

Lexical Semantics

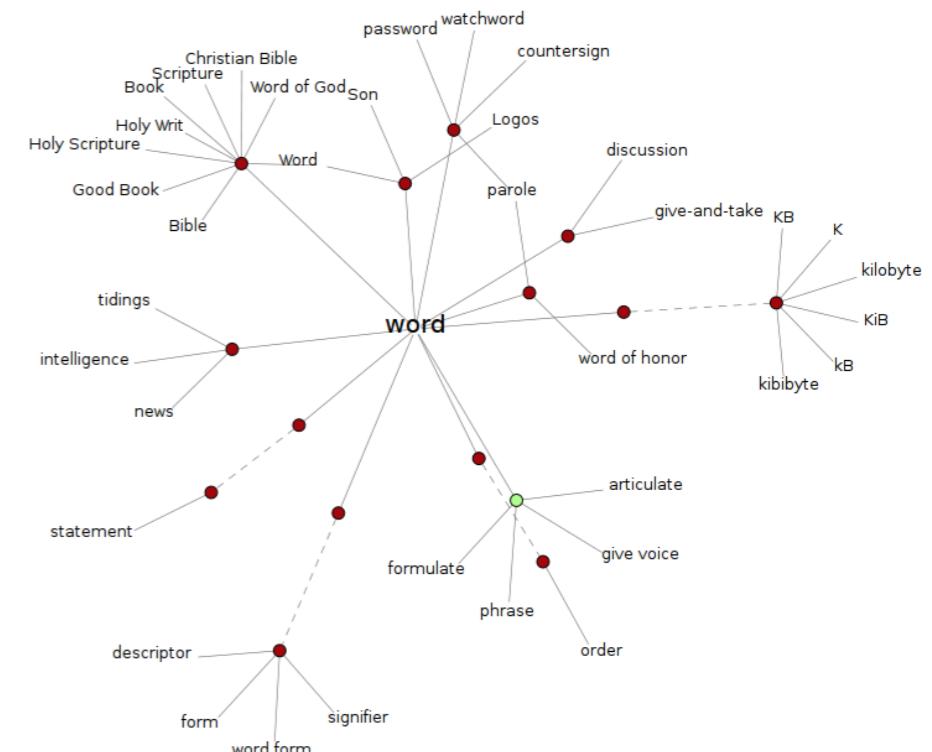
COMP90042

Natural Language Processing Lecture 9

Semester 1 2022 Week 5
Jey Han Lau



THE UNIVERSITY OF
MELBOURNE



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

mouse¹ : a *mouse* controlling a computer system in 1968.

mouse² : 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**

mouse¹ : a *mouse* controlling a computer system in 1968.

mouse² : a quiet animal like a *mouse*

bank¹ : ...a *bank* can hold the investments in a custodial account ...

bank² : ...as agriculture burgeons on the east *bank*, the river ...

Meaning Through Dictionary

- **Gloss:** **textual definition** of a sense, given by a dictionary
- *Bank* **bank¹** : ...a *bank* can hold the investments in a custodial account ...
 bank² : ...as agriculture burgeons on the east *bank*, the river ...
 - ▶ 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?

Hypernymy

- dragon and creature

Meronymy

- book and page

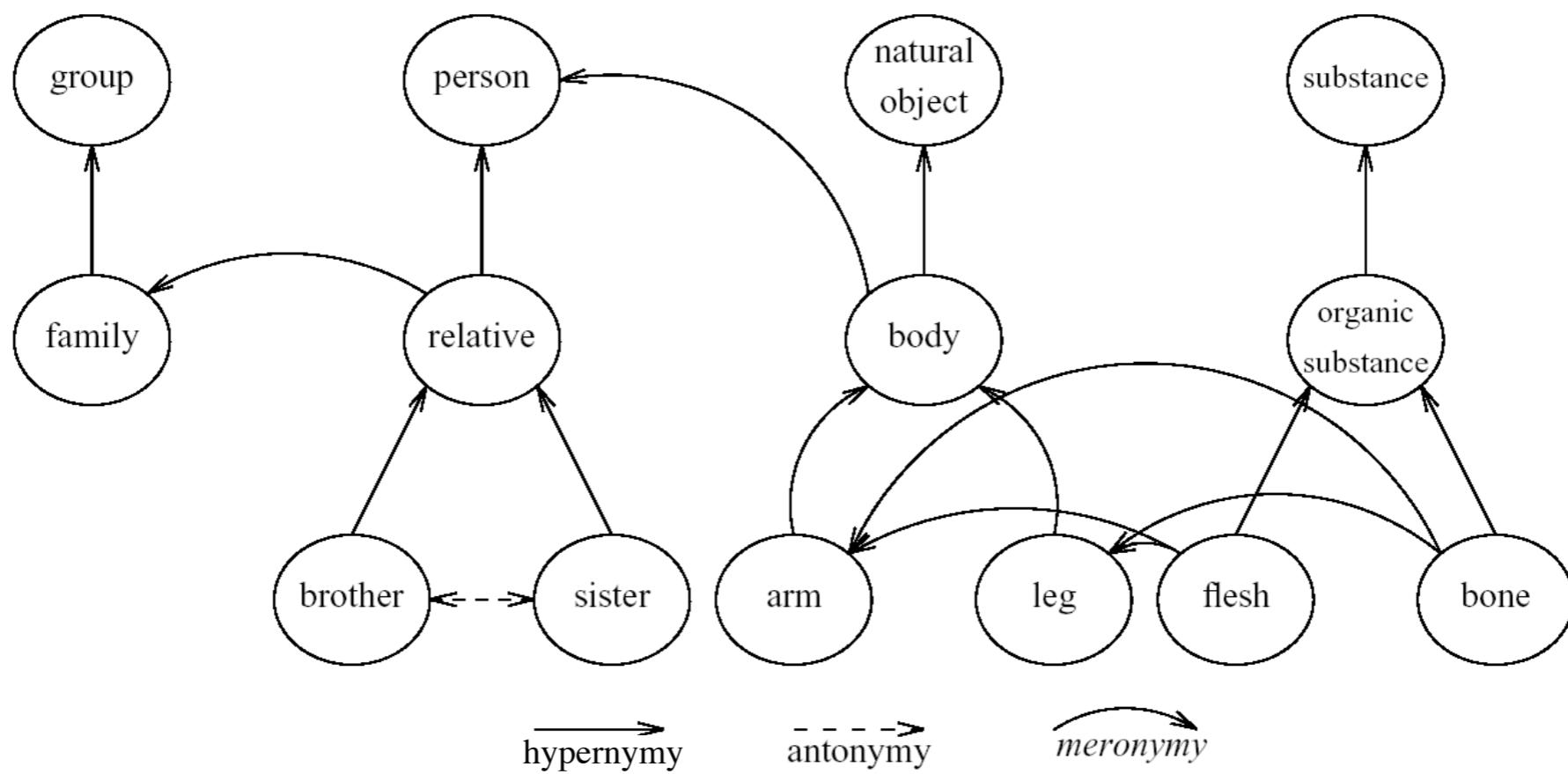
Antonymy

- 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. bass¹ - (the lowest part of the musical range) ←
2. bass², bass part¹ - (the lowest part in polyphonic music) ←
3. bass³, basso¹ - (an adult male singer with the lowest voice) ←
4. sea bass¹, bass⁴ - (the lean flesh of a saltwater fish of the family Serranidae) ←
5. freshwater bass¹, bass⁵ - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus)) ←
6. bass⁶, bass voice¹, basso² - (the lowest adult male singing voice) ←
7. bass⁷ - (the member with the lowest range of a family of musical instruments) ←
8. bass⁸ - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes) ←



Synsets

- Nodes of WordNet are not words or lemmas, but senses
- They 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']
```

Noun Relations in WordNet

Relation	Also Called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	<i>breakfast</i> ¹ → <i>meal</i> ¹
Hyponym	Subordinate	From concepts to subtypes	<i>meal</i> ¹ → <i>lunch</i> ¹
Instance Hypernym	Instance	From instances to their concepts	<i>Austen</i> ¹ → <i>author</i> ¹
Instance Hyponym	Has-Instance	From concepts to their instances	<i>composer</i> ¹ → <i>Bach</i> ¹
Part Meronym	Has-Part	From wholes to parts	<i>table</i> ² → <i>leg</i> ³
Part Holonym	Part-Of	From parts to wholes	<i>course</i> ⁷ → <i>meal</i> ¹
Antonym		Semantic opposition between lemmas	<i>leader</i> ¹ ↔ <i>follower</i> ¹
Derivation		Lemmas w/same morphological root	<i>destruction</i> ¹ ↔ <i>destroy</i> 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
- $\text{pathlen}(c_1, c_2) = 1 + \text{edge length in the shortest path between sense } c_1 \text{ and } c_2$
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

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

Examples

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

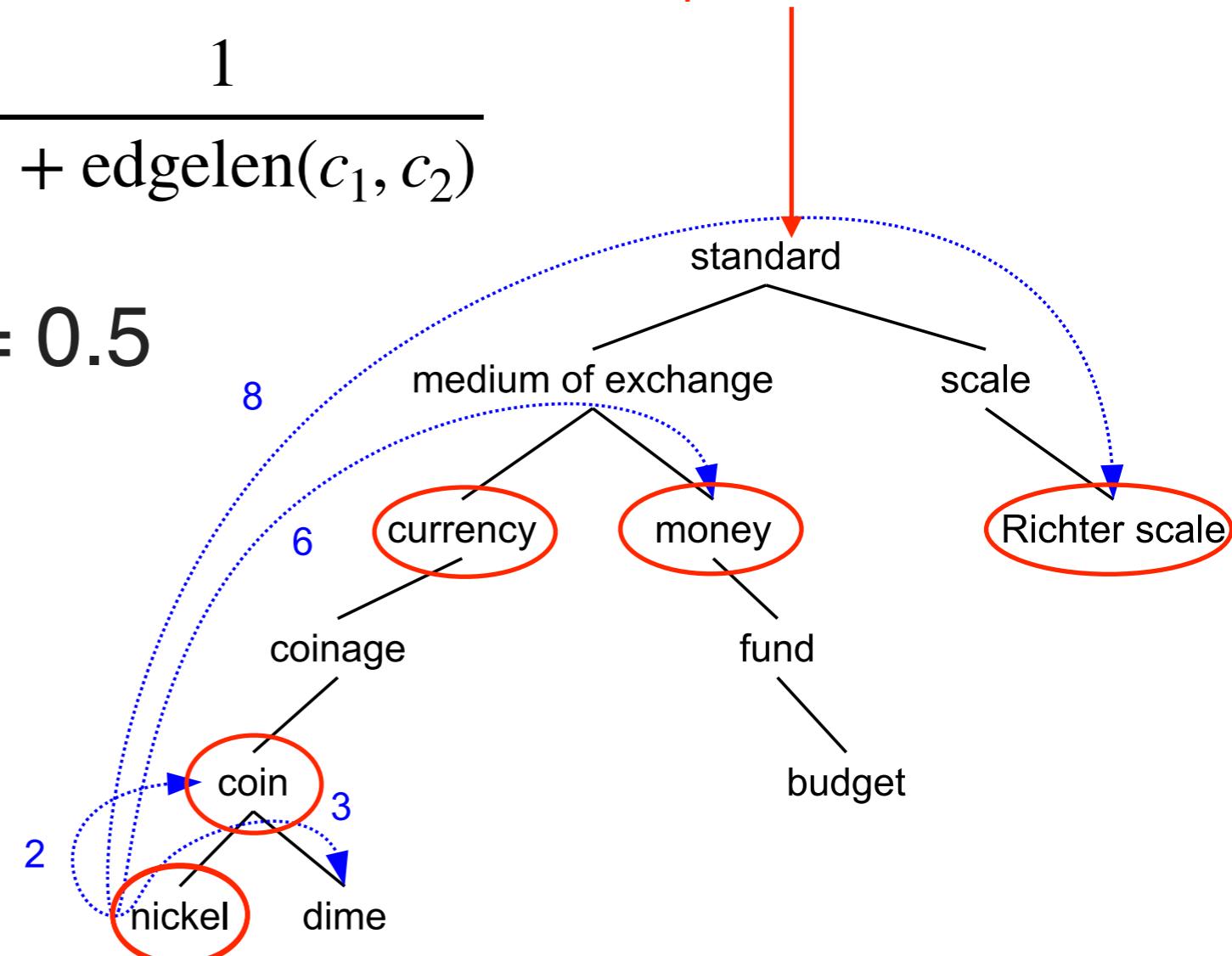
$$\text{simpath}(\text{nickel}, \text{coin}) = 1/2 = 0.5$$

$$\begin{aligned}\text{simpath}(\text{nickel}, \text{currency}) \\ = 1/4 = 0.25\end{aligned}$$

$$\begin{aligned}\text{simpath}(\text{nickel}, \text{money}) \\ = 1/6 = 0.17\end{aligned}$$

$$\begin{aligned}\text{simpath}(\text{nickel}, \text{Richter scale}) \\ = 1/8 = 0.13\end{aligned}$$

Each node is a synset!
For simplicity we use just
the representative word



Beyond Path Length

- $\text{simpath}(\text{nickel}, \text{money}) = 0.17$
- $\text{simpath}(\text{nickel}, \text{Richter scale}) = 0.13$
- 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

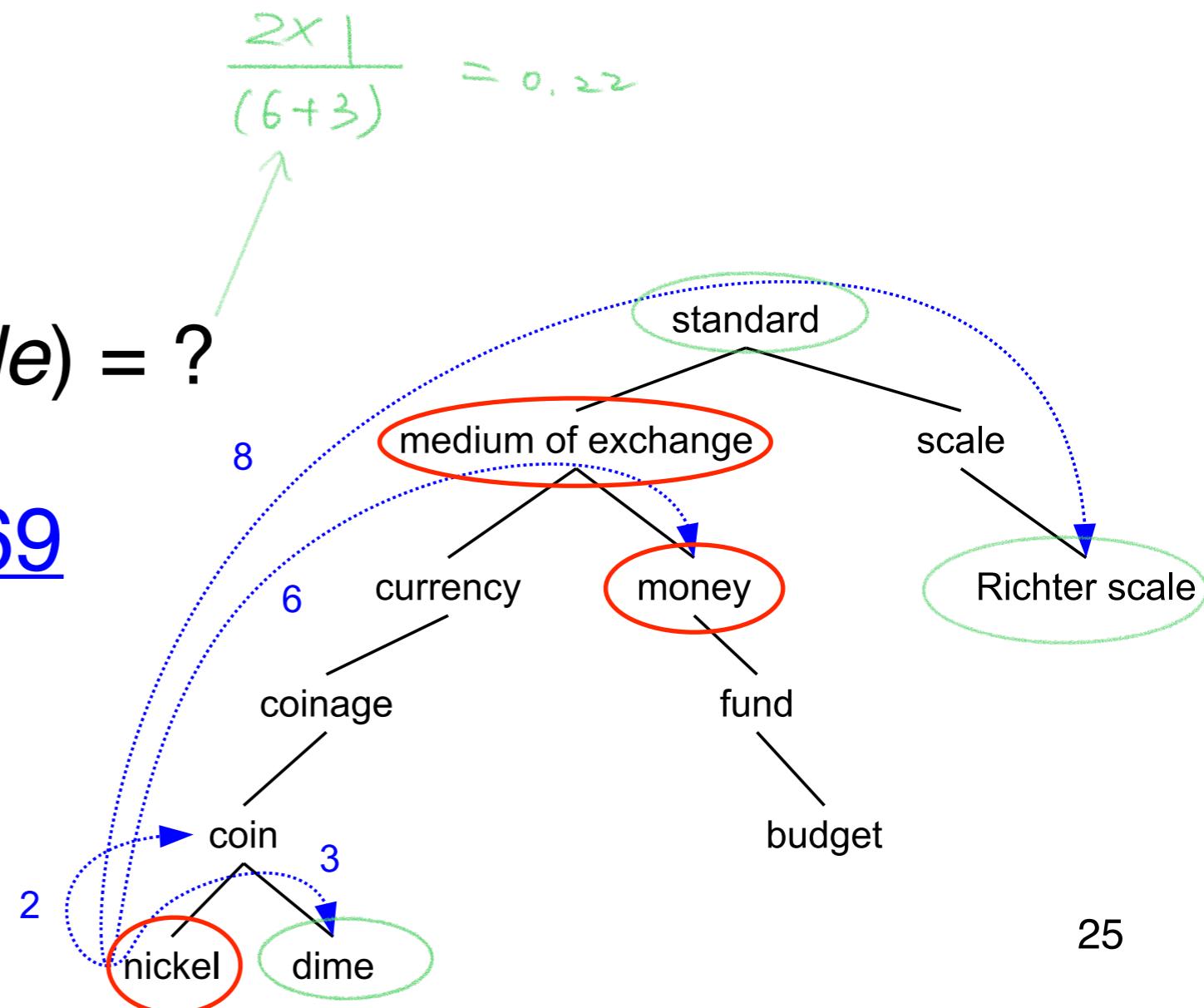
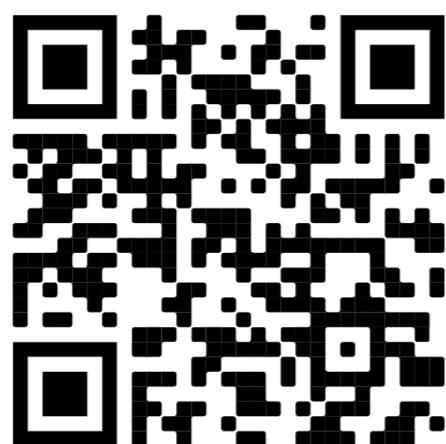
$$\text{simwup}(c_1, c_2) = \frac{2 \times \text{depth}(\text{LCS}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)}$$

$\text{simwup}(\text{nickel}, \text{money}) =$

$$2^*2 / (6+3) = 0.44$$

$\text{simwup}(\text{dime}, \text{Richter scale}) = ?$

PolIEv.com/jeyhanlau569



Abstract Nodes

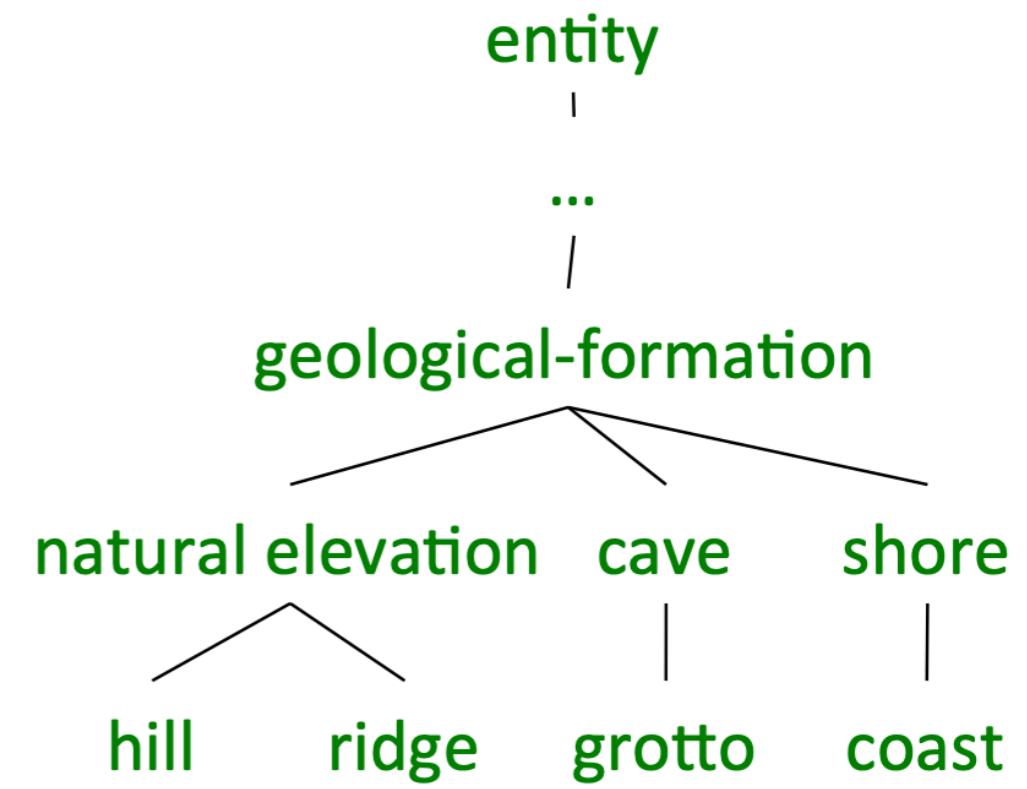
- But node depth is still poor semantic distance metric
 - ▶ $\text{simwup}(\textit{nickel}, \textit{money}) = 0.44$
 - ▶ $\text{simwup}(\textit{nickel}, \textit{Richter scale}) = 0.22$
- Nodes high in the hierarchy is very abstract or general
- How to better capture them?

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!

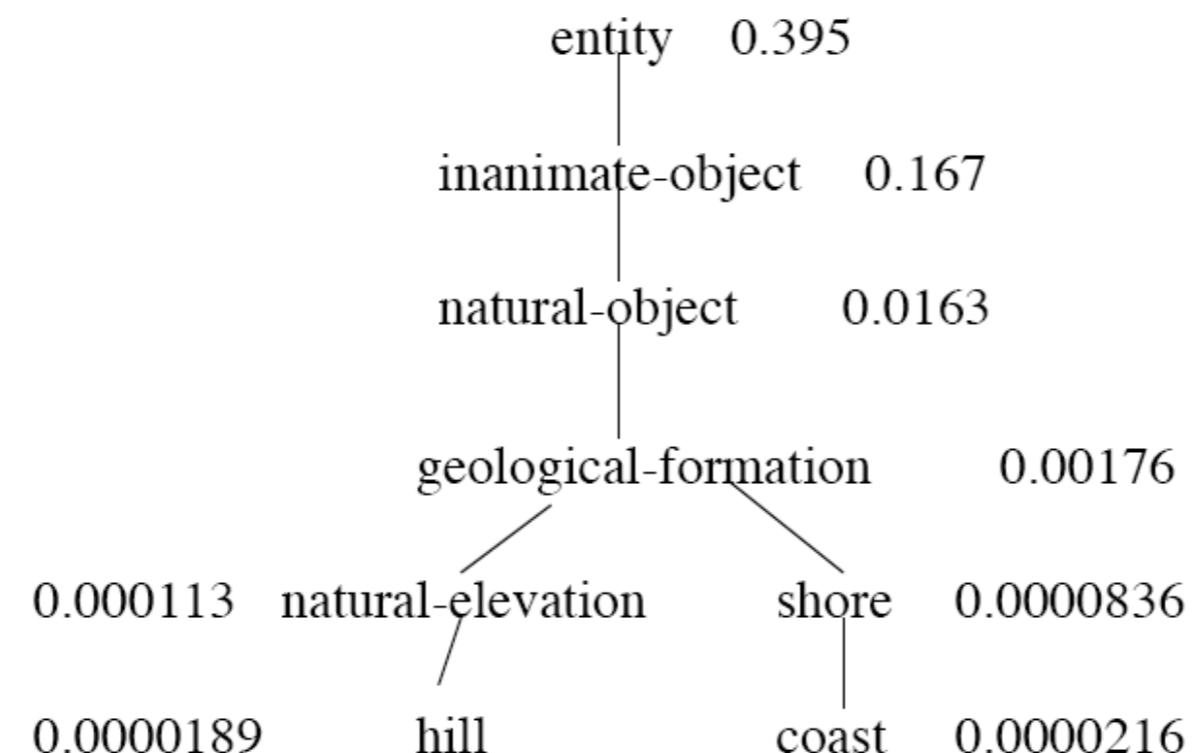
$$P(c) = \frac{\sum_{s \in \text{child}(c)} \text{count}(s)}{N}$$

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



Example

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



Similarity with Information Content

use IC instead of depth (simwup)

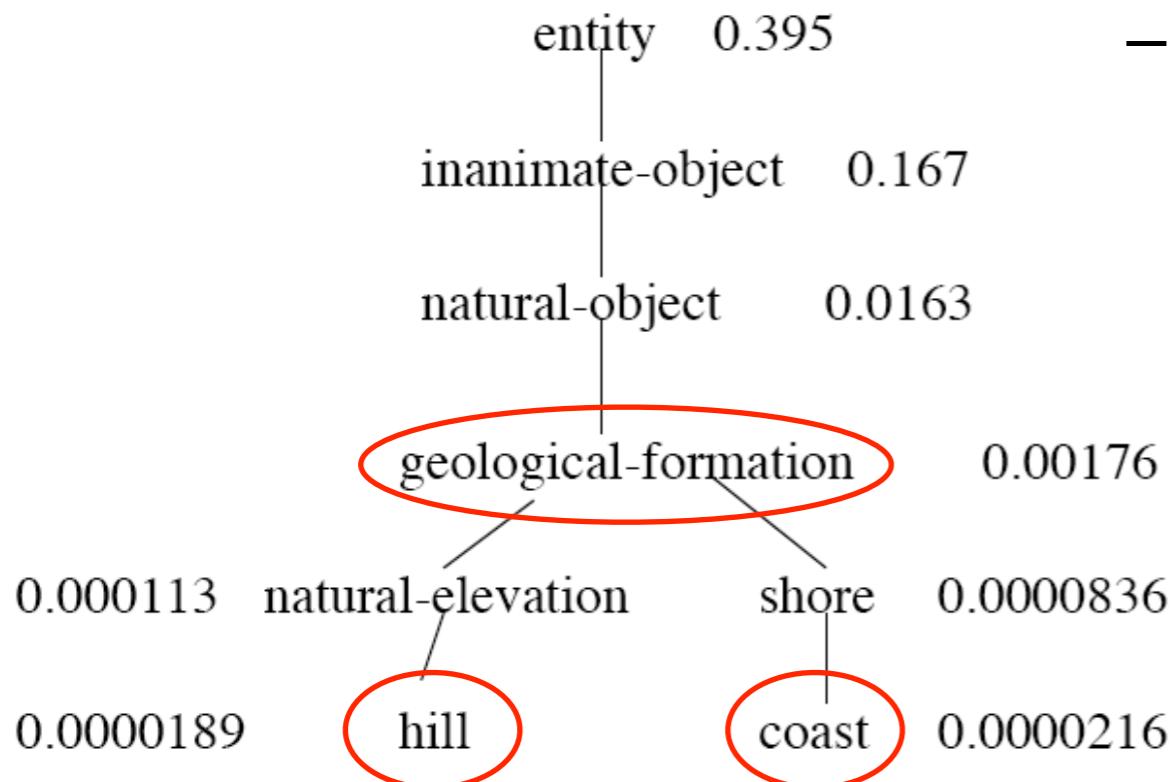
general concept = small values
narrow concept = large values

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

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

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

$$\begin{aligned} \text{simlin}(\text{hill}, \text{coast}) &= \frac{2 \times -\log P(\text{geological-formation})}{-\log P(\text{hill}) - \log P(\text{coast})} \\ &= \frac{-2 \log 0.00176}{-\log 0.0000189 - \log 0.0000216} \end{aligned}$$



If LCS is *entity*, $-2 \log(0.395)!$

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)
If big → trying to capture the theme, rather than the word meaning
- 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.*
- **bank¹**: 2 overlapping non-stopwords, *deposits* and *mortgage*
- **bank²**: 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”			

Unsupervised: Clustering

- Gather usages of the word
 - ... a **bank** is a financial institution that ...
 - ... reserve **bank**, or monetary authority is an institution...
 - ... bed of a river, or stream. The **bank** consists of the sides ...
 - ... right **bank** to the right. The river channel ...
- Perform clustering on context words to learn the different senses
 - ▶ Rationale: context words of the same sense should be similar

Unsupervised: Clustering

- Disadvantages:
 - Sense cluster not very interpretable
 - Need to align with dictionary senses

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 18-18.4.1