Intelligent Systems (AI-2)

Computer Science cpsc422, Lecture 24

March, 15, 2021

Slide credit: Satanjeev Banerjee Ted Pedersen 2003, Jurfsky & Martin 2008-2016

Lecture Overview

- **Semantic Similarity/Distance**
- **Concepts: Thesaurus/Ontology Methods**
- **Words: Distributional Methods**

Why words/concepts similarity is important ?

"**fast**" is similar to "**rapid**"

"**tall**" is similar to "**height**"

Question answering:

*Q: "How **tall** is Mt. Everest?"*
*Candidate A: "The official **height** of Mount Everest is 29029 feet"*

- Extends to sentence/paragraph similarity
- **Summarization**: identify and eliminate redundancy, aggregate similar phrase/sentences
- ....

WordNet: entry for **“table”**

gloss

The **noun** "table" has 6 senses in WordNet.

1. table, tabular array (a set of data arranged in rows and columns) *"see table 1"*
2. table (a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs) *"it was a sturdy table"*
3. table (a piece of furniture with tableware for a meal laid out on it) *"I reserved a table at my favorite restaurant"*
4. mesa, table (flat tableland with steep edges) *"the tribe was relatively safe on the mesa but they had to descend into the valley for water"*
5. table (a company of people assembled at a table for a meal or game) *"he entertained the whole table with his witty remarks"*
6. board, table (food or meals in general) *"she sets a fine table"; "room and board"*

The **verb** "table" has 1 sense in WordNet.

1. postpone, prorogue, hold over, put over, **table**, shelve, set back, defer, remit, put off – (hold back to a later time; "let's postpone the exam")

WordNet Relations (between synsets!)

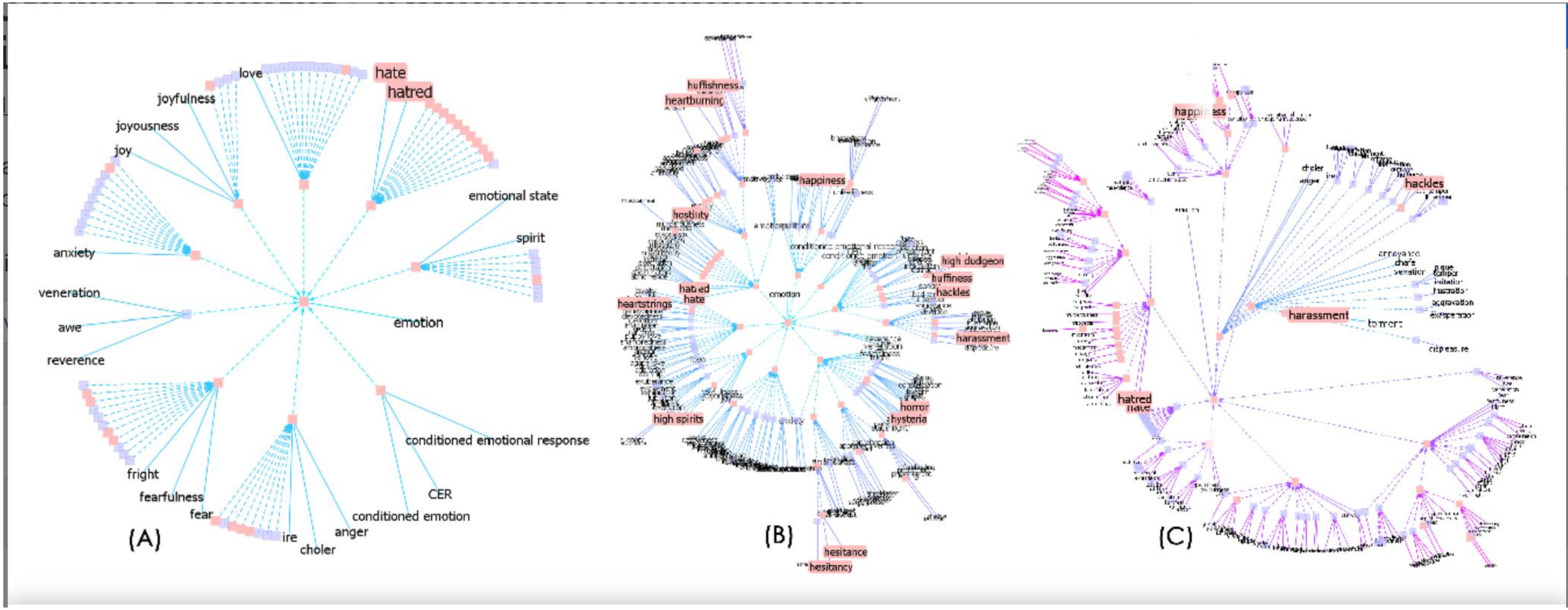
Nouns

Relation	Definition	Example
Hypernym	From concepts to superordinates	*breakfast* → *meal*
Hyponym	From concepts to subtypes	*meal* → *lunch*
Has-Member	From groups to their members	*faculty* → *professor*
Member-Of	From members to their groups	*copilot* → *crew*
Has-Part	From wholes to parts	*table* → *leg*
Part-Of	From parts to wholes	*course* → *meal*
Antonym	Opposites	*leader* → *follower*

Verbs

Relation	Definition	Example
Hypernym	From events to superordinate events	*fly* → *travel*
Troponym	From events to their subtypes	*walk* → *stroll*
Entails	From events to the events they entail	*snore* → *sleep*
Antonym	Opposites	*increase* ⟺ *decrease*

Visualizing Wordnet Relations



C. Collins, "WordNet Explorer: Applying visualization principles to lexical semantics," University of Toronto, Technical Report kmdi 2007-2, 2007.

Semantic Similarity/Distance: example

(n) table -- (a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs)

(n) mesa, table -- (flat tableland with steep edges)

sim

(n) hill (a local and well-defined elevation of the land)

(n) lamp (a piece of furniture holding one or more electric light bulbs)

dissimilar

Semantic Similarity/Distance

Between two concepts in an ontology, e.g., between two senses in Wordnet

What would you use to compute it ?

A. The distance between the two concepts in the underlying hierarchies / graphs

B. The glosses of the concepts

C. None of the above

D. Both of the above

Gloss Overlaps ≈ Relatedness

concepts

- Lesk's (1986) idea: Related word senses are (often) defined *using the same words.* E.g:
  - bank(1): "a financial institution"
  - bank(2): "sloping land beside a body of water"
  - lake: "a body of water surrounded by land"

Gloss Overlaps ≈ Relatedness

- Lesk's (1986) idea: Related word senses are (often) defined *using the same words*. E.g:
  - bank(1): "a financial institution"
  - bank(2): "sloping land beside a body of water"
  - lake: "a body of water surrounded by land"

Gloss Overlaps ≈ Relatedness

- Lesk's (1986) idea: Related word senses are (often) defined *using the same words*. E.g:
  - bank(1): "a financial institution"
  - bank(2): "sloping land beside a body of water"
  - lake: "a body of water surrounded by land"
- Gloss overlaps = # content words common to two glosses ≈ relatedness
  - Thus, relatedness (bank(2), lake) = 3
  - And, relatedness (bank(1), lake) = 0

Limitations of (Lesk's) Gloss Overlaps

- Most glosses are very short.
  - So not enough words to find overlaps with.

- Solution?

Extended gloss overlaps

- Add glosses of synsets connected to the input synsets.

Extending a Gloss

sentence: “the penalty meted out to one adjudged guilty”

bench: “persons who hear cases in a court of law”

overlapped words = 0

Extending a Gloss

final judgment: "a judgment disposing of the case before the court of law"

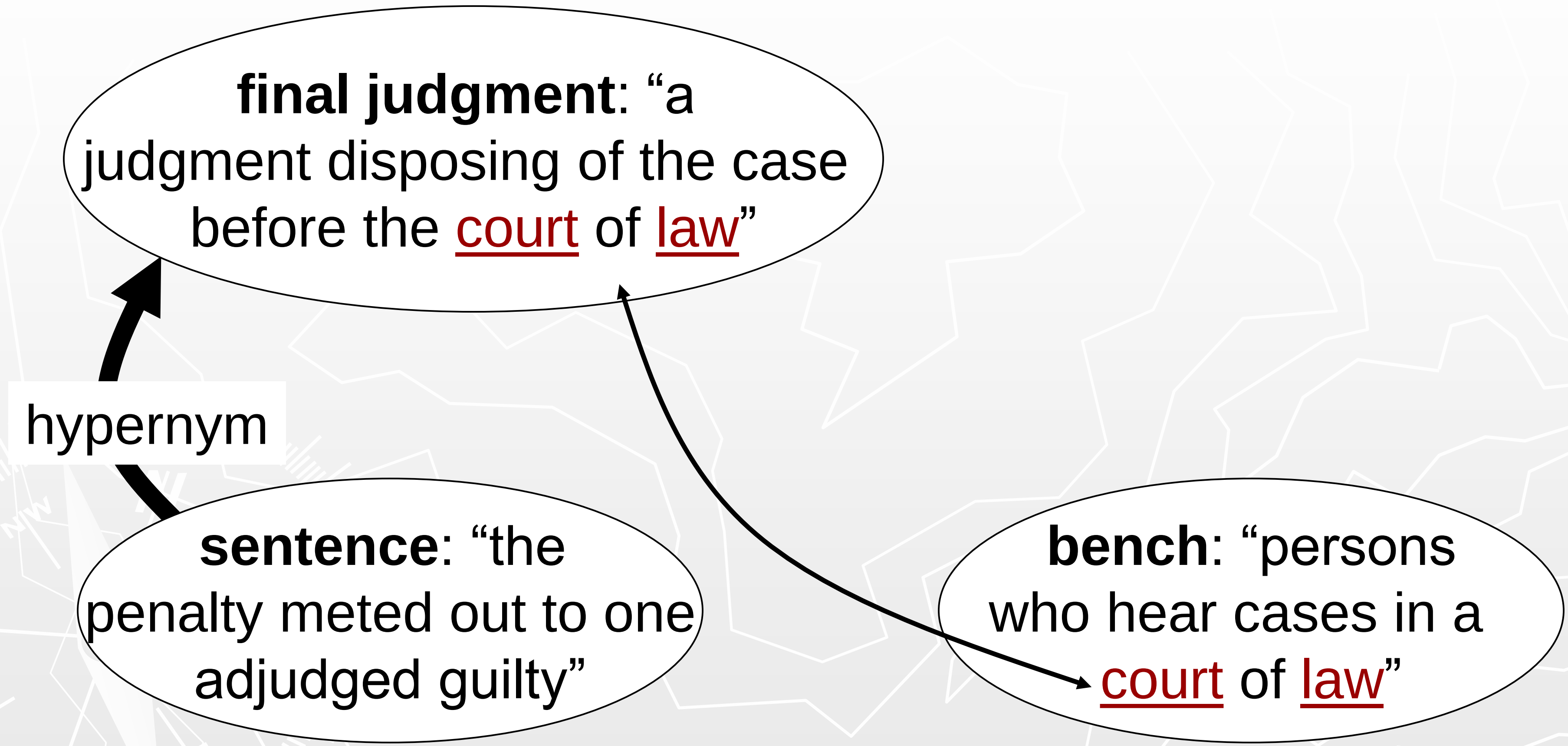
hypernym

sentence: "the penalty meted out to one adjudged guilty"

bench: "persons who hear cases in a court of law"

overlapped words = 0

Extending a Gloss



overlapped words = 2

Creating the Extended Gloss Overlap Measure

- How to measure overlaps?
- Which relations to use for gloss extension?

How to Score Overlaps?

- Lesk simply summed up overlapped words.
- But matches involving phrases – phrasal matches – are rarer, and more informative
  - E.g. "court of law" "body of water"
- Aim: Score of n words in a phrase > sum of scores of n words in shorter phrases
- Solution: Give a phrase of n words a score of n^2
  - "court of law" gets score of 9.

overlap of
9+1=10

- bank(2): "sloping <u>land</u> beside a <u>body of water</u>"
- lake: "a body of water surrounded by <u>land</u>"

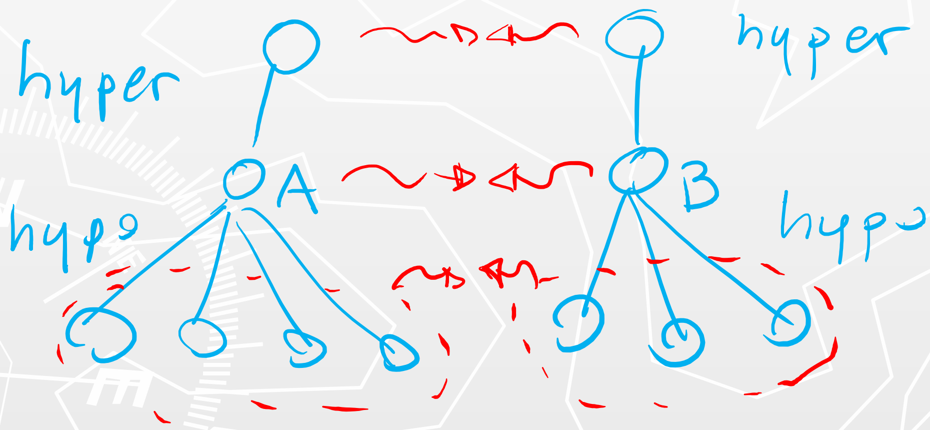
Which Relations to Use?

Typically include…

- Hypernyms [“car” → “vehicle”]
- Hyponyms [“car” → “convertible”]
- Meronyms [“car” → “accelerator”]
- …

Extended Gloss Overlap Measure

- Input two synsets A and B
- Find phrasal gloss overlaps between A and B
- For *each relation*, compute phrasal gloss overlaps between every synset connected to A, and every synset connected to B



compute phrasal
score overlap

- Add phrasal scores to get relatedness of A and B

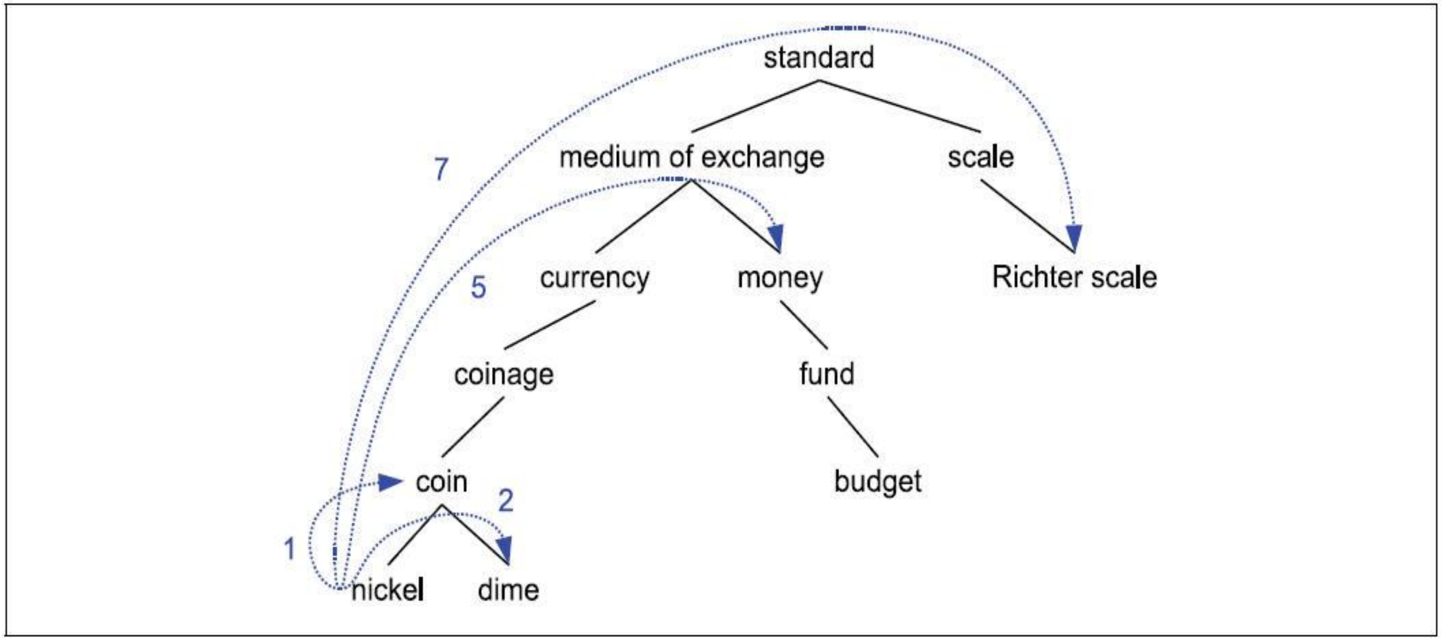
A and B can be from different parts of speech!

Distance: Path-length

Path-length sim based on is-a/hypernyms hierarchies

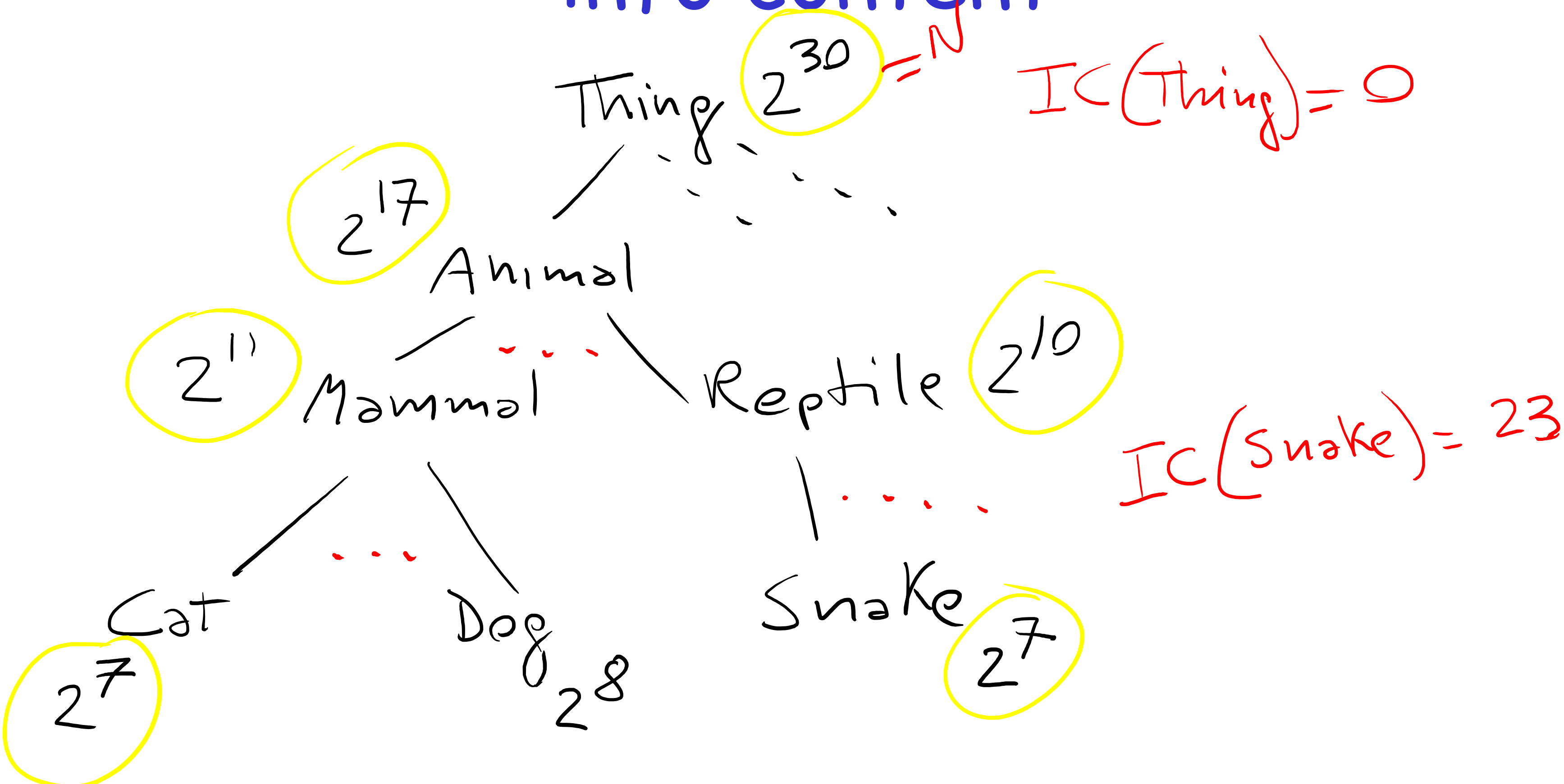
$$\boxed{\text{sim}_{\text{path}}(c_1, c_2) = 1/\, pathlen(c_1, c_2)}$$

c_1, c_2 are senses



But this is assuming that all the links are the same….
Encode the same semantic distance….

Probability of a concept/sense and its info content



probability

$$P(c) = \frac{count(c)}{N}$$

count of all Things

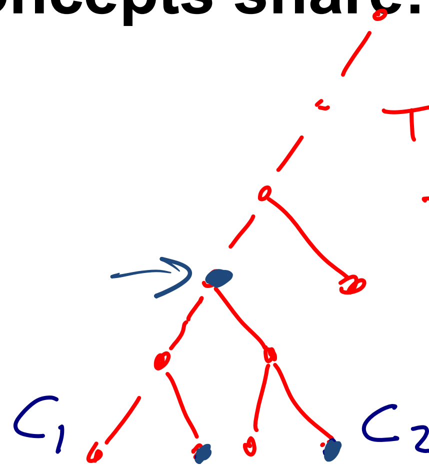
$$IC(c) = -\log P(c)$$

Information Content

Concept Distance: info content

- **Similarity should be proportional to the information that the two concepts share… what is that?**

P(root) = 1



The lower the concept/sense the lower its prob.

probability

$$\boxed{\mathrm{P}(c) = \frac{count(c)}{N}}$$

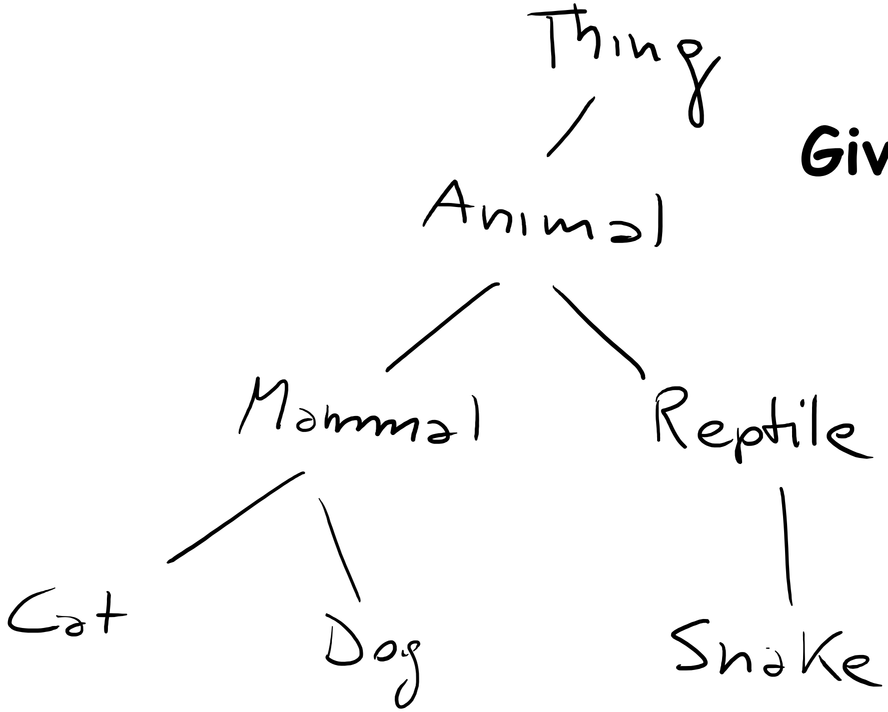
$$\boxed{\mathrm{IC}(c) = -\log P(c)}$$

Information Content

$$\boxed{LCS(c_1, c_2)}$$

Lowest Common Subsumer

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



Given this measure of similarity

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

Are these two the same?

$\mathrm{sim}_{\mathrm{resnik}}(Dog, Snake)$ $\mathrm{sim}_{\mathrm{resnik}}(Mammal, \mathrm{Reptile})$

A. Yes B. No C. Cannot tell

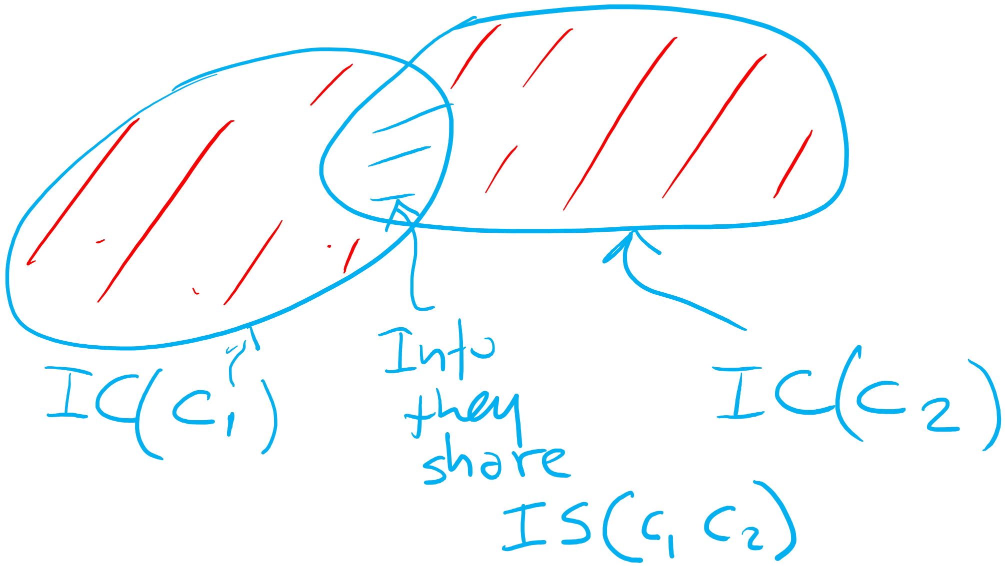
Is this reasonable?

A. Yes B. No C. Cannot tell

well ... we can consider better alternatives ...

Concept Distance: info content

- One of best performers – Jiang-Conrath distance
- How much information the two DO NOT share ?



$$\left(IC(c_1) - IS(c_1,c_2)\right) + \left(IC(c_2) - IS(c_1,c_2)\right)$$

$$IC(c_1) + IC(c_2) - 2*IS(c_1,c_2)$$

Concept Distance: info content

$$\left(IC(c_1) - IS(c_1,c_2)\right) + \left(IC(c_2) - IS(c_1,c_2)\right)$$

$$IC(c_1) + IC(c_2) - 2 * IS(c_1,c_2)$$

$$\boxed{dist_{JC}(c_1,c_2) = ((-\log P(c_1)) + (-\log P(c_2))) - (2 \times -\log P(LCS(c_1,c_2)))}$$

$$\boxed{dist_{JC}(c_1,c_2) = 2 \times \log P(LCS(c_1,c_2)) - (\log P(c_1) + \log P(c_2))}$$

- This is a measure of distance. Reciprocal for similarity! $1/dist_{JC}$
- Problem for measures working on hierarchies/graphs: only compare concepts associated with words of one part-of speech (typically nouns)

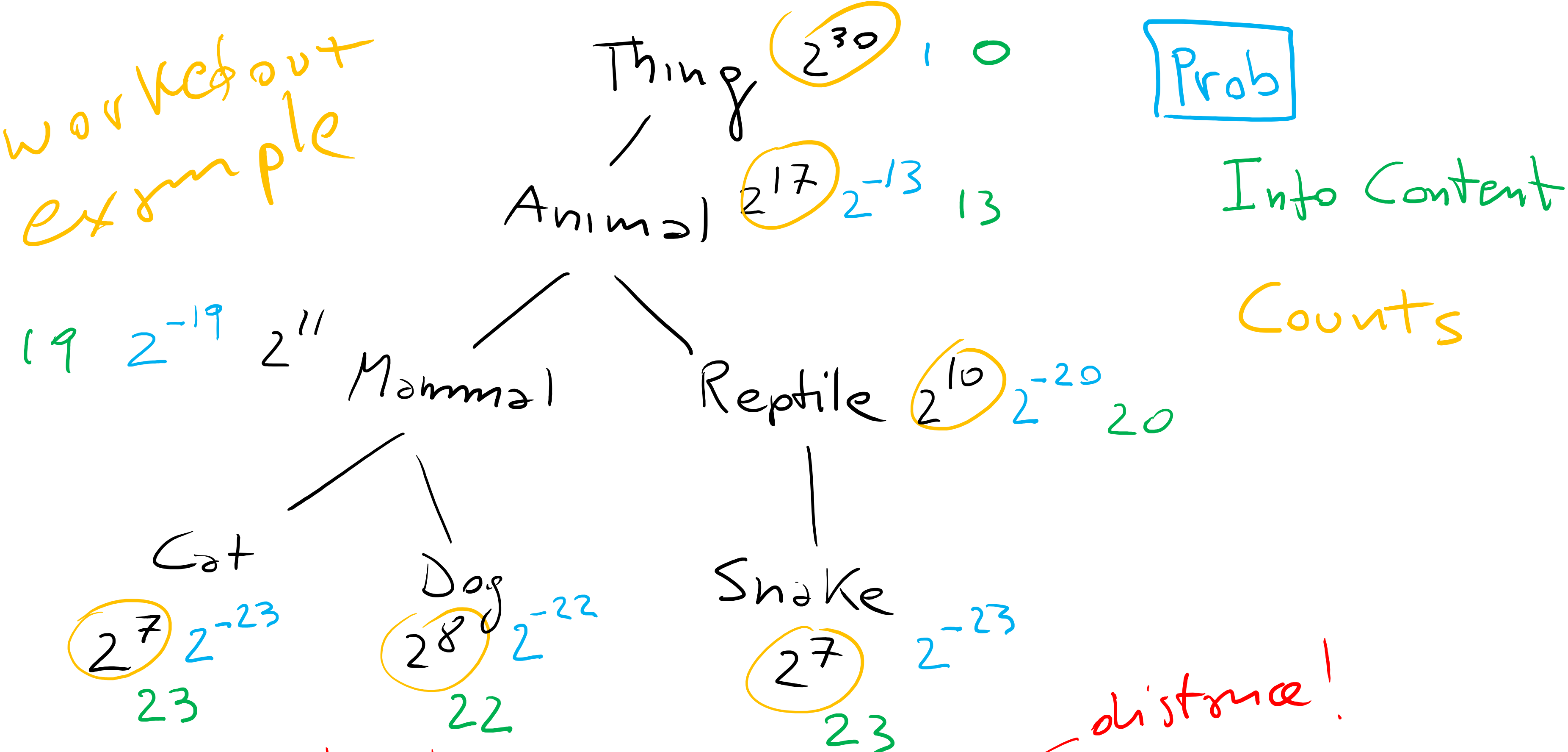
Concept Distance: info content

- One of best performers – Jiang-Conrath distance
- How much information the two DO NOT share

$$\boxed{dist_{JC}(c_1,c_2) = ((-\log P(c_1)) + (-\log P(c_2))) - (2 \times -\log P(LCS(c_1,c_2)))}$$

$$\boxed{dist_{JC}(c_1,c_2) = 2 \times \log P(LCS(c_1,c_2)) - (\log P(c_1) + \log P(c_2))}$$

- This is a measure of distance. Reciprocal for similarity! $1/dist_{JC}$
- Problem for measures working on hierarchies/graphs: only compare concepts associated with words of one part-of speech (typically nouns)



similarity!

$$sim_{res}(Dog, Snake) = 13$$

$$sim_{res}(Mammal, Reptile) = 13$$

$$dist_{JC}(Dog, Snake) = (2\times -13) + (22+23) = 19$$

$$dist_{JC}(Mammal, Reptile) = (2\times -13) + (19+20) = 13$$

Lecture Overview

- Semantic Similarity/Distance
- Concepts: Thesaurus/Ontology Methods
- **Words: Distributional Methods – Word Similarity (WS)**

Word Similarity: Distributional Methods

- Do not have any thesauri/ontologies for target language (e.g., Russian)
- If you have thesaurus/ontology, still
  - Missing domain-specific (e.g., technical words)
  - Poor hyponym knowledge (for V) and nothing for Adj and Adv
  - Difficult to compare senses from different hierarchies (although extended Lesk can do this)
- Solution: extract similarity from corpora

- Basic idea: two words are similar if they appear in similar contexts

Intuition of distributional word similarity

- Example: Suppose I asked you what is ***tesgüino***?

```
A bottle of tesgüino is on the table
Everybody likes tesgüino
Tesgüino makes you drunk
We make tesgüino out of corn.
```

- From context words humans can guess ***tesgüino*** means
  - an alcoholic beverage like beer
- Intuition for algorithm:
  - **Two words are similar if they have similar word contexts.**

WS Distributional Methods (1)

Word-Word matrix: Sample contexts $\pm$ 7 words

sugar, a sliced lemon, a tablespoonful of	**apricot**	preserve or jam, a pinch each of,
their enjoyment. Cautiously she sampled her first	**pineapple**	and another fruit whose taste she likened
well suited to programming on the digital	**computer**.	In finding the optimal R-stage policy from
for the purpose of gathering data and	**information**	necessary for the study authorized in the
...	...	...

- Portion of matrix from the Brown corpus

	aardvark	computer	data	pinch	result	sugar	…
apricot	0	0	0	1	0	1	
pineapple	0	0	0	1	0	1	
digital	0	2	1	0	1	0	
information	0	1	6	0	4	0	
…	…						

Simple example of Vectors Models aka "embeddings".

- Model the meaning of a word by "embedding" in a vector space.
- The meaning of a word is a vector of numbers

WS Distributional Methods (2)

- **More informative values (referred to as weights or measure of association in the literature)**
  - **Point-wise Mutual Information**

$= P(w)P(w_i)$

if independent/unrelated

$$assoc_{PMI}(w, w_i) = \log_2 \frac{\boxed{P(w, w_i)}}{P(w)P(w_i)}$$

- **t-test**

$$assoct_{t-test}(w, w_i) = \frac{P(w, w_i) - P(w)P(w_i)}{\sqrt{P(w)P(w_i)}}$$

$\sqrt{\frac{s^2}{N}}$

Positive Pointwise Mutual Information

- PMI ranges from $-\infty$ to $+\infty$
- But the negative values are problematic
  - Things are co-occurring less than we expect by chance
  - Unreliable without enormous corpora
    - Imagine w1 and w2 whose probability is each 10^{-6}
    - Hard to be sure p(w1,w2) is significantly different than 10^{-12}
  - Plus it's not clear people are good at "unrelatedness"
- So we just replace negative PMI values by 0
- Positive PMI (PPMI) between word1 and word2:

$$\text{PPMI}(word_1, word_2) = \max\left(\log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}, 0\right)$$

PMI example

$$assoc_{PMI}(w, w_i) = \log_2 \frac{P(w, w_i)}{P(w)P(w_i)}$$

Assume w, w_i appear with equal frequency $\frac{1}{2^{10}}$

in a large set of documents

$P(w) = 2^{-10}$

$P(w_i) = 2^{-10}$

$P(w, w_i) =$

A $2^{-10} * 2^{-10} = 2^{-20}$ if the words are completely independent

B 2^{-10} if the words appear always together

A $assoc_{PMI} = \log_2 \frac{2^{-20}}{2^{-10} * 2^{-10}} = \log_2 1 = 0$

B $assoc_{PMI} = \log_2 \frac{2^{-10}}{2^{-10} * 2^{-10}} = \log_2 2^{10} = 10$

Other popular vector representations

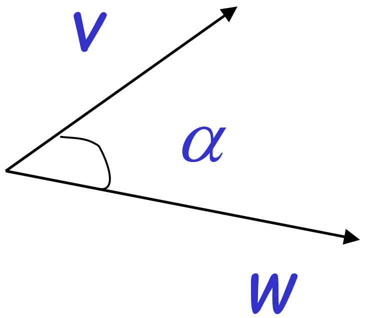
Dense vector representations (less dimensions):

1. **Singular value decomposition applied to word-word PointWise-MI matrix**
2. **Neural-Network-inspired models (skip-grams, CBOW)**

WS Distributional Methods (3)

- **Similarity between vectors**

$$sim_{cosine}(\vec{v},\vec{w}) = \frac{\vec{v}}{\|\vec{v}\|} \bullet \frac{\vec{w}}{\|\vec{w}\|} = \frac{\vec{v} \bullet \vec{w}}{\|\vec{v}\| \times \|\vec{w}\|} = \cos(\alpha)$$



Not sensitive to extreme values

$$sim_{Jaccard}(\vec{v},\vec{w}) = \frac{\sum_{i=1}^{N} \min(v_i, w_i)}{\sum_{i=1}^{N} \max(v_i, w_i)}$$

Normalized (weighted) number of overlapping features e.g.

$\vec{v}$ 2 1 0 3

$\vec{w}$ 3 1 1 2

$$\frac{2+1+0+2}{3+1+1+3} = \frac{5}{8}$$

Learning Goals for today's class

You can:

- Describe and Justify metrics to compute the similarity/distance of two concepts in an ontology
- Describe and Justify distributional metrics to compute the similarity/distance of two words (or phrases) in a Natural Language

Assignment-3 out – due Mar 30
(8-18 hours – working in pairs on programming parts is strongly advised)

Midterm marking almost done! Likely completed by tomorrow

Next class Wed

- **Natural language Processing (Context free grammars and parsing)**