

# 自然语言处理

## week-7

凌震华

2024年4月25日



# □ Lexical Semantics (词汇语义学)

- Word Sense
- WordNet
- Word Similarity
- Beyond WordNet



# Lexical Semantics

- Meanings of words
  - What does a word represent in the 'real world'?
- Words (词), **lemmas (词目)**, stems (词干)
  - Surface form, dictionary form, root
  - Celebrations, celebration, celebrate



# Word sense (词义)

- Aspect of meaning of a word (lemma)
- Bank: financial institution; sloping mound
- Just say that there are 2 senses.
- word+sense = "word"
  - bank1, bank2



# Word senses in sentences

The patron orders a meal

- 2 senses of patron, 4 of orders, 2 of meal  
= 16 senses of sentence
- How do we resolve the ambiguity?



# Relations among words/senses

- Homonyms (同形/同音异义词)
  - same surface form, diff semantics
- Polysemy (多义关系)
  - semantic relation (blood bank, money bank)
  - Metonymy (借喻): systematic relationship between diff aspects of same entity
    - “I love Jane Austen” : Author→works of author



# Homonymy

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

- **bank<sub>1</sub>**: financial institution, **bank<sub>2</sub>**: sloping land
- **bat<sub>1</sub>**: club for hitting a ball, **bat<sub>2</sub>**: nocturnal flying mammal

1. Homographs 同形异义 (bank/bank, bat/bat)

2. Homophones 同音异义

1. **write** and **right**
2. **piece** and **peace**



# Homonymy causes problems for NLP applications

- Information retrieval
  - “bat care”
- Machine Translation
  - bat: **murciéago** (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

- 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” (轭式搭配法) 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”**



# More relations

- Homophones: sound the same (wood/would)
- Homographs: bass(fish) and bass(sound)
- Synonym 同义词, antonym 反义词: obvious
- Hyponym 下位词: sub-class (dog, animal)
- Hypernym (superordinate) 上位词: super-class (dog, hound)
- Meronym 部分词: part/whole (wheel, car)
- Holonym 整体词: whole/part (car wheel)



# 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<sub>2</sub>O
- A more formal definition: Two words are synonyms
  - if they can be substituted for each other in all situations



# 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<sub>2</sub>O
  - Big/large



# 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** (反向词)
    - 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*

<b>Superordinate/hyper</b>	vehicle	fruit	furniture
<b>Subordinate/hyponym</b>	car	mango	chair



# Hyponymy more formally

- Extensional 外延:
  - The class denoted by the superordinate extensionally includes the class denoted by the hyponym
- Entailment 蕴含:
  - 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** (层级)
  - A **IS-A** B (or A **ISA** B)
  - B **subsumes** A



# Hyponyms 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...



# Meronymy

- The part-whole relation
  - A *leg* is part of a *chair*, a *wheel* is part of a *car*.
- *Wheel* is a **meronym** of *car*, and *car* is a **holonym** of *wheel*.



## □ Lexical Semantics (词汇语义学)

- Word Sense
- WordNet
- Word Similarity
- Beyond WordNet



# WordNet

- WordNet is a database of facts about words
  - Meanings and the relations among them
- Different from traditional lexicon (e-dictionary)
  - Meaning-based traversal 遍历 (similar to thesaurus)
- <http://wordnet.princeton.edu>
  - Currently over 100,000 nouns, 11,000 verbs, 20,000 adjectives, and 4,000 adverbs
  - Arranged in separate files (DBs)



# 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”



# Knowledge about concepts

- Minimal knowledge about concepts:
  - ◆ X is a kind of Y
  - ◆ X has part Y
  - ◆ X is Y/has property Y
- Can most (all?) concepts referring to entities/events/properties be linked by means of a few relations?





# Structure of WordNet

- Lexical entries linked by semantic relations
  - (Very) large semantic graph
- Database interpretation based on synsets (to be seen in later slides)



# WordNet Noun Relations

Relation	Also Called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	<i>breakfast</i> <sup>1</sup> → <i>meal</i> <sup>1</sup>
Hyponym	Subordinate	From concepts to subtypes	<i>meal</i> <sup>1</sup> → <i>lunch</i> <sup>1</sup>
Instance Hypernym	Instance	From instances to their concepts	<i>Austen</i> <sup>1</sup> → <i>author</i> <sup>1</sup>
Instance Hyponym	Has-Instance	From concepts to concept instances	<i>composer</i> <sup>1</sup> → <i>Bach</i> <sup>1</sup>
Member Meronym	Has-Member	From groups to their members	<i>faculty</i> <sup>2</sup> → <i>professor</i> <sup>1</sup>
Member Holonym	Member-Of	From members to their groups	<i>copilot</i> <sup>1</sup> → <i>crew</i> <sup>1</sup>
Part Meronym	Has-Part	From wholes to parts	<i>table</i> <sup>2</sup> → <i>leg</i> <sup>3</sup>
Part Holonym	Part-Of	From parts to wholes	<i>course</i> <sup>7</sup> → <i>meal</i> <sup>1</sup>
Substance Meronym		From substances to their subparts	<i>water</i> <sup>1</sup> → <i>oxygen</i> <sup>1</sup>
Substance Holonym		From parts of substances to wholes	<i>gin</i> <sup>1</sup> → <i>martini</i> <sup>1</sup>
Antonym		Semantic opposition between lemmas	<i>leader</i> <sup>1</sup> ⇔ <i>follower</i> <sup>1</sup>
Derivationally Related Form		Lemmas w/same morphological root	<i>destruction</i> <sup>1</sup> ⇔ <i>destroy</i> <sup>1</sup>



# WordNet Verb Relations

Relation	Definition	Example
Hypernym	From events to superordinate events	<i>fly</i> <sup>9</sup> $\rightarrow$ <i>travel</i> <sup>5</sup>
Troponym	From events to subordinate event (often via specific manner)	<i>walk</i> <sup>1</sup> $\rightarrow$ <i>stroll</i> <sup>1</sup>
Entails	From verbs (events) to the verbs (events) they entail	<i>snore</i> <sup>1</sup> $\rightarrow$ <i>sleep</i> <sup>1</sup>
Antonym	Semantic opposition between lemmas	<i>increase</i> <sup>1</sup> $\iff$ <i>decrease</i> <sup>1</sup>
Derivationally Related Form	Lemmas with same morphological root	<i>destroy</i> <sup>1</sup> $\iff$ <i>destruction</i> <sup>1</sup>



# Synsets

- Most important lexical relation in WordNet
- Key WordNet concept/tool
- Senses grouped with near-synonyms
- In WordNet, relations are between synsets
- Concepts instantiated (例示) by synsets



# Synset and “sense” ?

- The **synset (synonym set)**, the set of near-synonyms, instantiates a sense or concept, with a **gloss** (注释)
- 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<sup>1</sup>, fool<sup>2</sup>, gull<sup>1</sup>, mark<sup>9</sup>, patsy<sup>1</sup>, fall guy<sup>1</sup>, sucker<sup>1</sup>, soft touch<sup>1</sup>, mug<sup>2</sup>
- Each of **these** senses have this same gloss
  - (Not **every** sense; sense 2 of gull is the aquatic bird)

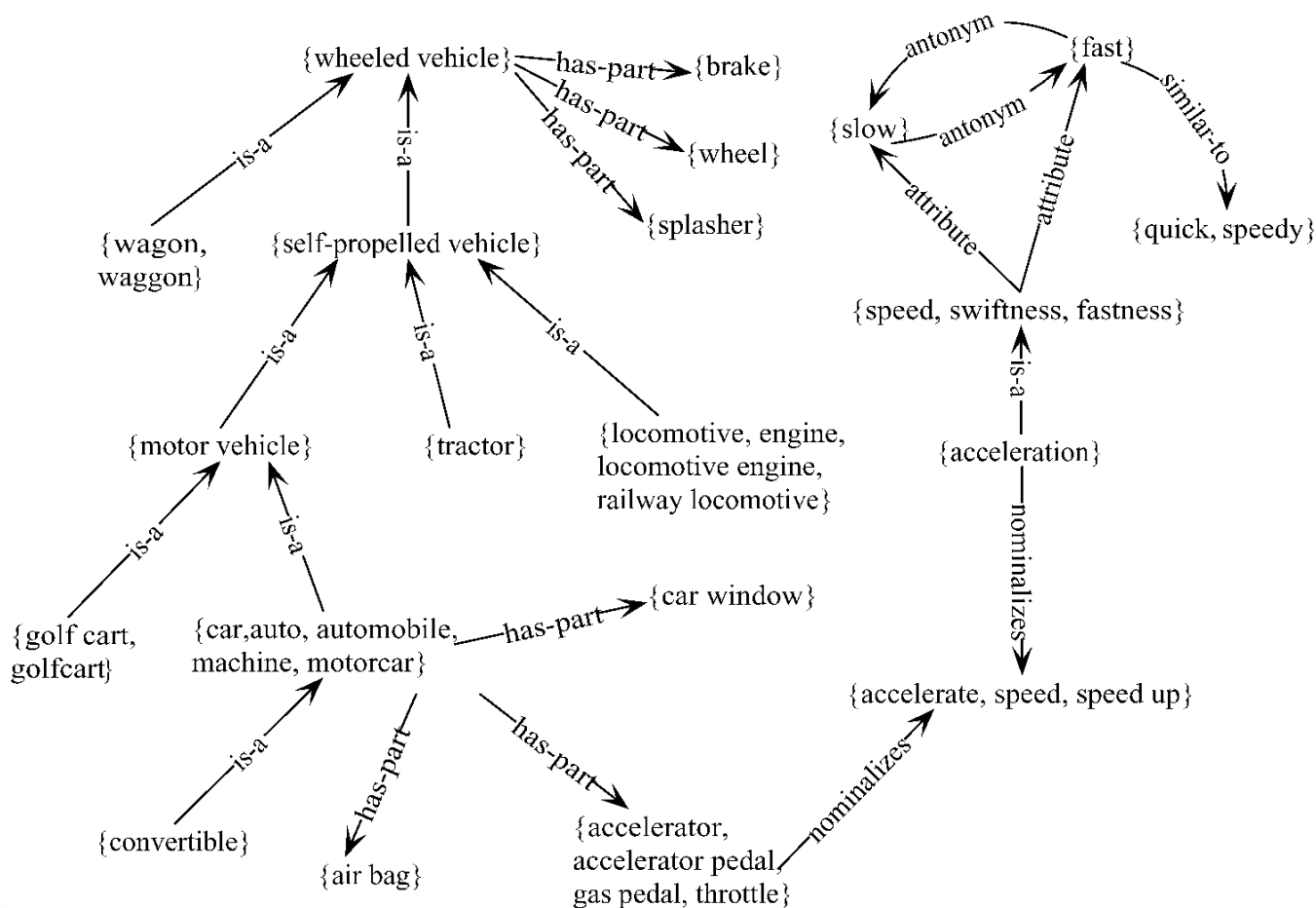


# WordNet Hypernym Hierarchy for "bass"

- **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: Viewed as a graph



# "Supersenses": The top level hypernyms in the hierarchy

(counts from Schneider and Smith 2013's Streusel corpus)

| Noun       |      |                   |              | Verb |                   |             |                      |
|------------|------|-------------------|--------------|------|-------------------|-------------|----------------------|
| GROUP      | 1469 | <i>place</i>      | BODY         | 87   | <i>hair</i>       | STATIVE     | 2922 <i>is</i>       |
| PERSON     | 1202 | <i>people</i>     | STATE        | 56   | <i>pain</i>       | COGNITION   | 1093 <i>know</i>     |
| ARTIFACT   | 971  | <i>car</i>        | NATURAL OBJ. | 54   | <i>flower</i>     | COMMUNIC.*  | 974 <i>recommend</i> |
| COGNITION  | 771  | <i>way</i>        | RELATION     | 35   | <i>portion</i>    | SOCIAL      | 944 <i>use</i>       |
| FOOD       | 766  | <i>food</i>       | SUBSTANCE    | 34   | <i>oil</i>        | MOTION      | 602 <i>go</i>        |
| ACT        | 700  | <i>service</i>    | FEELING      | 34   | <i>discomfort</i> | POSSESSION  | 309 <i>pay</i>       |
| LOCATION   | 638  | <i>area</i>       | PROCESS      | 28   | <i>process</i>    | CHANGE      | 274 <i>fix</i>       |
| TIME       | 530  | <i>day</i>        | MOTIVE       | 25   | <i>reason</i>     | EMOTION     | 249 <i>love</i>      |
| EVENT      | 431  | <i>experience</i> | PHENOMENON   | 23   | <i>result</i>     | PERCEPTION  | 143 <i>see</i>       |
| COMMUNIC.* | 417  | <i>review</i>     | SHAPE        | 6    | <i>square</i>     | CONSUMPTION | 93 <i>have</i>       |
| POSSESSION | 339  | <i>price</i>      | PLANT        | 5    | <i>tree</i>       | BODY        | 82 <i>get...done</i> |
| ATTRIBUTE  | 205  | <i>quality</i>    | OTHER        | 2    | <i>stuff</i>      | CREATION    | 64 <i>cook</i>       |
| QUANTITY   | 102  | <i>amount</i>     |              |      |                   | CONTACT     | 46 <i>put</i>        |
| ANIMAL     | 88   | <i>dog</i>        |              |      |                   | COMPETITION | 11 <i>win</i>        |
|            |      |                   |              |      |                   | WEATHER     | 0 —                  |





# Supersenses

- A word's supersense can be a useful coarse-grained 粗粒度的 representation of word meaning for NLP tasks

I googled<sub>communication</sub> restaurants<sub>GROUP</sub> in the area<sub>LOCATION</sub> and Fuji\_Sushi<sub>GROUP</sub>  
came\_up<sub>communication</sub> and reviews<sub>COMMUNICATION</sub> were<sub>stative</sub> great so I made\_a  
carry\_out<sub>possession</sub> \_order<sub>communication</sub>



# Accessing WordNet

- Where it is:
  - <http://wordnet.princeton.edu>
- Libraries
  - Python: WordNet from NLTK
    - <http://www.nltk.org/>
  - Java:
    - JWNL, extJWNL on sourceforge



# Database view of WordNet

- Huge boolean table with columns being words, rows being synsets
  - Very sparse table
  - Efficiently implemented by recording non-zero entries as binary relations: word  $j$  belongs to synset  $l$
  - Sparse table organization methods applied



# Limitations of WordNet

- Links mostly restricted to within a POS category (related terms must be substitutable)
- Paradigmatic (聚合) relations (within POS) vs/ syntagmatic (组合) relations (across POS)
- 4 largely unconnected WordNets: nouns, verbs, adjectives, adverbs



# Network issues

- Very sparse even within a POS category
  - Work on connecting all synsets
- Links not weighted
- (Most) links not directed
- Evocation 唤起 not implemented
  - Word-associations based on psycholinguistic studies

## **Adding Dense, Weighted Connections to WORDNET**

Jordan Boyd-Graber and Christiane Fellbaum and Daniel Osherson and Robert Schapire  
Princeton University

October 9, 2005



# Types vs Roles

- WordNet does not implement roles separate from types
- Dog → poodle (type relation)
- Dog → pet (role relation)



# WordNet activity

- WordNet conferences since 2002
  - Standardization, applications, improvements to WordNet
- Improve precision of search by expanding queries based on synsets
- Use in language learning
- Translation aid based on semantic distance



## □ Lexical Semantics (词汇语义学)

- Word Sense
- WordNet
- Word Similarity
- Beyond WordNet





# 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<sup>1</sup> is similar to fund<sup>3</sup>
  - Bank<sup>2</sup> is similar to slope<sup>5</sup>
- But we'll compute similarity over both words and senses



# Why word similarity

- A practical component in lots of NLP tasks
  - Question answering
  - Natural language generation
  - Automatic essay grading
  - Plagiarism detection
- A theoretical component in many linguistic and cognitive tasks
  - Historical semantics
  - Models of human word learning
  - Morphology and grammar induction



# 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



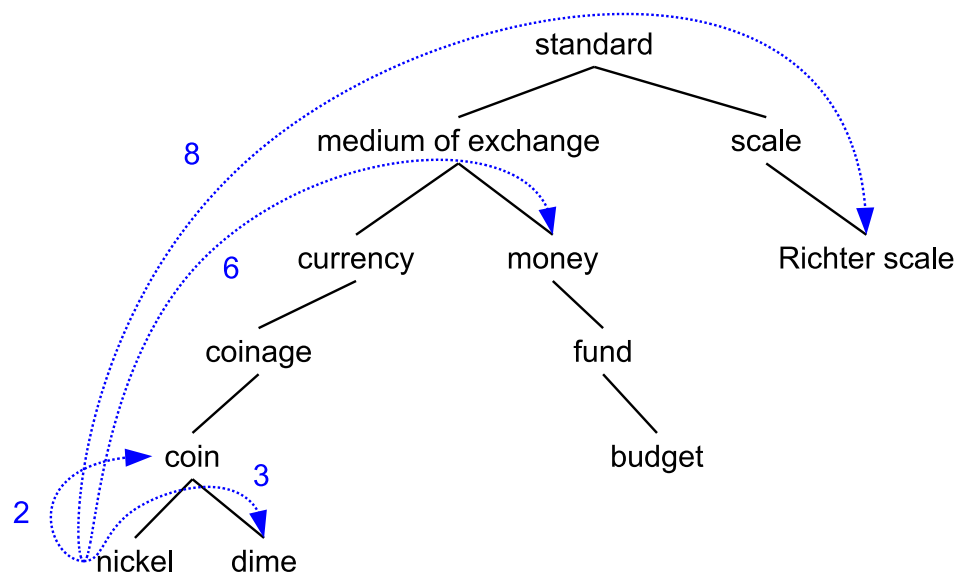
# Two classes of similarity algorithms

- Thesaurus-based algorithms
  - Are words “nearby” in hypernym hierarchy?
  - Do words have similar glosses (definitions)?
- Distributional algorithms
  - Do words have similar distributional contexts?
  - Distributional (Vector) semantics!



# 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



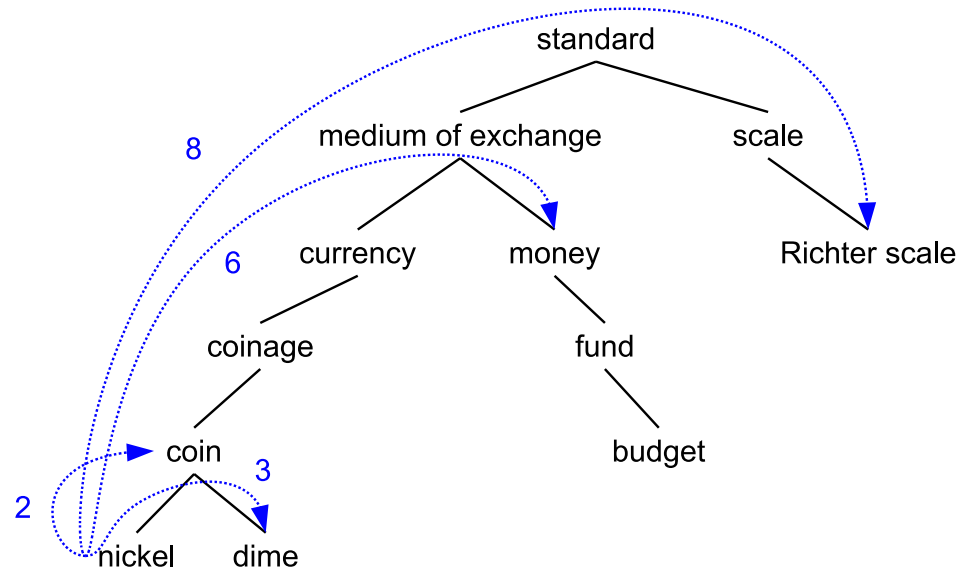
# Refinements to path-based similarity

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



# Example: path-based similarity

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



$$\text{simpath}(\textit{nickel}, \textit{coin}) = 1/2 = .5$$

$$\text{simpath}(\textit{fund}, \textit{budget}) = 1/2 = .5$$

$$\text{simpath}(\textit{nickel}, \textit{currency}) = 1/4 = .25$$

$$\text{simpath}(\textit{nickel}, \textit{money}) = 1/6 = .17$$

$$\text{simpath}(\textit{coinage}, \textit{Richter scale}) = 1/6 = .17$$



# 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





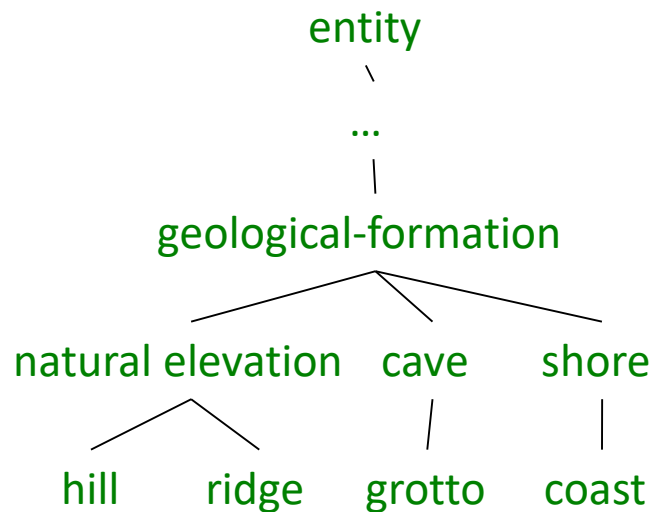
# Information content similarity metrics

Resnik 1995

- 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  $1-P(c)$
  - 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



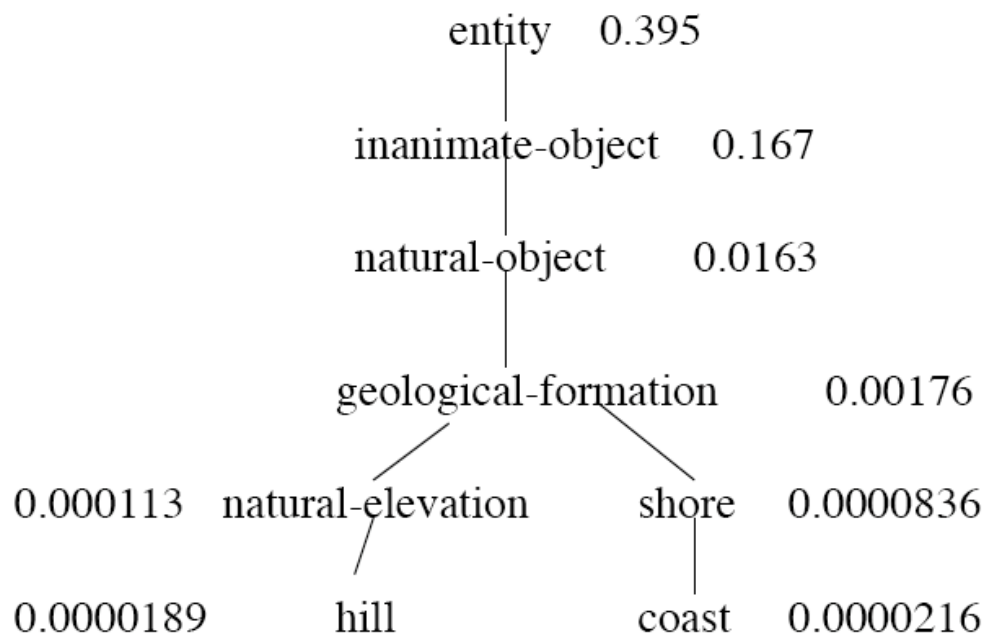
- Train by counting in a corpus
  - Each instance of `hill` counts toward frequency of *natural elevation*, *geological formation*, *entity*, etc
  - Let `words(c)` be the set of all words that are children of node `c`
    - `words( "geo-formation" ) = {hill,ridge,grotto,coast,cave,shore,natural elevation}`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

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

# Information content similarity

- WordNet hierarchy augmented with probabilities  $P(c)$

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



# Information content and probability

- The **self-information** of an event, also called its **surprisal** 惊异度
  - how surprised we are to know it; how much we learn by knowing it.
  - The more surprising something is, the more it tells us when it happens
  - We'll measure self-information in **bits**.

$$I(w) = -\log_2 P(w)$$

- I flip a coin;  $P(\text{heads}) = 0.5$
- How many bits of information do I learn by flipping it?
  - $I(\text{heads}) = -\log_2(0.5) = -\log_2(1/2) = \log_2(2) = 1 \text{ bit}$
- I flip a biased coin:  $P(\text{heads}) = 0.8$  I don't learn as much
  - $I(\text{heads}) = -\log_2(0.8) = -\log_2(0.8) = .32 \text{ bits}$



# Information content: definitions

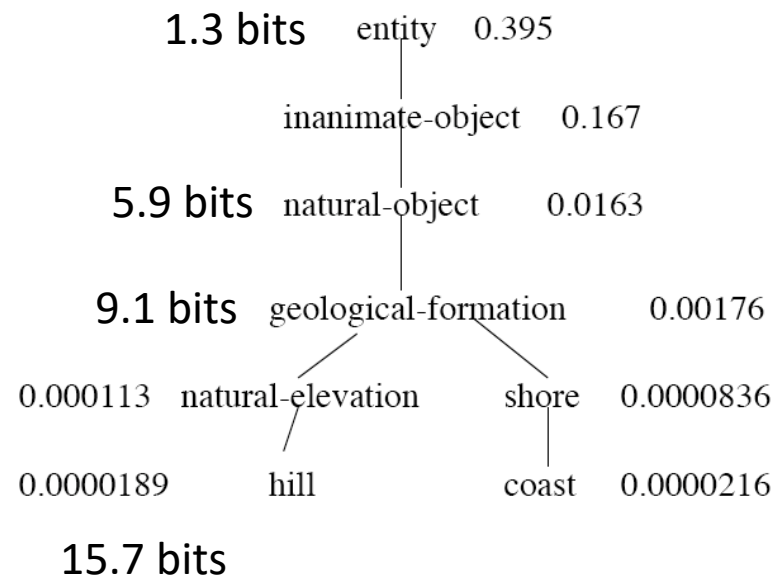
- Information content:

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

- Most informative subsumer  
(Lowest common subsume  
最小公共包含)

$$LCS(c_1, c_2) =$$

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
  - $\text{sim}_{\text{resnik}}(c_1, c_2) = -\log P(\text{LCS}(c_1, c_2))$



# 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:  $IC(\text{common}(A,B))$
- Difference:  $IC(\text{description}(A,B)) - IC(\text{common}(A,B))$



# 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

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

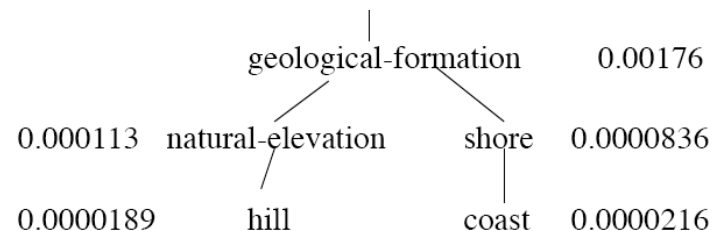
- Lin (altering Resnik) defines  $IC(common(A, B))$  as 2 x information of the LCS

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





# Lin similarity function



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

$$\begin{aligned} sim_{Lin}(\text{hill}, \text{coast}) &= \frac{2 \log P(\text{geological-formation})}{\log P(\text{hill}) + \log P(\text{coast})} \\ &= \frac{2 \ln 0.00176}{\ln 0.0000189 + \ln 0.0000216} \\ &= .59 \end{aligned}$$



# 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

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

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

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

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



## □ Lexical Semantics (词汇语义学)

- Word Sense
- WordNet
- Word Similarity
- Beyond WordNet



# Beyond WordNet: 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” 标题

- **Hemoglobins**

*Synset*

**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 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, and](#)
  - [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 \[D26\]](#) +
  - [Chemical Actions and Uses \[D27\]](#) +
5. + Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. + Psychiatry and Psychology [F]
7. + Phenomena and Processes [G]

[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 \[D12.776.124.400.141\]](#)

[Erythrocyruorins \[D12.776.124.400.220\]](#)

# 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

