Lexical Semantics

Lexicon vs. Vocabulary

LEXICON
VERSUS
VOCABULARY

LEXICON	VOCABULARY
List of words along with its related knowledge on their linguistic significance and usage	List of words a person knows of a particular language
Has a broader coverage on the linguistic significance and the usage of the words in a language	Defines the list of words a person knows of a language
Often used in a linguistic context	Used on a general context on daily use among people

Visit www.PEDIAA.com

Definitions

Lexical Semantics

Lexical semantics looks at the meaning related connections among **lexical items**.

Lexical item

"lexical item" is just a fancy way of saying a "word"

Definitions

Lexical Semantics
Lexical semantics looks at the meaning related connections among **lexical items**.

Lexical item
"lexical item" is just a fancy way of saying a "word"
To identify the semantics of lexical items, we need to focus on the notion of **lexeme**, an individual entry in the lexicon.

Lexeme
Lexeme contains a pairing of a particular orthographic and phonological form with some sort of symbolic meaning representation (Sense).

Examples
Lexeme: RUN
Lexical item: run, runs, ran and running

Example

verge [1] | vərj |

noun

an edge or border: *they came down to the verge of the lake.*
• an extreme limit beyond which something specified will happen: *I was **on the verge of** tears.*
• Brit. a grass edging such as that by the side of a road or path.
• Architecture an edge of tiles projecting over a gable.

verb [no obj.] (**verge on**)
approach (something) closely; be close or similar to (something): *despair verging on the suicidal.*

ORIGIN late Middle English: via Old French from Latin ***virga 'rod.'*** The current verb sense dates from the late 18th cent.

verge [2] | vərj |

noun

a wand or rod carried before a bishop or dean as an emblem of office.

ORIGIN late Middle English: from Latin ***virga 'rod.'***

verge [3] | vərj |

verb [no obj.]
incline in a certain direction or toward a particular state: *his style verged into the art nouveau school.*

ORIGIN early 17th cent. (in the sense ***'descend (to the horizon)'***): from Latin ***vergere 'to bend, incline.'***

Extracting senses from definitions

Definitions from the American Heritage Dictionary (Morris, 1985)

- **right** *adj.* located near the right hand esp. being on the right when facing the same direction as the observer
- **left** *adj.* located near to this side of the body than the right
- **red** *n.* the color of blood or a ruby
- **blood** *n.* the red liquid that circulates in the heart, arteries and veins of animals

Extracting senses from definitions

- **right** *adj.* located near the right hand esp. being on the right when facing the same direction as the observer
- **left** *adj.* located near to this side of the body than the right
- **red** *n.* the color of blood or a ruby
- **blood** *n.* the red liquid that circulates in the heart, arteries and veins of animals

Observations

- Define a sense through its relationship with other senses
- One sense of *right* and *left* represents some kind of alternation, or opposition, to one another.
- We can glean that *red* is a color, it can be applied to both *blood* and *rubies*, and that *blood* is a liquid.

Relations Between Senses

- Homonymy
- Polysemy
- Synonymy
- Antonymy
- Hypernymy
- Hyponymy
- Meronymy

Homonymy

Definition

Homonymy is defined as a relation that holds between words that have the same form (sound alike or are spelled alike) with unrelated senses.

- Bat (wooden stick-like thing) vs Bat (flying mammal thing)
- Bank (financial institution) vs Bank (riverside)

Homonymy

Definition

Homonymy is defined as a relation that holds between words that have the same form (sound alike or are spelled alike) with unrelated senses.

- Bat (wooden stick-like thing) vs Bat (flying mammal thing)
- Bank (financial institution) vs Bank (riverside)

Homophones and Homographs

Homophones are the words with the same pronunciation (different spellings) but different senses.

- write vs right
- piece vs peace

Homographs are the lexemes with the same orthographic form but different senses. Ex: Bass

Problems for NLP applications

Text-to-Speech
Same orthographic form but different phonological form

Information Retrieval
Different meaning but same orthographic form

Speech Recognition
to, two, too
Perfect homonyms are also problematic

Polysemy

Multiple **related** senses within a single lexeme.

- The *bank* was constructed in 1875 out of local red brick.
- I withdrew the money from the *bank*.

Polysemy

Multiple **related** senses within a single lexeme.

- The *bank* was constructed in 1875 out of local red brick.
- I withdrew the money from the *bank*.

Are those the same sense?

- Sense 1: "The building belonging to a financial institution"
- Sense 2: "A financial institution"

Another example

- Heavy snow caused the roof of the *school* to collapse. The
- *school* hired more teachers this year than ever before.

Polysemy: multiple related senses

Often, the relationships are systematic

E.g., building vs. organization

school, university, hospital, church, supermarket

More examples:

- Author (Jane Austen wrote Emma) ↔Works of Author (I really love Jane Austen)
- Animal (The chicken was domesticated in Asia) ↔Meat (The chicken was overcooked)
- Tree (Plums have beautiful blossoms) ↔Fruit (I ate a preserved plum yesterday)

Polysemy: multiple related meanings

Zeugma test

- Which of these flights *serve* breakfast?
- Does Midwest Express *serve* Philadelphia?

*Does Midwest Express serve breakfast and San Jose?

Combine two separate uses of a lexeme into a single example using conjunction

Since it sounds weird, we say that these are two different senses of *serve*.

Synonymy

Words that have the identical, or nearly identical sense

- couch / sofa
- big / large
- automobile / car
- vomit / throw up
- water / H_2O

Two lexemes are synonyms if they can be successfully substituted for each other in all situations.

Synonymy: A relation between senses

Consider the lexemes *big* and *large*.

Are they synonyms?

- How **big** is that plane?
- Would I be flying on a **large** or small plane?

Synonymy: A relation between senses

Consider the lexemes *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, for instance, became a kind of **big** sister to Benjamin.
- *Miss Nelson, for instance, became a kind of **large** sister to Benjamin.

Synonymy: A relation between senses

Consider the lexemes *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, for instance, became a kind of **big** sister to Benjamin.
- *Miss Nelson, for instance, became a kind of **large** sister to Benjamin.

Why?

- *big* has a sense that means being older, or grown up
- *large* lacks this sense

Synonyms

Shades of meaning

- What is the cheapest first class *fare*?
- *What is the cheapest first class *price*?

Collocational constraints

- We frustate 'em and frustate 'em, and pretty soon they make a *big* mistake.
- *We frustate 'em and frustate 'em, and pretty soon they make a *large* mistake.

Antonyms

- Senses that are opposites with respect to one feature of their meaning
- Otherwise, they are similar!
  - dark / light
  - short / long
  - hot / cold
  - up / down
  - in / out

More formally: antonyms can

- define a binary opposition or at opposite ends of a scale (*long/short*, *fast/slow*)

- Be **reversives**: *rise/fall*

Hyponymy and Hypernymy

Hyponymy

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*
- *dog* is a hyponym of *animal*
- *mango* is a hyponym of *fruit*

Hyponymy and Hypernymy

Hyponymy

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*
- *dog* is a hyponym of *animal*
- *mango* is a hyponym of *fruit*

Hypernymy

Conversely

- *vehicle* is a hypernym/superordinate of *car*
- *animal* is a hypernym of *dog*
- *fruit* is a hypernym of *mango*

Hyponymy more formally

Entailment

Sense A is a hyponym of sense B if being an A entails being a B.

Ex: dog, animal

Transitivity

A hypo B and B hypo C entails A hypo C

Meronyms and holonyms

Definition

Meronymy: an asymmetric, transitive relation between senses.

X is a **meronym** of Y if it denotes a part of Y.

The inverse relation is **holonymy**.

meronym	holonym
porch	house
wheel	car
leg	chair
nose	face

WordNet

https://wordnet.princeton.edu/wordnet/

- A hierarchically organized lexical database
- A machine-readable thesaurus, and aspects of a dictionary
- Versions for other languages are under development

| part of speech	no. synsets
noun	82,115
verb	13,767
adjective	18,156
adverb	3,621

Synsets in WordNet

- A **synset** is a set of synonyms representing a sense
- Example: chump as a noun to mean 'a person who is gullible and easy to take advantage of'

  $\{$chump^1, fool^2, gull^1, mark^9, patsy^1, fall guy^1, sucker^1, soft touch^1, mug$^2\}$

- Each of these senses share this same gloss.
- For WordNet, the meaning of this sense of chump is this list.

lemma vs. synsets
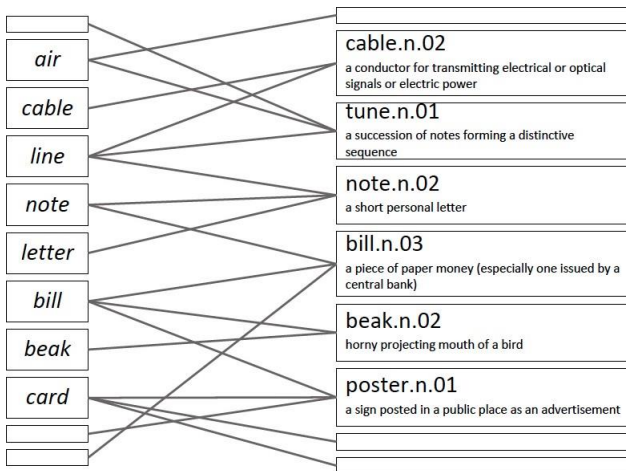
All relations in WordNet

```
searchtype is at least one of the following:
        -ants{n|v|a|r}          Antonyms
        -hype{n|v}              Hypernyms
        -hypo{n|v}, -tree{n|v}  Hyponyms & Hyponym Tree
        -entav                  Verb Entailment
        -syns{n|v|a|r}          Synonyms (ordered by estimated frequency)
        -smemn                  Member of Holonyms
        -ssubn                  Substance of Holonyms
        -sprtn                  Part of Holonyms
        -membn                  Has Member Meronyms
        -subsn                  Has Substance Meronyms
        -partn                  Has Part Meronyms
        -meron                  All Meronyms
        -holon                  All Holonyms
        -causv                  Cause to
        -pert{a|r}              Pertainyms
        -attr{n|a}              Attributes
        -deri{n|v}              Derived Forms
        -domn{n|v|a|r}          Domain
        -domt{n|v|a|r}          Domain Terms
        -faml{n|v|a|r}          Familiarity & Polysemy Count
        -framv                  Verb Frames
        -coor{n|v}              Coordinate Terms (sisters)
        -simsv                  Synonyms (grouped by similarity of meaning)
        -hmern                  Hierarchical Meronyms
        -hholn                  Hierarchical Holonyms
        -grep{n|v|a|r}          List of Compound Words
        -over     _             Overview of Senses
```

Wordnet noun and verb relations

Relation	Also called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	$breakfast^1 \rightarrow meal^1$
Hyponym	Subordinate	From concepts to subtypes	$meal^1 \rightarrow lunch^1$
Member Meronym	Has-Member	From groups to their members	$faculty^2 \rightarrow professor^1$
Has-Instance		From concepts to instances of the concept	$composer^1 \rightarrow Bach^1$
Instance		From instances to their concepts	$Austen^1 \rightarrow author^1$
Member Holonym	Member-Of	From members to their groups	$copilot^1 \rightarrow crew^1$
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		Opposites	$leader^1 \rightarrow follower^1$

Relation	Definition	Example
Hypernym	From events to superordinate events	$fly^9 \rightarrow travel^5$
Troponym	From a verb (event) to a specific manner elaboration of that verb	$walk^1 \rightarrow stroll^1$
Entails	From verbs (events) to the verbs (events) they entail	$snore^1 \rightarrow sleep^1$
Antonym	Opposites	$increase^1 \iff decrease^1$

WordNet Hierarchies

```
Synonyms/Hypernyms (Ordered by Estimated Frequency) of noun mouse

4 senses of mouse

Sense 1
mouse
      => rodent, gnawer
         => placental, placental mammal, eutherian, eutherian mammal
            => mammal, mammalian
               => vertebrate, craniate
                  => chordate
                     => animal, animate being, beast, brute, creature, fauna
                        => organism, being
                           => living thing, animate thing
                              => whole, unit
                                 => object, physical object
                                    => physical entity
                                       => entity


      Sense 4
      mouse, computer mouse
            => electronic device
               => device
                  => instrumentality, instrumentation
                     => artifact, artefact
                        => whole, unit
                           => object, physical object
                              => physical entity
                                 => entity
```

Word Similarity

- Synonymy is a binary relation
  - Two words are either synonymous or not

Word Similarity

- Synonymy is a binary relation
    - Two words are either synonymous or not
- We want a looser metric
    - Word similarity or
    - Word distance

Word Similarity

- Synonymy is a binary relation
  - Two words are either synonymous or not
- We want a looser metric
  - Word similarity or
  - Word distance
- Two words are more similar
  - If they share more features of meaning

Word Similarity

- Synonymy is a binary relation
  - Two words are either synonymous or not
- We want a looser metric
  - Word similarity or
  - Word distance
- Two words are more similar
  - If they share more features of meaning
- Actually these are really relations between **senses**:
  - Instead of saying "bank is like fund"
  - We say
    - Bank[1] is similar to fund[3]
    - Bank[2] is similar to slope[5]

Word Similarity

- Synonymy is a binary relation
  - Two words are either synonymous or not
- We want a looser metric
  - Word similarity or
  - Word distance
- Two words are more similar
  - If they share more features of meaning
- Actually these are really relations between **senses**:
  - Instead of saying "bank is like fund"
  - We say
    - Bank[1] is similar to fund[3]
    - Bank[2] is similar to slope[5]
- We will compute similarity over both words and senses
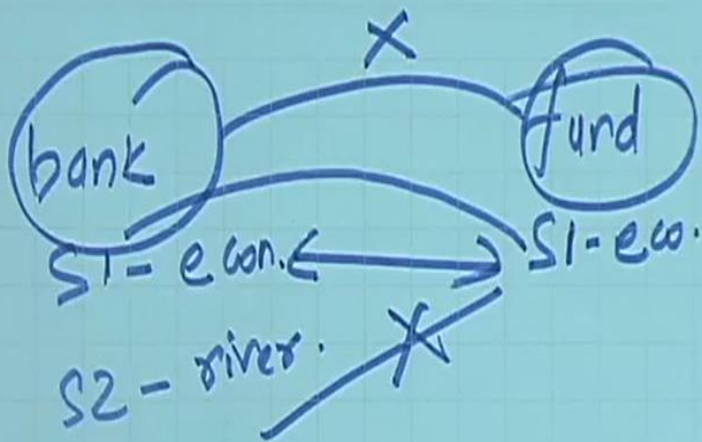
Two classes of algorithms

Distributional algorithms
By comparing words based on their distributional context in the corpora

Thesaurus-based algorithms
Based on whether words are "nearby" in WordNet

Thesaurus-based Word Similarity

- We could use anything in the thesaurus:
    - Meronymy, hyponymy, troponymy
    - Glosses and example sentences
- In practice, "thesaurus-based" methods usually use:
    - the is-a/subsumption/hypernymy hierarchy
    - and sometimes the glosses too
- Word similarity vs. word relatedness
    - Similar words are near-synonyms
    - Related words could be related any way
        - car, gasoline : related, but nor similar
        - car, bicycle: similar

Path-based similarity

Basic Idea

- Two words are similar if they are nearby in the hypernym graph

Path-based similarity

Basic Idea

- Two words are similar if they are nearby in the hypernym graph
- $pathlen(c_1, c_2)$ = number of edges in shortest path (in hypernym graph) between senses c_1 and c_2

Path-based similarity

Basic Idea

- Two words are similar if they are nearby in the hypernym graph
- $pathlen(c_1, c_2)$ = number of edges in shortest path (in hypernym graph) between senses c_1 and c_2
- $sim_{path}(c_1, c_2) = \frac{1}{1 + pathlen(c_1, c_2)}$

$W_1 \longleftarrow W_2 \longrightarrow \max \; \text{sim}(S_{1i}, S_{2j})$

S_{21}

S_{11}

$S_{12} \longrightarrow S_{22}$

S_{13}
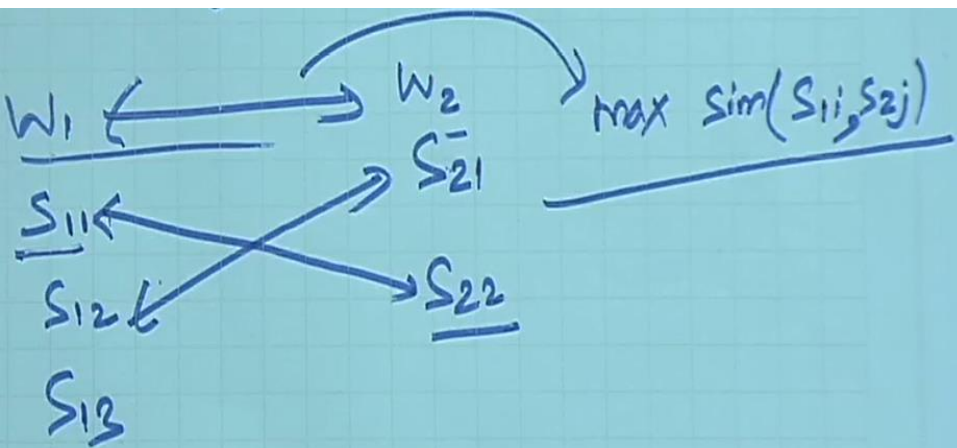
Path-based similarity

Basic Idea

- Two words are similar if they are nearby in the hypernym graph
- $pathlen(c_1, c_2)$ = number of edges in shortest path (in hypernym graph) between senses c_1 and c_2
- $sim_{path}(c_1, c_2) = \frac{1}{1 + pathlen(c_1, c_2)}$
- $sim(w_1, w_2) = max_{c_1 \in senses(w_1), c_2 \in senses(w_2)} sim(c_1, c_2)$

Shortest path in the hierarchy
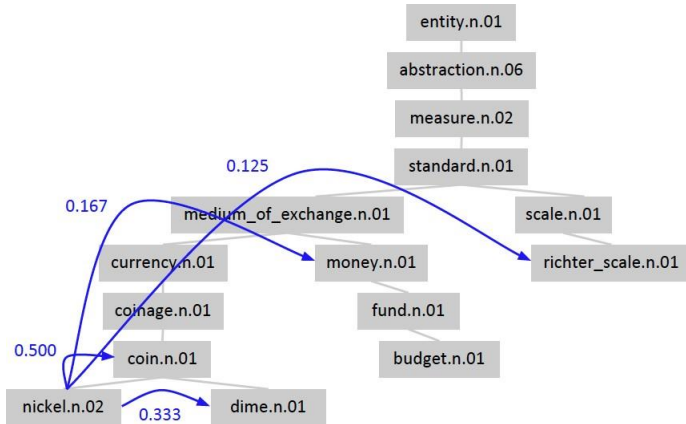
Leacock-Chodorow (L-C) Similarity

L-C similarity

$$sim_{LC}(c_1, c_2) = -log(pathlen(c_1, c_2)/2d)$$

d: maximum depth of the hierarchy

Problems with L-C similarity

- Assumes each edge represents a uniform distance
- 'nickel-money' seems closer than 'nickel-standard'
- We want a metric which lets us assign different "lengths" to different edges - but how?

Concept probability models

Cencept probabilities

- For each concept (synset) c, let $P(c)$ be the probability that a randomly selected word in a corpus is an instance (hyponym) of c

Concept probability models

Cencept probabilities

- For each concept (synset) c, let $P(c)$ be the probability that a randomly selected word in a corpus is an instance (hyponym) of c
- $P(ROOT) = 1$
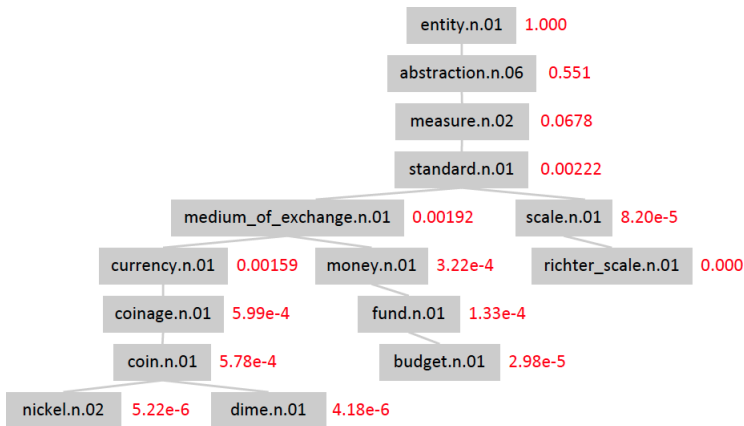- The lower a node in the hierarchy, the lower its probability

Concept probability models

Cencept probabilities

- For each concept (synset) c, let $P(c)$ be the probability that a randomly selected word in a corpus is an instance (hyponym) of c
- $P(ROOT) = I$
- The lower a node in the hierarchy, the lower its probability

Estimating concept probabilities

- Train by counting "concept activations" in a corpus
- Each occurrence of *dime* also increments counts for *coin, currency, standard*, etc.

ROOT -entity

Example : concept count

entity.n.01 — 1,915,712
abstraction.n.06 — 1,055,337
measure.n.02 — 129,897
standard.n.01 — 4,253
medium_of_exchange.n.01 — 3,680
scale.n.01 — 157
currency.n.01 — 3,043
money.n.01 — 616
richter_scale.n.01 — 0
coinage.n.01 — 1,149
fund.n.01 — 254
coin.n.01 — 1,108
budget.n.01 — 57
nickel.n.02 — 10
dime.n.01 — 8
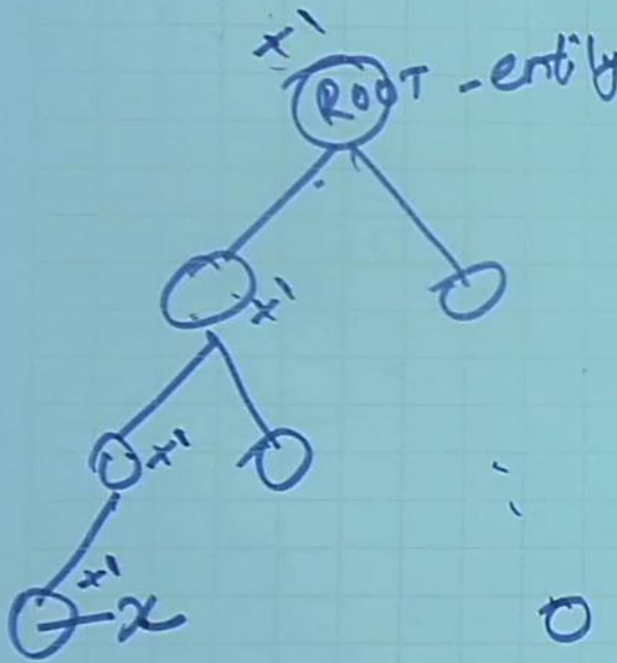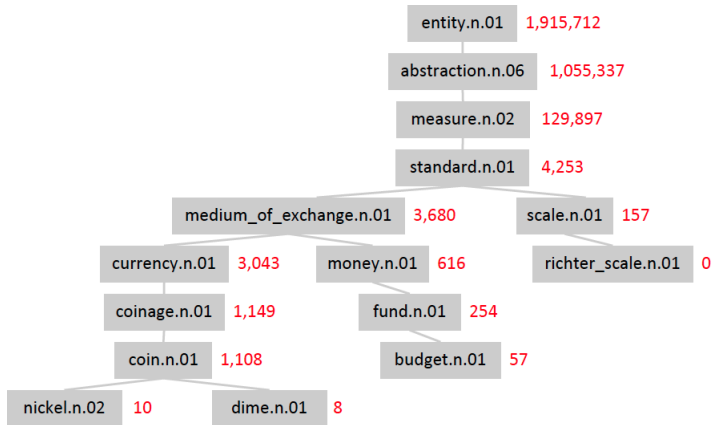
Example : concept probabilities

- entity.n.01 1.000
- abstraction.n.06 0.551
- measure.n.02 0.0678
- standard.n.01 0.00222
- medium_of_exchange.n.01 0.00192
- scale.n.01 8.20e-5
- currency.n.01 0.00159
- money.n.01 3.22e-4
- richter_scale.n.01 0.000
- coinage.n.01 5.99e-4
- fund.n.01 1.33e-4
- coin.n.01 5.78e-4
- budget.n.01 2.98e-5
- nickel.n.02 5.22e-6
- dime.n.01 4.18e-6

Information content

Information content

- Information content: $IC(c) = -logP(c)$
- Lowest common subsumer : $LCS(c_1, c_2)$: the lowest node in the hierarchy that subsumes (is a hypernym of) both c_1 and c_2
- We are now ready to see how to use information content (IC) as a similarity metric.

Example : Information content

Node	Value
entity.n.01	0.000
abstraction.n.06	0.596
measure.n.02	2.691
standard.n.01	6.110
medium_of_exchange.n.01	6.255
scale.n.01	9.409
currency.n.01	6.445
money.n.01	8.042
richter_scale.n.01	Inf
coinage.n.01	7.419
fund.n.01	8.928
coin.n.01	7.455
budget.n.01	10.423
nickel.n.02	12.163
dime.n.01	12.386
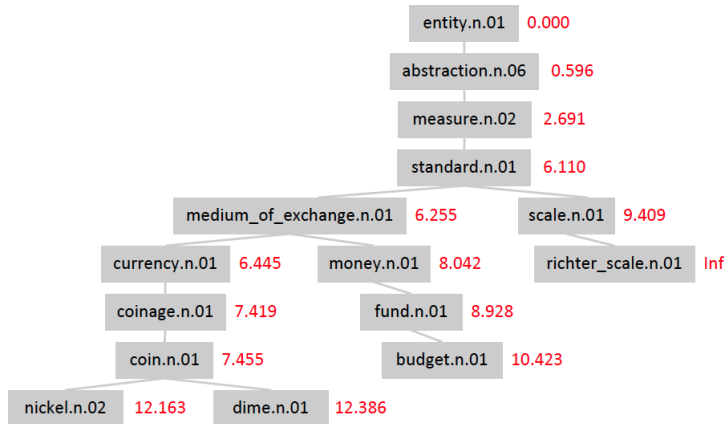
Resnik Similarity

Resnik Similarity

- Intuition: how similar two words are depends on how much they have in common

- It measures the commonality by the information content of the lowest common subsumer

- $sim_{resnik}(c_1, c_2) = IC(LCS(c_1, c_2)) = -logP(LCS(c_1, c_2))$

Example: Resnik similarity

Node	Value
entity.n.01	0.000
abstraction.n.06	0.596
measure.n.02	2.691
standard.n.01	6.110
scale.n.01	9.409
medium_of_exchange.n.01	6.255
currency.n.01	6.445
money.n.01	8.042
richter_scale.n.01	Inf
coinage.n.01	7.419
fund.n.01	8.928
coin.n.01	7.455
budget.n.01	10.423
nickel.n.02	12.163
dime.n.01	12.386

Arc labels: 6.110, 6.255, 6.255, 7.455, 12.163, 7.455
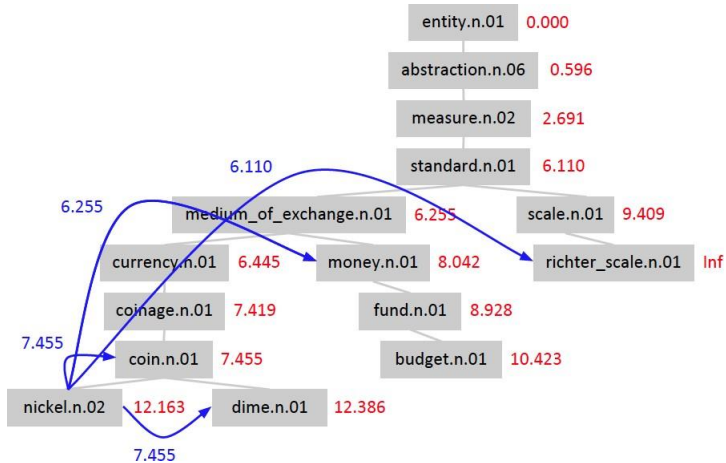
Lin similarity

Proportion of shared information

- It's not just about commonalities - it's also about differences!
- **Resnik:** The more information content they share, the more similar they are
- **Lin:** The more information content they don't share, the less similar they are
- Not the *absolute* quantity of shared information but the *proportion* of shared information

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

The information content common to c_1 and c_2, normalized by their average information content.

Example: Lin similarity

entity.n.01   0.000

abstraction.n.06   0.596

measure.n.02   2.691

standard.n.01   6.110

0.000

0.619

medium_of_exchange.n.01   6.255

scale.n.01   9.409

currency.n.01   6.445   money.n.01   8.042   richter_scale.n.01   Inf

coinage.n.01   7.419   fund.n.01   8.928

0.760

coin.n.01   7.455   budget.n.01   10.423

nickel.n.02   12.163   dime.n.01   12.386

0.607
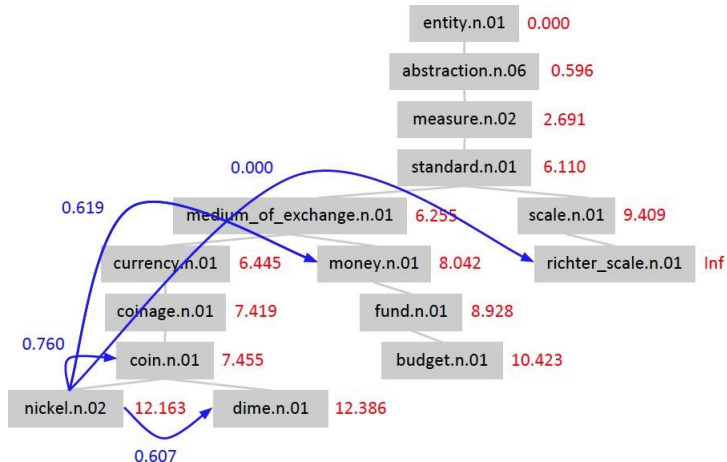
Jiang-Conrath distance

JC similarity

We can use IC to assign lengths to graph edges:

$$dist_{JC}(c, hypernym(c)) = IC(c) - IC(hypernym(c))$$

$$
\begin{aligned}
dist_{JC}(c_1, c_2) &= dist_{JC}(c_1, LCS(c_1, c_2)) + dist_{JC}(c_2, LCS(c_1, c_2)) \\
&= IC(c_1) - IC(LCS(c_1, c_2)) + IC(c_2) - IC(LCS(c_1, c_2)) \\
&= IC(c_1) + IC(c_2) - 2 \times IC(LCS(c_1, c_2))
\end{aligned}
$$

$$sim_{JC}(c_1, c_2) = \frac{1}{IC(c_1) + IC(c_2) - 2 \times IC(LCS(c_1, c_2))}$$
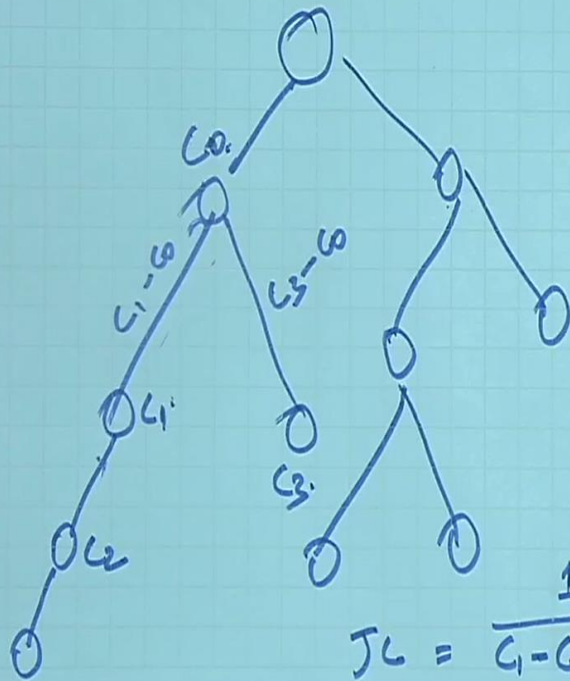
Resnic

$$Sim(c_1, c_3) = IC(c_0)$$

$$Lin = \frac{2 \cdot IC(c_0)}{IC(c_1) + IC(c_3)}$$

$$JC = \frac{1}{c_1 - c_0 + c_3 - c_0} \qquad IC(c_1) + IC(c_3)$$

Example: Jiang-Conrath distance
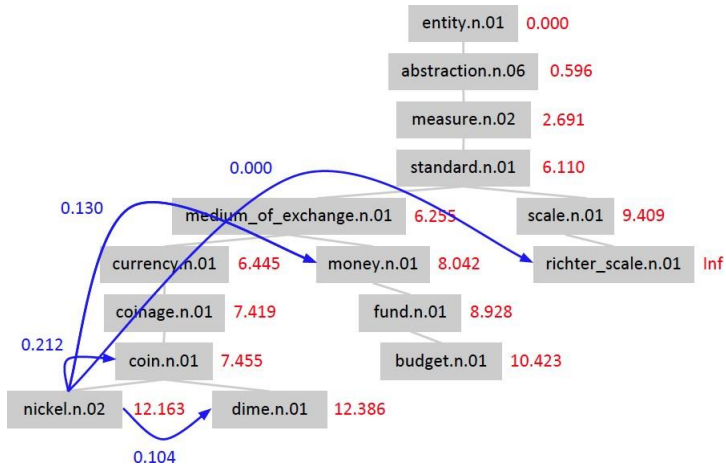
Example:

Consider the hypothetical WordNet noun taxonomy with their information content as shown in fig
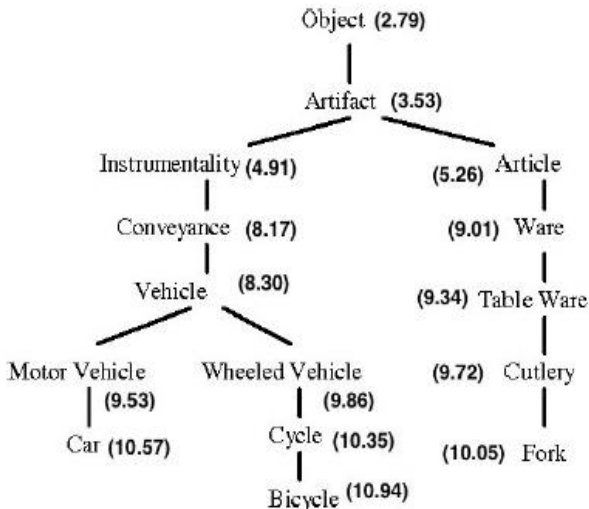
1. **What is Resnik similarity between "Bicycle" and "Fork"?**

Resnik Similarity:

$$-\log P(LCS(c_1, c_2))$$

   Ans: 3.53

2. **What is the Lin Similarity between "Vehicle" and "Table ware"?**

 Lin Similarity:

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

   = (2*3.53)/(8.30 + 9.34) = 0.40

3. **What is the Jiang – Conrath distance between "Car" and "Article"?**

$$\left(IC(c_1) + IC(c_2) - 2 \times IC(LCS(c_1, c_2)) \right)^{-1}$$

Ans: 1/((10.57+5.26) − 2*3.53) = 0.114

The (extended) Lesk Algorithm

- 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*

The (extended) Lesk Algorithm

- 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 occurs in both glosses, add a score of n^2

The (extended) Lesk Algorithm

- 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 occurs in both glosses, add a score of n^2
- **paper** and **specially prepared** → $1 + 4 = 5$

A wordnet of Indian languages

http://www.cfilt.iitb.ac.in/indowordnet/