parts	*table*^{2} $\rightarrow$ *leg*^{3}
Part Holonym	Part-Of	From parts to wholes	*course*^{7} $\rightarrow$ *meal*^{1}
Antonym		Semantic opposition between lemmas	*leader*^{1} $\Longleftrightarrow$ *follower*^{1}
Derivation		Lemmas w/same morphological root	*destruction*^{1} $\Longleftrightarrow$ *destroy*^{1}

Hypernymy Chain

```
bass³, 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
bass⁷ (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
- similarity between two senses:
  - $\text{simpath}(c_1, c_2) = \dfrac{1}{\text{pathlen}(c_1, c_2)}$
- similarity between two words
  - $\text{wordsim}(w_1, w_2) = \max\limits_{c_1 \in \text{senses}(w_1), c2 \in \text{senses}(w_2)} \text{simpath}(c_1, c_2)$

Examples

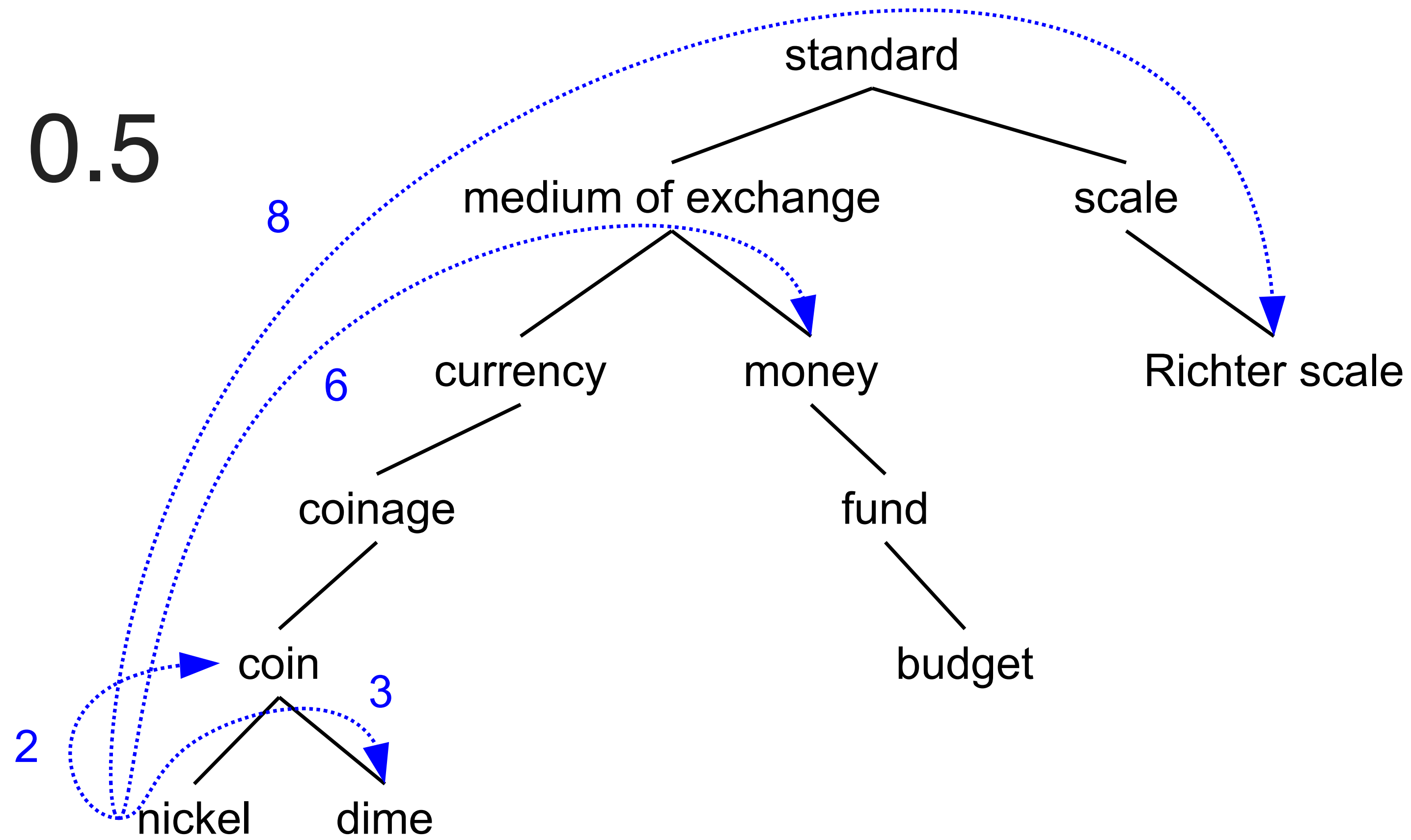
$$\text{simpath}(c_1, c_2) = \frac{1}{\text{pathlen}(c_1, c_2)}$$

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

- 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

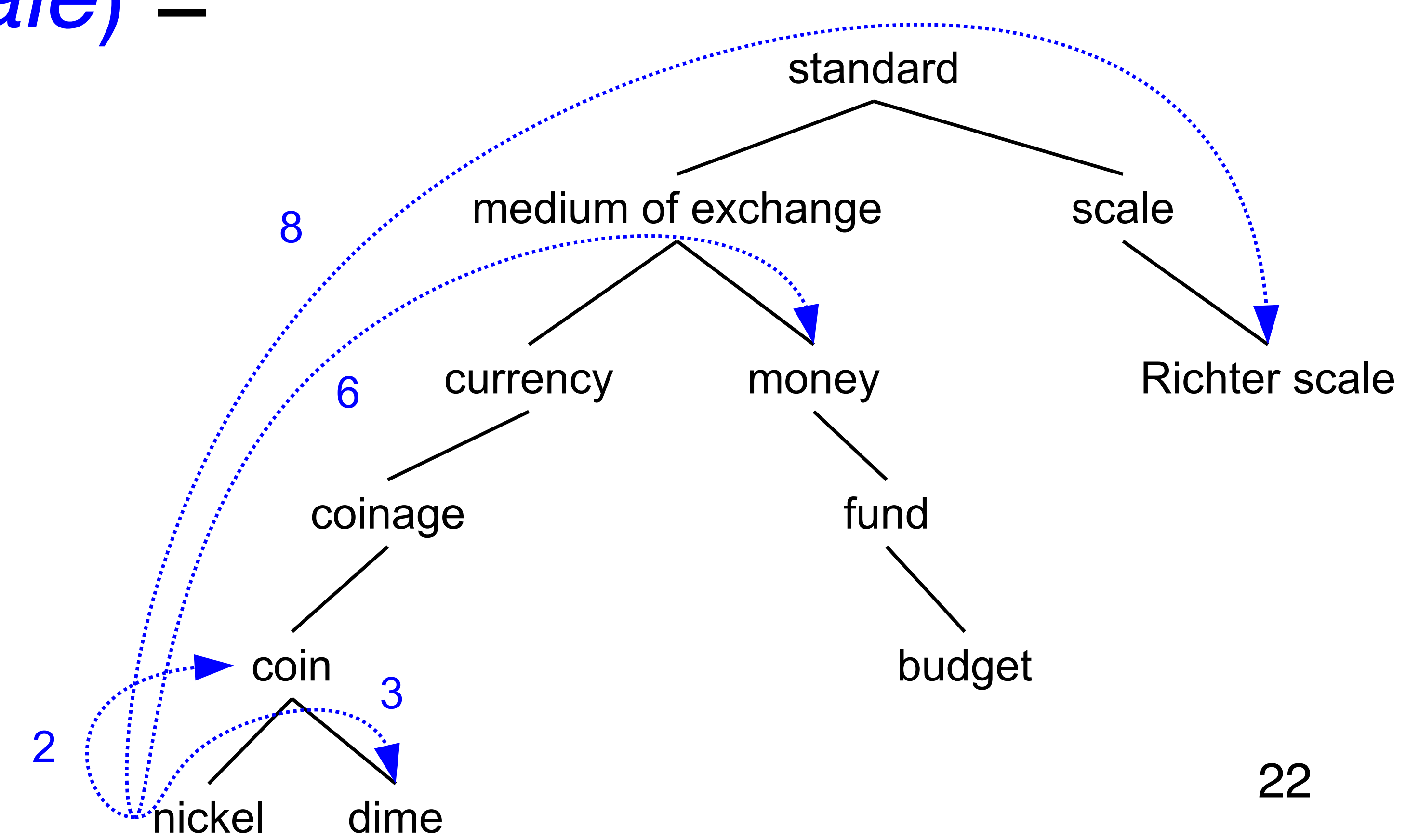
$$\text{simwup}(c_1,c_2) = \frac{2 * \text{depth}(\text{LCS}(c1,c2))}{\text{depth}(c1) + \text{depth}(c2)}$$

Examples

$$\text{simwup}(c_1,c_2) = \frac{2 * \text{depth}(\text{LCS}(c1,c2))}{\text{depth}(c1) + \text{depth}(c2)}$$

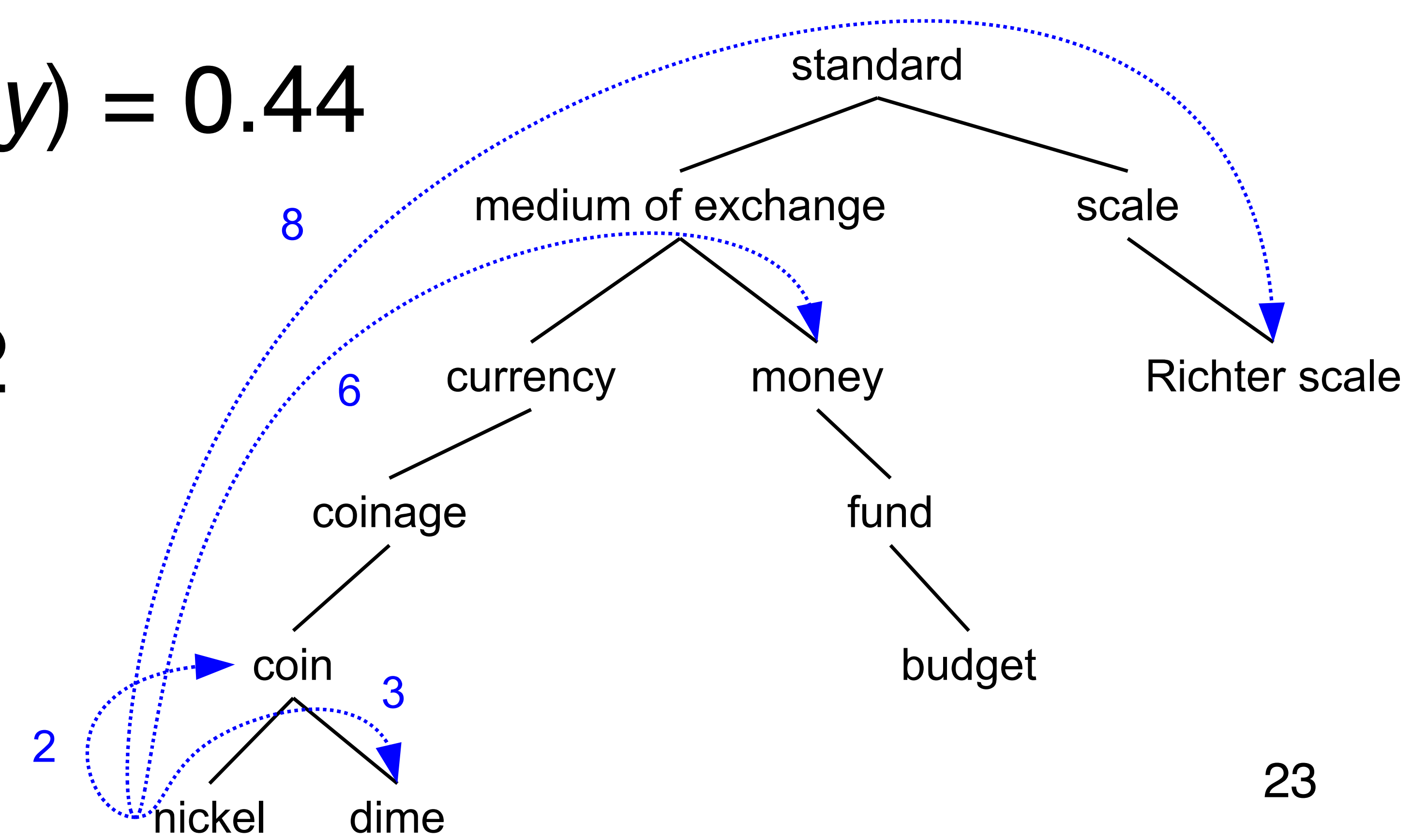
simwup(*nickel,money*) = 2*2/(3+ 6) = 0.44

simwup(*nickel,Richter scale*) =
2*1/(3 + 6) = 0.22



Abstract Nodes

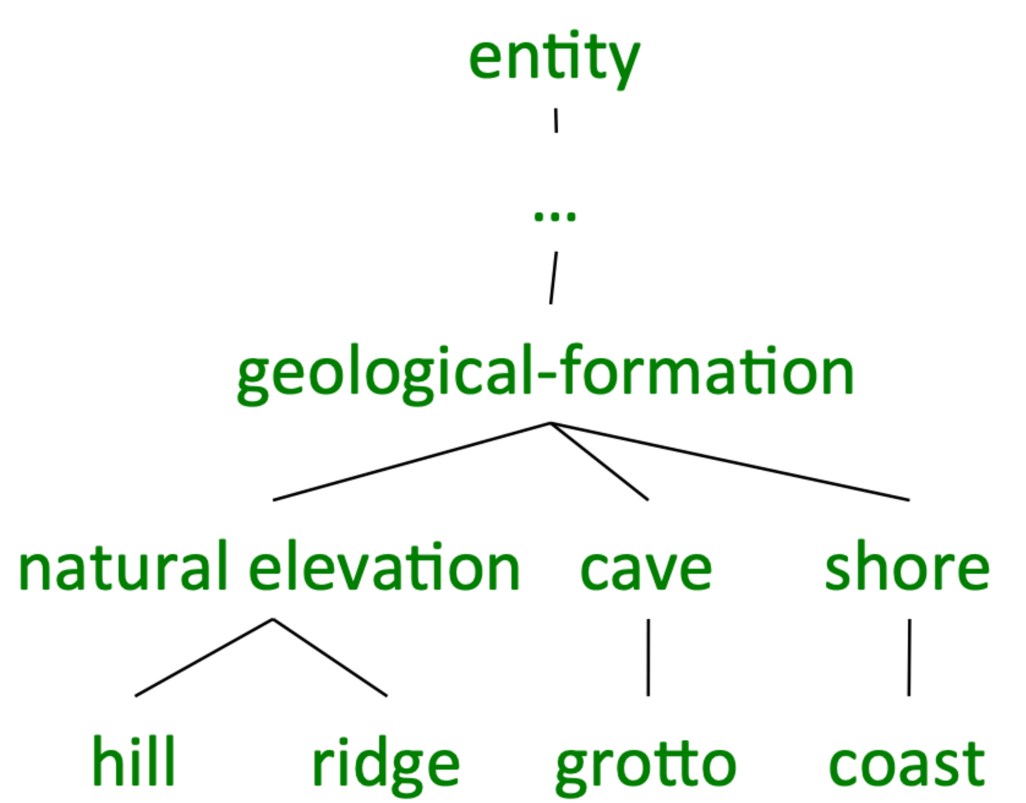
- But count of edges or node depth is still poor semantic distance metric
- Nodes high in the hierarchy is very abstract/general
- How do we make words that connect through very abstract nodes much less similar
    - simwup(*nickel,money*) = 0.44
    - simwup(*nickel, Richter scale*) = 0.22



Concept Probability

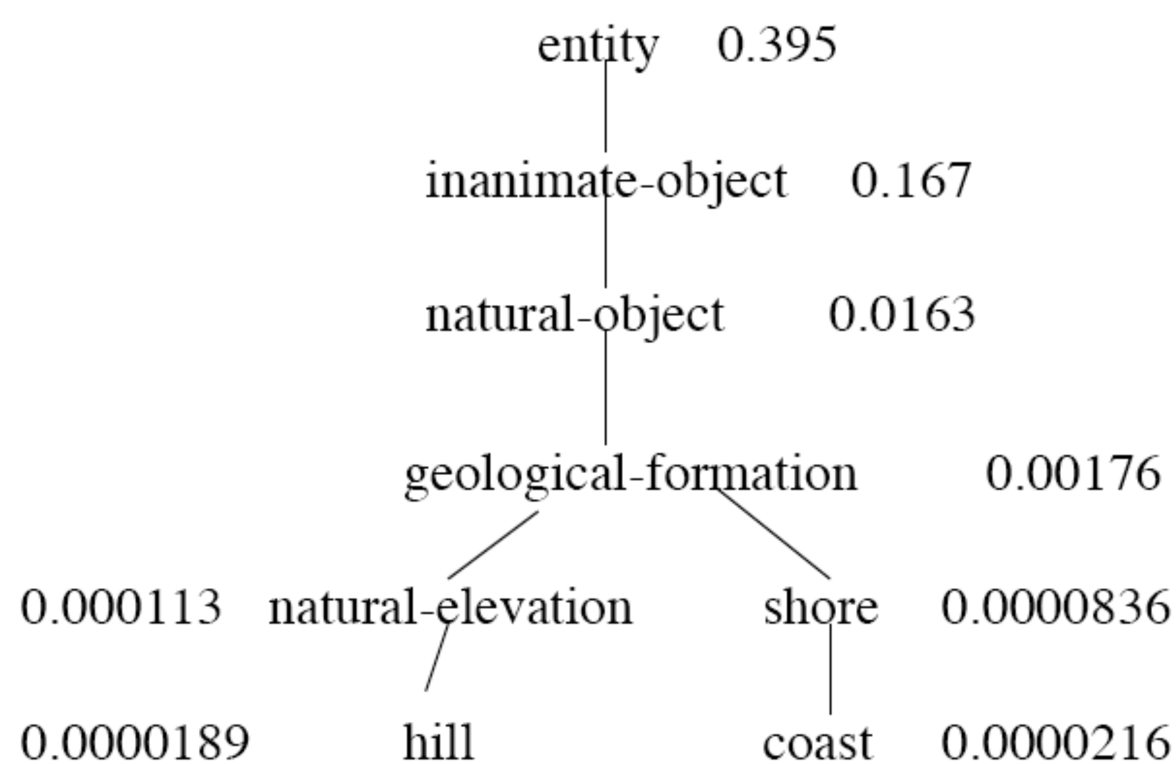
$$P(c) = \frac{\sum_{w \in \text{words}(c))} \text{count}(w)}{N}$$

- P(c): probability that a randomly selected word in a corpus is an instance of concept *c*
- words(c): set of all words that are children of *c*
- words(*geological-formation*) = {*hill*, *ridge*, *grotto*, *coast*, *natural elevation, cave, shore*}
- words(*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 in simwup

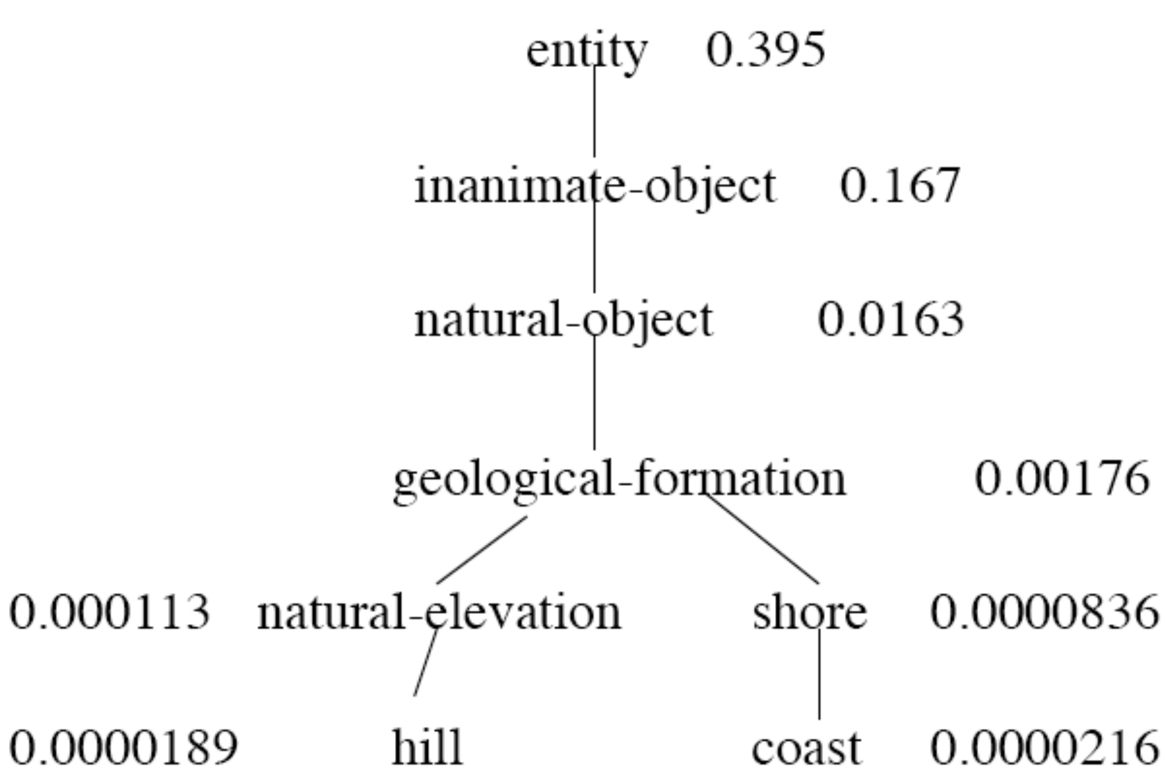
$$IC(c) = -\log P(c)$$

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

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

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

$$= 0.587$$



if LCS node is very high up in the hierarchy (say P(c) = 0.99), then IC will be very low (0.01 in this case)

Sentiment Analysis Revisited

- "This is a great movie." → ☺️
- "This is a wonderful film." → ?
- Comparing words using WordNet paths work well if our classifier is based on word similarities (such as kNN)
- But what if we want sense as a general feature representation, so we can employ other classifiers?
- Solution: map words in text to senses in WordNet explicitly.

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 in NLP
  - In practice, often ignored because good WSD too hard
  - Active research area

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? (typically very small)
- Requires sense-tagged corpora
  - E.g. SENSEVAL, SEMCOR (available in NLTK)
  - Very time consuming to create!

Less Supervised Approaches

- Lesk: Choose sense whose dictionary gloss from WordNet most overlaps 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"

Other Databases - FrameNet

- Based on **frame semantics**
  - *Mary bought a car from John*
  - *John sold a car to Mary*
  - Same situation (semantic frame), just different perspective
- A lexical database of **frames**, typically prototypical situations
  - E.g. “commerce_buy”, “apply_heat”

FrameNet

- Includes lists of *lexical units* that evoke the frame
  - E.g. *cook*, *fry*, *bake*, *boil*, etc.
- Lists of *semantic roles* or *frame elements*
  - E.g. “the cook”, “the food”, “the container”, “the instrument”
- Semantic relationships among frames
  - “apply_heat” is Causative of “absorb_heat”, is Used by “cooking_creation”

Moving On To The Corpus

- Manually-tagged lexical resources an important starting point for text analysis
- But much modern work attempts to derive semantic information directly from corpora, without human intervention
- Distributional semantics!

Reading

- JM3 Ch 19.1-19.3, 19.4.1, 19.5.1