Lecture 19

Paper Presentations

- **Adaptive Information Extraction from Text by Rule Induction and Generalisation by Fabio Ciravegna**

- **Finding and Linking Incidents in News Ao Feng and James Allan**

Modelling word relatedness

- Concepts. } formalisms.
  - → Synsets
  - → Linear algebraic (LSA)
  - → Set Theoretic (FCA)
  - Wikipedia (ESA)
  - → Logic (PSL)

A WordNet Primer

(Ack.: Harrold Somers for part or whole of some slides)

Limitations of dictionaries

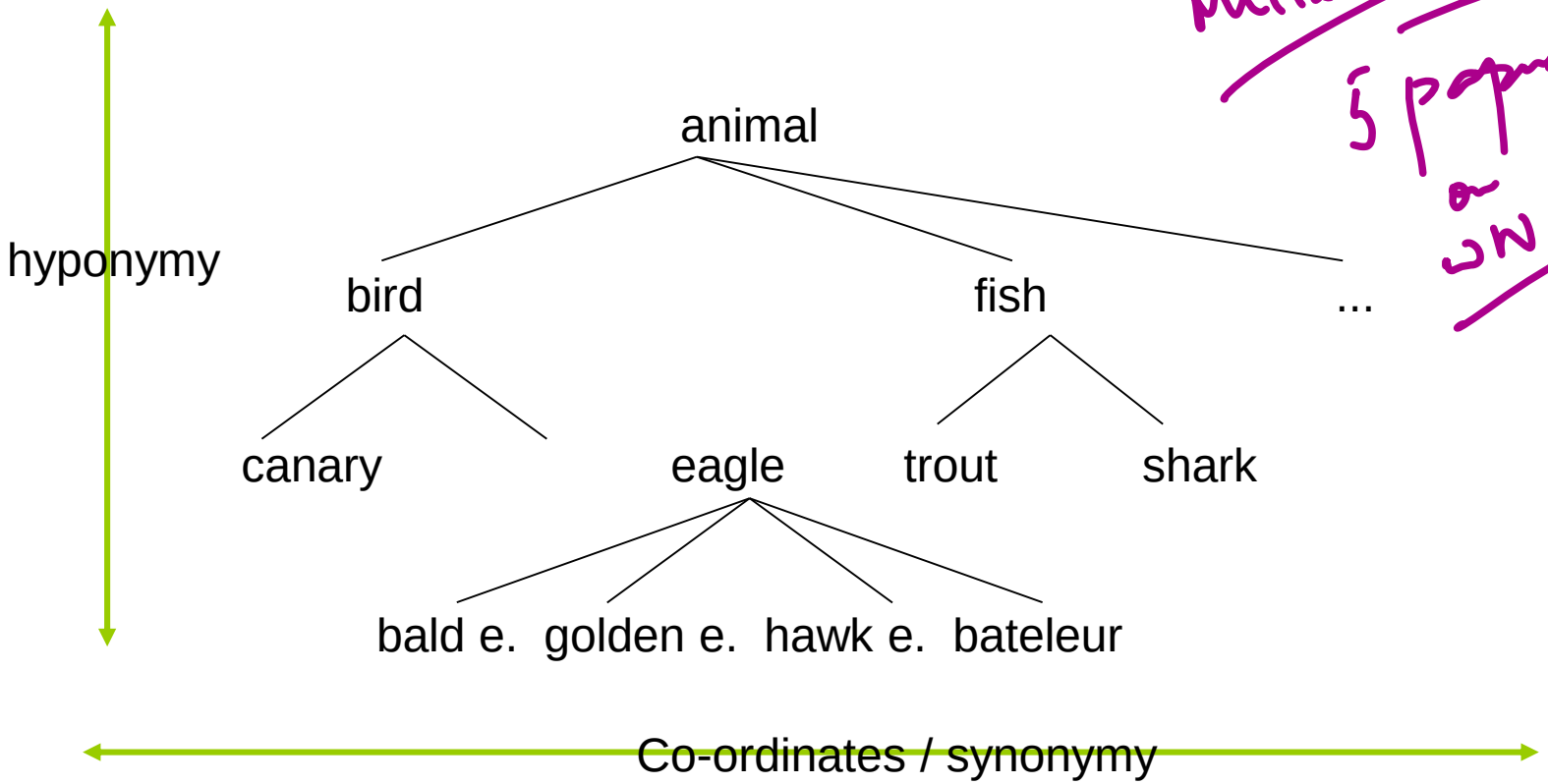
- No elaborate listing of features, only a pointer to the superconcept
- No information on co-ordinate terms or hyponyms
- Limits itself to a definition, Not encyclopaedic

WordNet: History

- 1985: a group of psychologists and linguists start to develop a “lexical database”
  - Princeton University
- theoretical basis: results from psycholinguistics and psycholexicology
  - What are properties of the “mental lexicon”?

Miller

5 papers on WN



- Psycholinguistic basis for WordNet : Do we have hierarchies in our heads ?

Global organisation

- division of the lexicon into five categories:
  - Nouns
  - Verbs
  - Adjectives
  - Adverbs
  - function words (“probably stored separately as part of the syntactic component of language”) [Miller et al.]

WordNet

- A hierarchically organized lexical database
- On-line thesaurus + aspects of a dictionary

Category	Unique Forms	# of Senses
Noun	117,097	145,104
Verb	11,488	24,890
Adjective	22,141	31,302
Adverb	4,601	5,720

- Several relations defined over each of these 4 categories

Meaning vs. form matrix

Illustrating the Concept of a Lexical Matrix:

F_1 and F_2 are synonyms; F_2 is polysemous

Word Meanings	Word Forms F_1	F_2	F_3	. . .	F_n
M_1	$E_{1,1}$	$E_{1,2}$			
M_2		$E_{2,2}$			
M_3			$E_{3,3}$		
$\vdots$				. . .	
M_m					$E_{m,n}$

Lexical semantics

- How are word meanings represented in WordNet?
  - synsets (synonym sets) as basic units
  - a word ‘meaning’ is represented by simply listing the word forms that can be used to express it
- example: senses of *board*
  - a piece of lumber vs. a group of people assembled for some purpose
  - synsets as unambiguous designators:
  - {board, plank, ...} vs. {board, committee, ...}
- Members of synsets are rarely true synonyms
  - WordNet does not attempt to capture subtle distinctions among members of the synset

Synsets

- synsets often sufficient for differential purposes
  - if an appropriate synonym is not available a short gloss may be used
    - e.g. {board, (a person's meals, provided regularly for money)}
  - Preferable for cardinality of synset to be >1
  - WordNet also gives a gloss for each word meaning, and (often) an example

Relations

Semantic relations → between synset (meanings)

lexical relations → between words (forms)

Antonymy.

rise ↔ fall

ascend ↔ descend

Relations in WordNet

- WordNet is organized by semantic relations.
  - It is characteristic of semantic relations that they are reciprocated
  - if there is a semantic relation R between meaning {x1, x2, ...} and meaning {y1, y2, ...}, then there is a relation R′ between {y1,y2, ...} and {x1, x2, ...}
  - Individual relations may or may not be
    - Reflexive R(A,A) is true (synonymy is, antonymy isn't)
    - Symmetric R(A,B) $\supset$ R(B,A) (eg synonymy, not hyponymy)
    - Transitive R(A,B) & R(B,C) $\supset$ R(A,C) (eg synonymy may be)

Synonymy

- similarity of meaning
  - Leibniz: two expressions are synonymous if the substitution of one for the other never changes the truth value of a sentence in which the substitution is made
- such global synonymy is rare (it would be redundant)
  - synonymy *relative to a context*: two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value
  - consequence of this synonymy in terms of substitutability: words in different syntactic categories cannot be synonyms

Antonymy

- antonym of a word x is sometimes not-x, but not always
  - *rich* and *poor* are antonyms
  - but: ***not*** *rich* does not imply *poor*
  - (because many people consider themselves neither rich nor poor)
- antonymy is a <u>lexical</u> relation between word <u>forms</u>, not a semantic relation between word meanings
  - meanings {rise, ascend} and {fall, descend} are conceptual opposites, but they are not antonyms [rise/fall] and [ascend/descend] are pairs of antonyms

Hypernymy/hyponymy

- hyponymy is a semantic relation between word *meanings*
  - {maple} is a hyponym of {tree}

Sub concept

Superconcept

- inverse: hypernymy
  - {tree} is a hypernym of {maple}
- also called: subordination/superordination; subset/superset; ISA relation
- test for hyponomy:
  - native speaker must accept sentences built from the frame “An x is a (kind of) y”
- called troponomy when applied to verbs
- Asymmetric and transitive

Holonymy/meronymy

- A concept represented by the synset {x1, x2,...} is a **meronym** of a concept represented by the synset {y1, y2, ...} if native speakers of English accept sentences constructed from such frames as “A y has an x (as a part)”, “An x is a part of y”.
- inverse relation: holonymy
- HAS-AS-PART
  - part hierarchy
  - part-of is asymmetric and (with caution) transitive

Meronymy

Winston et al. (1987) differentiate six types of meronyms:

component-object (*branch/tree*)
member-collection (*tree/forest*)
portion-mass (*slice/cake*)
stuff-object (*aluminum/airplane*)
feature-activity (*paying/shopping*) and
place-area (*Princeton/New Jersey*).

Chaffin, Hermann, and Winston (1988) add a seventh: phase-process (*adolescence/growing up*).

Only three of these types of meronymy are coded in WordNet:
Wm #p → *Wh* indicates that *Wm* is a component part of *Wh*;
Wm #m → *Wh* indicates that *Wm* is a member of *Wh*; and
Wm #s → *Wh* indicates that *Wm* is the stuff that *Wh* is made from.
Of these three, the 'is a component of' relation '#p' is by far the most frequent

Meronymy

- failures of transitivity caused by different part-whole relations, e.g.
  - *A musician has an arm*.
  - *An orchestra has a musician*.
  - but: ? *An orchestra has an arm*.
- Knowing where to stop is important :
  - Is atom a meronym of everything?
- If wheel is a meronym of vehicle, vehicles without wheels may inherit that property. Separate synset “wheeled vehicles” created

Noun

- S: (n) **orchestra** (a musical organization consisting of a group of instrumentalists including string players)
  - *direct hyponym* / *full hyponym*
    - S: (n) chamber orchestra (small orchestra; usually plays classical music)
    - S: (n) string orchestra (an orchestra playing only stringed instruments)
    - S: (n) symphony orchestra, symphony, philharmonic (a large orchestra; can perform symphonies) *"we heard the Vienna symphony"*
  - *part meronym*
    - S: (n) section (a division of an orchestra containing all instruments of the same class)
  - *direct hypernym* / *inherited hypernym* / ***sister term***
    - S: (n) musical organization, musical organisation, musical group (an organization of musicians who perform together)
      - S: (n) chorus (a group of people assembled to sing together)
      - S: (n) ensemble (a group of musicians playing or singing together) *"a string ensemble"*
      - S: (n) section (a division of an orchestra containing all instruments of the same class)
      - S: (n) duet, duette, duo (two performers or singers who perform together)
      - S: (n) trio (three performers or singers who perform together)
      - S: (n) quartet, quartette (four performers or singers who perform together)
      - S: (n) quintet, quintette (five performers or singers who perform together)
      - S: (n) sextet, sextette, sestet (six performers or singers who perform together)
      - S: (n) septet, septette (seven performers or singers who perform together)
      - S: (n) octet, octette (eight performers or singers who perform together)
      - S: (n) **orchestra** (a musical organization consisting of a group of instrumentalists including string players)
      - S: (n) band (instrumentalists not including string players)
      - S: (n) dance band, band, dance orchestra (a group of musicians playing popular music for dancing)
  - *derivationally related form*
- S: (n) **orchestra** (seating on the main floor in a theater)

WordNet’s noun hierarchy

- noun hierarchy partitioned into separate hierarchies with unique top hypernyms
- vague abstractions would be semantically empty, e.g. {entity} with immediate hyponyms {object, thing} and {idea}

- {act,action,activity}
- {animal,fauna}
- {artifact}
- {attribute,property}
- {body,corpus}
- {cognition,knowledge}
- {communication}
- {event,happening}
- {feeling,emotion}
- {food}
- {group,collection}
- {location,place}
- {motive}
- {natural object}
- {natural phenomenon}
- {person,human being}
- {plant,flora}
- {possession}
- {process}
- {quantity,amount}
- {relation}
- {shape}
- {state,condition}
- {substance}
- {time}

Nouns in WordNet

- Distinguising features :
  - Parts (small, yellow)
  - Attributes (beak, wings)
  - Functions (sing, fly)
- noun hierarchy as lexical inheritance system
  - seldom goes more than ten levels deep,
  - the deepest examples usually contain technical levels that are not part of everyday vocabulary
  - shallowest levels are too vague
  - “Inherited hypernym” option shows full hierarchy

S: (n) **pony** (any of various breeds of small gentle horses usually less than five feet high at the shoulder)

- *direct hyponym* / *full hyponym*
- *direct hypernym* / ***inherited hypernym*** / *sister term*
  - S: (n) horse, Equus caballus (solid-hoofed herbivorous quadruped domesticated since prehistoric times)
    - S: (n) equine, equid (hoofed mammals having slender legs and a flat coat with a narrow mane along the back of the neck)
      - S: (n) odd-toed ungulate, perissodactyl, perissodactyl mammal (placental mammals having hooves with an odd number of toes on each foot)
        - S: (n) ungulate, hoofed mammal (any of a number of mammals with hooves that are superficially similar but not necessarily closely related taxonomically)
          - S: (n) placental, placental mammal, eutherian, eutherian mammal (mammals having a placenta; all mammals except monotremes and marsupials)
            - S: (n) mammal, mammalian (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)
              - S: (n) vertebrate, craniate (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
                - S: (n) chordate (any animal of the phylum Chordata having a notochord or spinal column)
                - S: (n) animal, animate being, beast, brute, creature, fauna (a living organism characterized by voluntary movement)
                - 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))

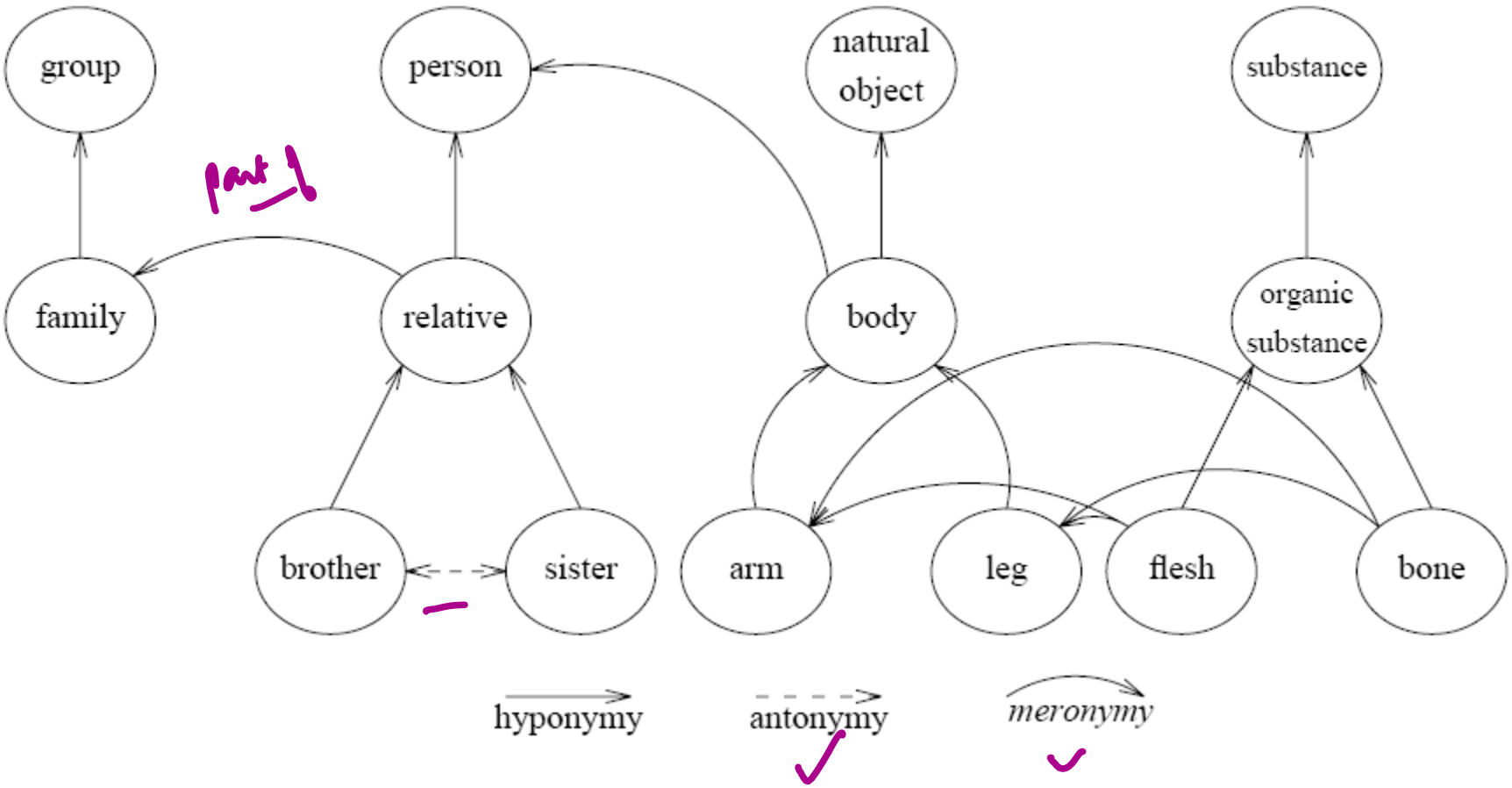
deep

shallow

Nouns in WordNet

- man-made artefacts: sometimes six or seven levels deep
  - roadster → car → motor vehicle → wheeled vehicle → vehicle → conveyance → artefact
- hierarchy of persons: about three or four levels
  - One of the deepest is: televangelist → evangelist → preacher → clergyman → spiritual leader → person
- Like all thesaurus structures, words can have multiple hypernyms

Noun Relations



Verbs

- Verbs are more polysemous than nouns: the nouns in *Collins* have on the average 1.74 senses, whereas verbs average 2.11 senses

- The most frequently used verbs (*have, be, run, make, set, go, take*, and others) are also the most polysemous
  - *I have a Mercedes* and *I have a headache*

- WordNet has 21,000 verb word forms and approximately 8,400 word meanings

- 15 files (categories) : verbs of bodily care and functions, change, cognition, communication, competition, consumption, contact, creation, emotion, motion, perception, possession, social interaction, weather verbs, states

Verb : synonymy

- Few true synonyms : {close, shut}
- {Ascend, rise} : what about temperature?
- {begin, commence}, {end, terminate}, {spit, expectorate}
- Verb synsets in Wordnet often contain periphrasatic expressions rather than lexicalized synonyms
  - {*whiten, become white*}, {*enrich, make rich*}

Verb: lexical entailment

- the relation between two verbs *V1* and *V2* that holds when the sentence *Someone V1* logically entails the sentence *Someone V2*;
- *He is snoring* entails *He is sleeping*
- Lexical entailment is a unilateral relation: if a verb *V1* entails another verb *V2*, then it cannot be that case that *V2* entails *V1*.
  - Exception : synonyms
- Negation reverses direction of entailment
- Temporal inclusion and entailment
  - Buy, pay
  - Sleep, snore

Verb : hyponymy

The *troponymy* relation between two verbs can be expressed by the formula

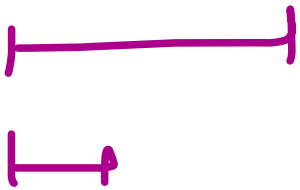
To V1 is to V2 in some particular manner

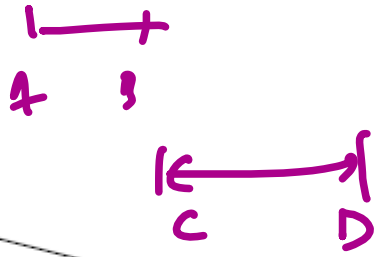
Troponyms of ‘fight’ : {battle, war, tourney, joust, duel, feud}

Relation with entailment :

Temporally co-extensive, but not proper inclusion

Contrast {limp,walk} and {snore, sleep} or {buy, pay}





Entailment

+Temporal Inclusion | **-Temporal Inclusion**

+Troponymy	**-Troponymy**	**Backward Presupposition**	**Cause**
(Co-extensiveness)	**(Proper Inclusion)**	*succeed-try*	*raise-rise*
limp-walk	*snore-sleep*	*untie-tie*	*give-have*
lisp-talk	*buy-pay*		

Four kinds of entailment relations among verbs

[Allen's Interval algebra]

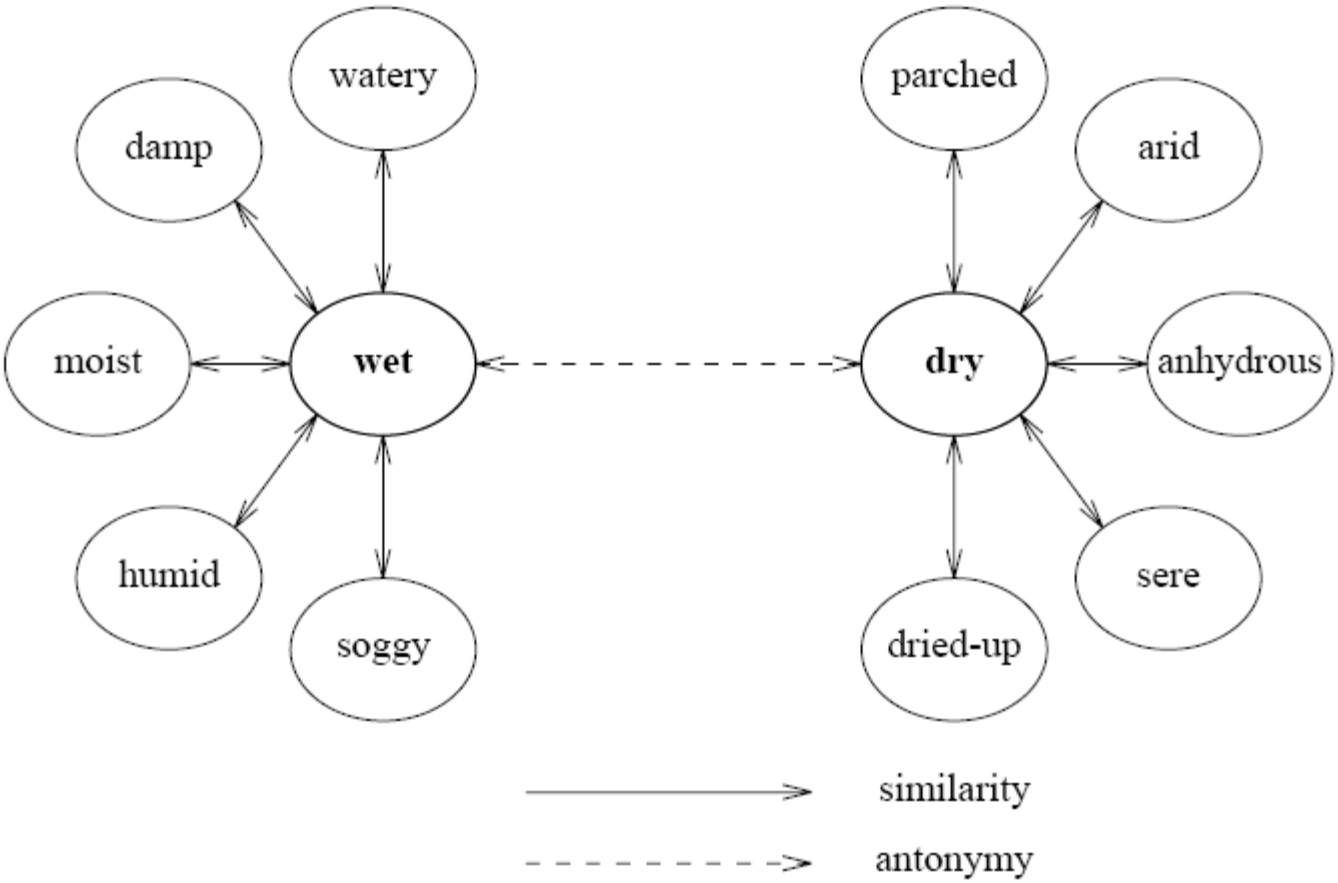
Adjectives

WordNet presently contains approximately 19,500 adjective word forms, organized into approximately 10,000 word meanings (synsets).

WordNet contains descriptive adjectives (such as *big*, *interesting*, *possible*) and relational adjectives (such as *presidential* and *nuclear*).

Relations between adjectives very different from relations between nouns, e.g. what does an is_a relation mean ?

Adjective Relations in WordNet



Bipolar Adjective Structure

Gradation

SIZE	WHITENESS	AGE	VIRTUE	VALUE	WARMTH
astronomical	snowy	ancient	saintly	superb	torrid
huge	**white**	**old**	**good**	great	**hot**
large	ash-gray	middle-aged	worthy	**good**	warm
standard	gray	mature	ordinary	mediocre	tepid
small	charcoal	adolescent	unworthy	**bad**	cool
tiny	**black**	**young**	**evil**	awful	**cold**
infinitesimal	pitch-black	infantile	fiendish	atrocious	frigid

Not coded in WordNet

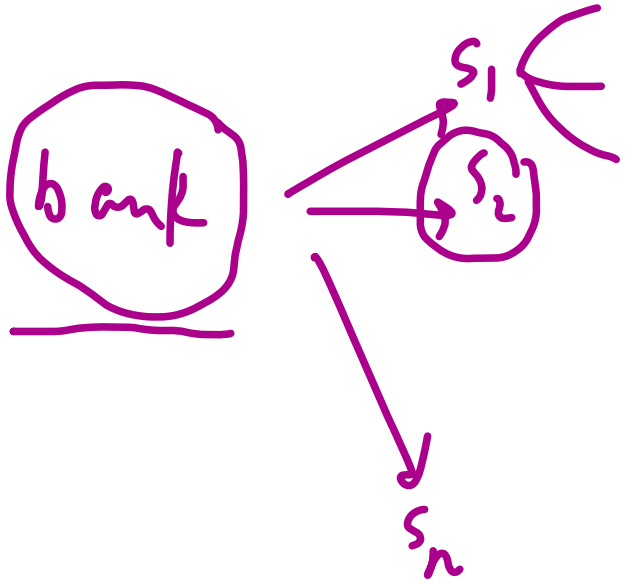
It was estimated that not more than 2% of the more than 2,500 adjective clusters could be organized in that way

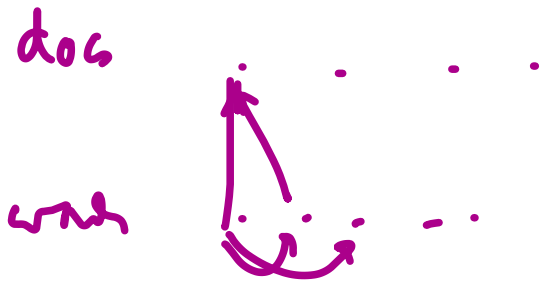
The WordNet System

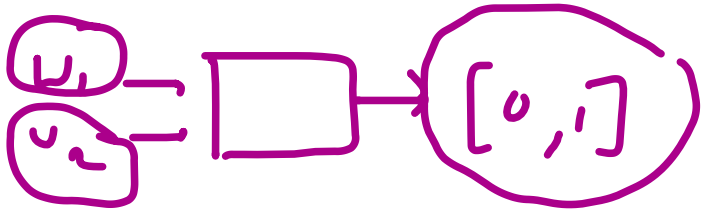
- Lexicographer’s source files
- Software to convert these files into the WordNet lexical database
- The WordNet lexical database
- A suite of software tools to access the database

What can WordNet be used for?

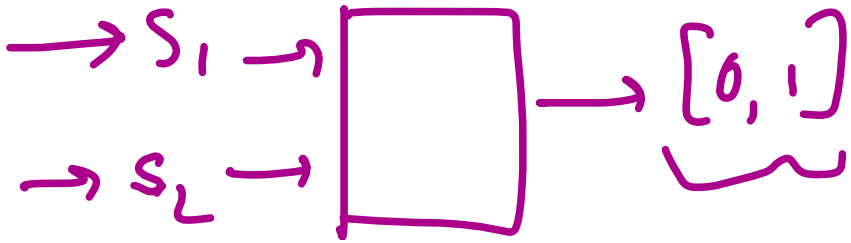
- As a lexical resource, an online dictionary, for human use
- Word-sense disambiguation
- Document classification
  - What is this text about? Look for recurring hypernyms
- Document retrieval
  - eg looking for texts about sports cars, search for synonyms and hyponyms of *sports car*
- Open-domain Q/A
  - Searching texts (eg WWW) to answer questions expressed in natural language
  - eg http://uk.ask.com/
- Textual entailment
  - Answering questions implied by text


bank
s_1
s_2
s_n





synsets.



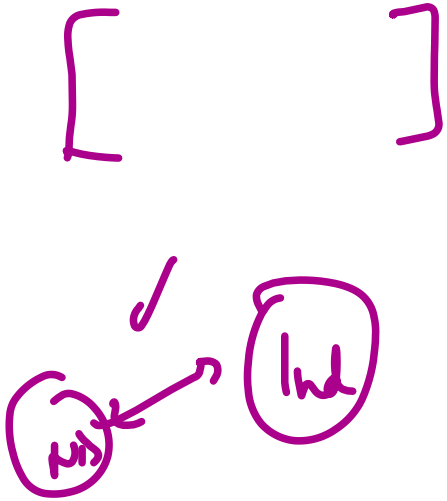
- Path-based measure
- Information theoretic measures.

bank $\leq m$

shore $\leq n$

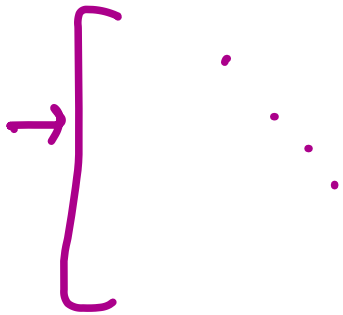
mn

Ontology



taxonomy.

is-a relationships.



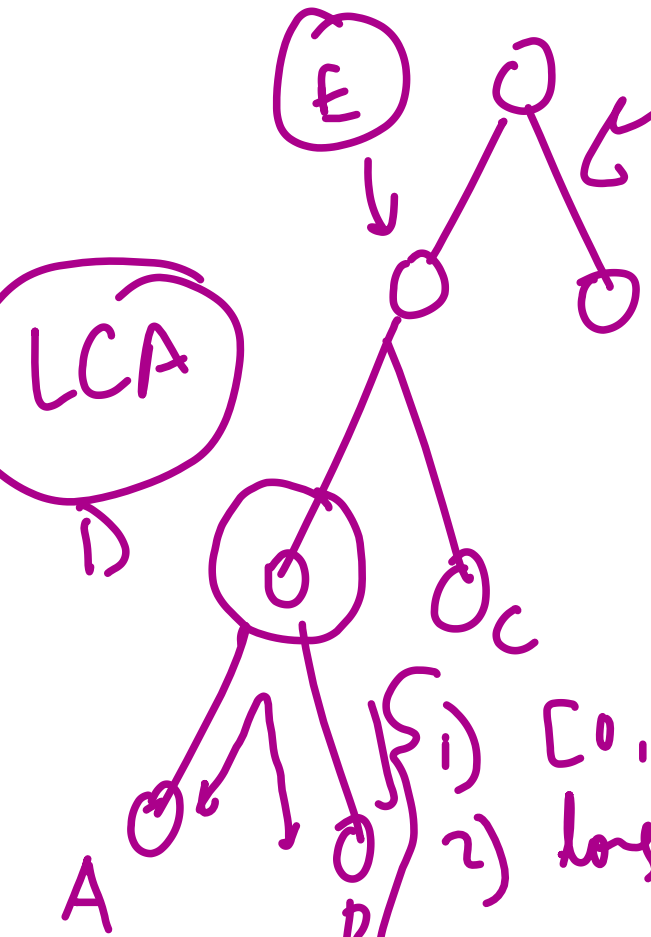
Path based measures.

IS_a.

Synset

VERY IMPORT

Sim (A, B) =



distances.

range

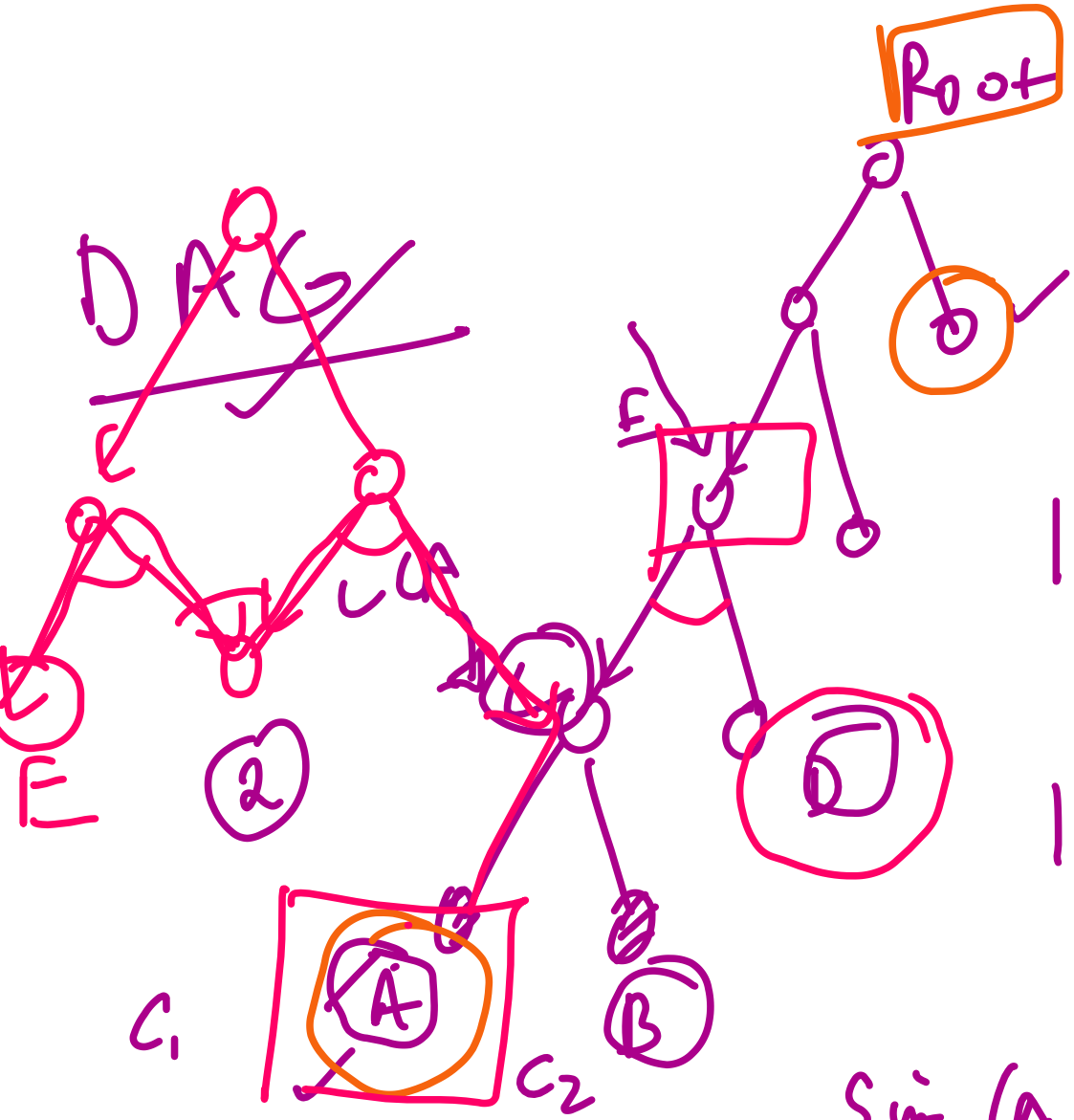
Similarity.

[0, 1]

1) [0, 1]

2) longer the path length, lesser the sim.

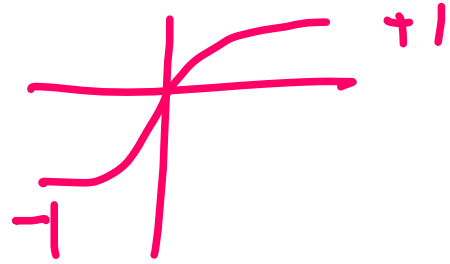
3) the deeper the LCA, the more sim the nodes

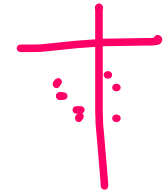


$$\mathrm{sim}(C_1, C_2) = 1 - \left[\frac{\text{Sum of distance to LCA}}{\text{Sum of dist. to root}}\right]$$

$$1 - \frac{1+1}{4+4} = 1 - \frac{2}{8} = .75$$

$$\mathrm{sim}(A, C) = 1 - \frac{1+0}{4+3} = 1 - \frac{1}{7}$$





Hirst & St. Onge

$$\mathrm{sim}(c_1, c_2) = C - \text{Path-length}(c_1, c_2) - k \times d$$

Const

Const

$[0, 1]$

no. of turns.

Liu

$$e^{-\alpha L}\left[\frac{e^{\beta H} - e^{-\beta H}}{e^{\beta H} + e^{-\beta H}}\right]$$

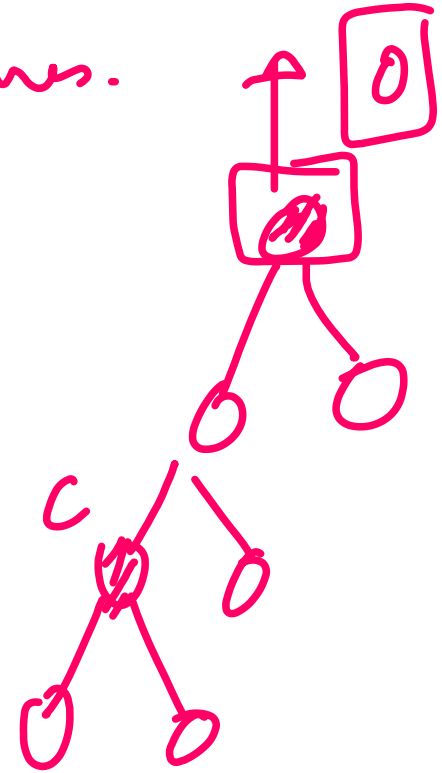
tanh

L: path length

H: depth of LCA

Information Theoretic Measures.

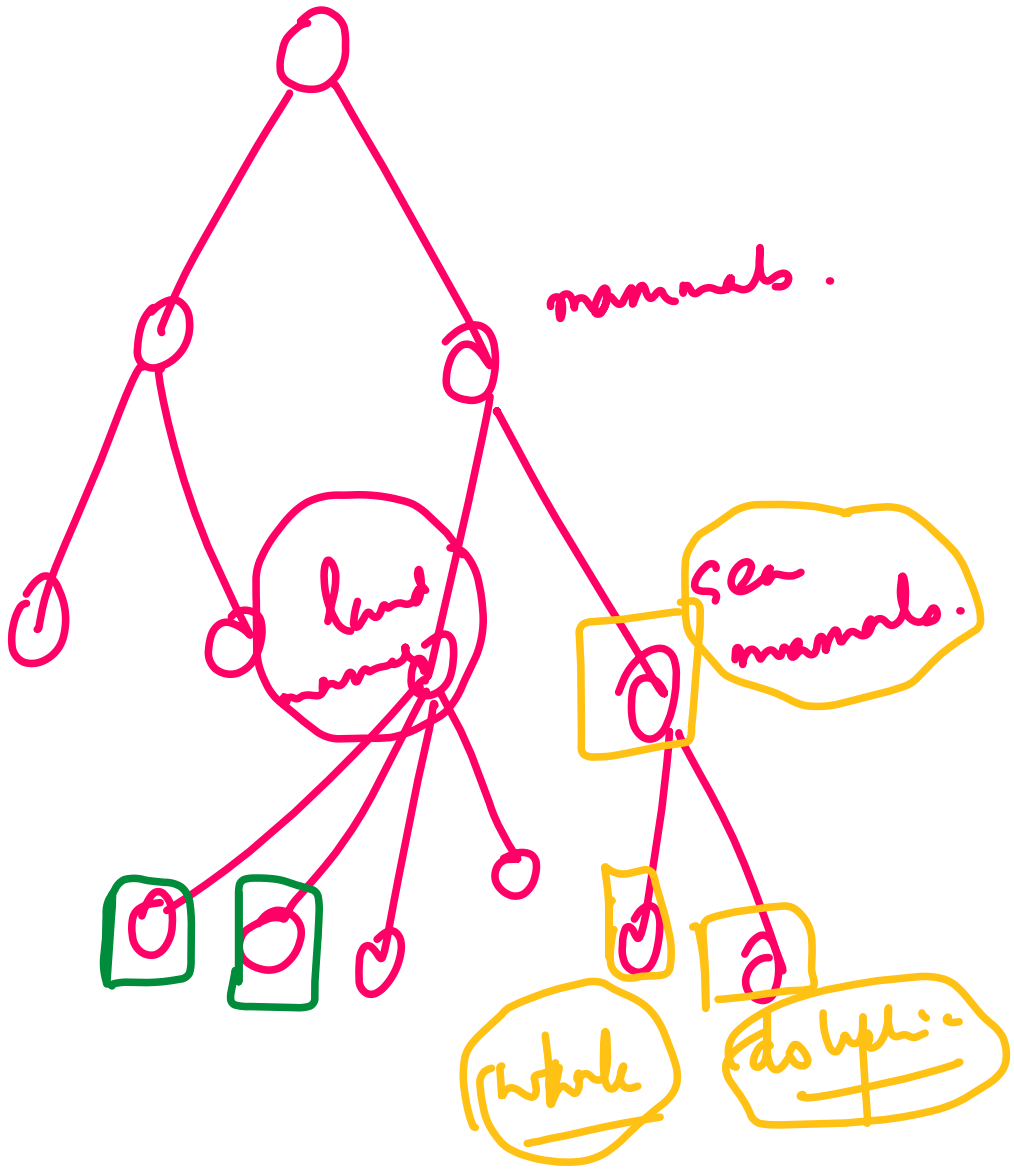
$$\boxed{IC(c) = \log \frac{1}{p(c)}}$$



$$p(c) = \frac{n_c}{N}$$

Resnik measure.

Sem. relatedness $(c_1, c_2) = IC(LCA(c_1, c_2))$


mammals.
land mammals
sea mammals.
whale
dolphin

apple

fruit

edible

entity

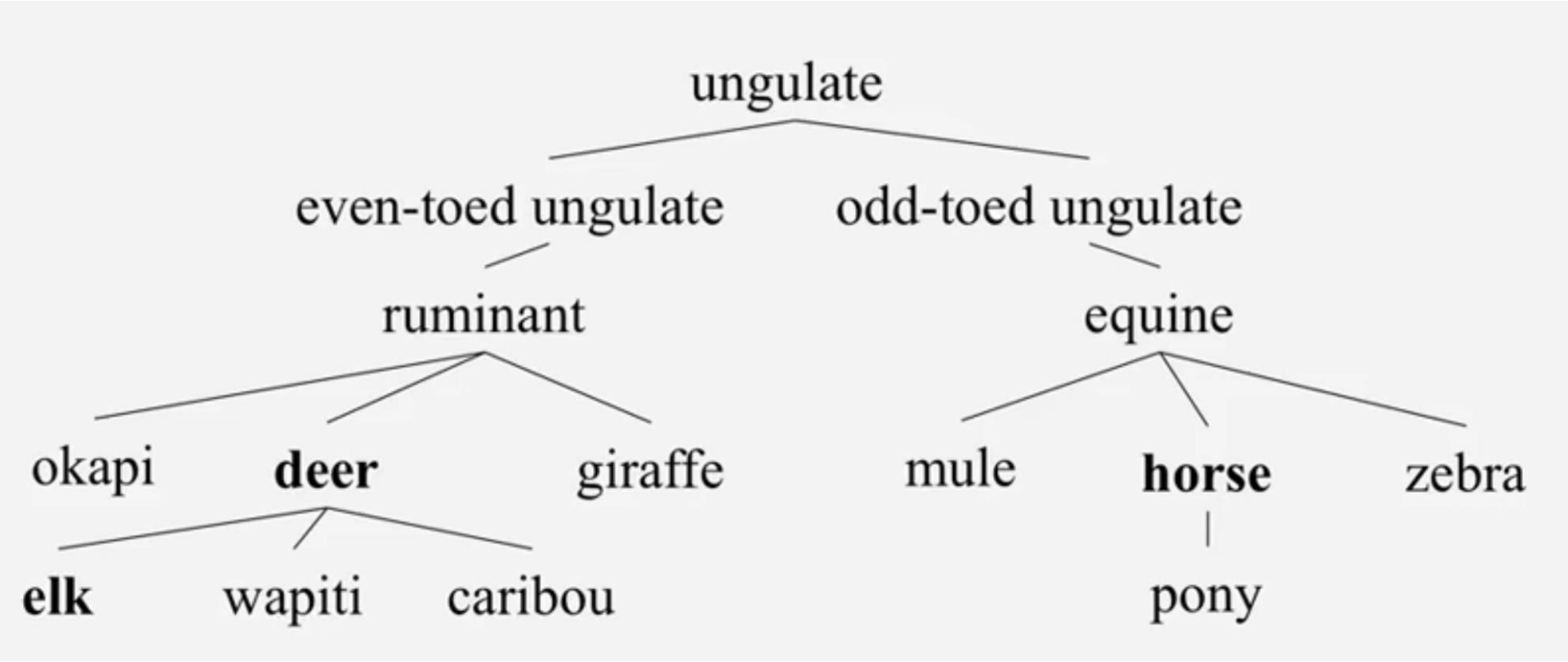
Specific

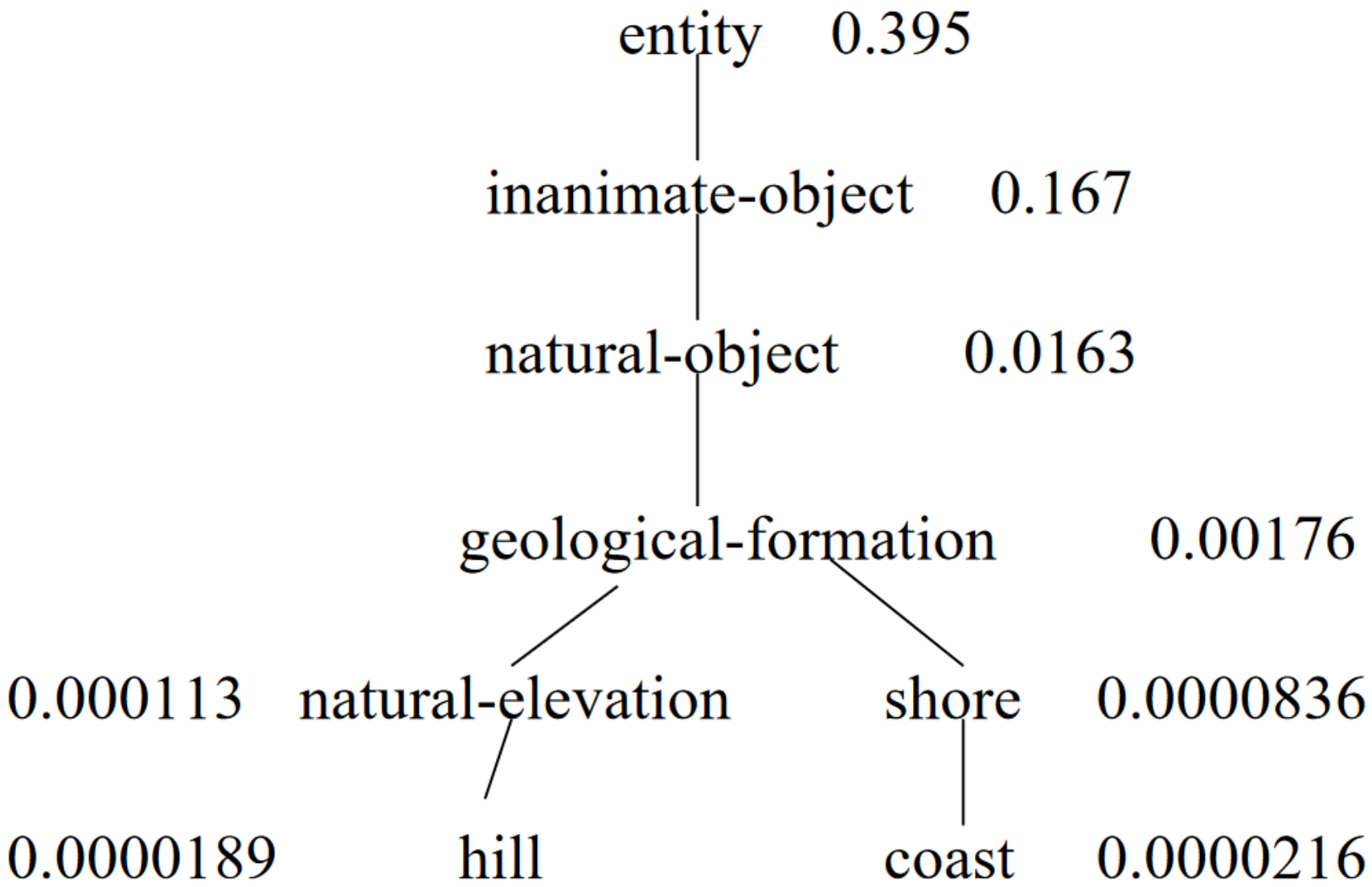
trick in

Counting

Lin measure

$$\left[\frac{2 \log p(LCA(c_1, c_2))}{\log p(c_1) + \log p(c_2)}\right]$$


ungulate
even-toed ungulate
odd-toed ungulate
ruminant
equine
okapi
deer
giraffe
mule
horse
zebra
elk
wapiti
caribou
pony


entity 0.395
inanimate-object 0.167
natural-object 0.0163
geological-formation 0.00176
0.000113 natural-elevation
shore 0.0000836
0.0000189 hill
coast 0.0000216

Adapted Lesk (2002)

- 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

Ack: Steve Vincent