Natural Language Processing
SoSe 2014

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Lexical Semantics, Word Sense Disambiguation and Word Similarity

Dr. Mariana Neves

May 21th, 2014

(based on the slides of Dr. Saeedeh Momtazi)

Outline

- Lexical Semantics
  - WordNet
- Word Sense Disambiguation
- Word Similarity

Outline

- Lexical Semantics
  - WordNet
- Word Sense Disambiguation
- Word Similarity

Word Meaning

- Considering the meaning(s) of a word in addition to its written form
- Word Sense
  - A discrete representation of an aspect of the meaning of a word

Berlin, Maryland

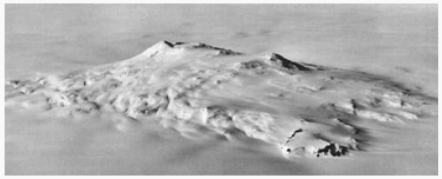
Town

Downtown Berlin, Maryland

Berlin

Berlin, 1982. L-R: David Diamond, Rob Brill, Terri Nunn, John Crawford, Matt Reid, and Ric Olsen.

Mount Berlin



Aerial view of Mount Berlin from the northwest

Elevation	3,478 m (11,411 ft)
Location	
Location	Marie Byrd Land, Antarctica

Berlin

State of Germany

Word

- Lexeme
  - An entry in a lexicon consisting of a pair: a form with a single meaning representation
    - Berlin (Germany's capital)
    - Berlin (music band)
- Lemma
  - The grammatical form that is used to represent a lexeme
    - Berlin

Homonymy

- Words which have similar form but different meanings
  - Homographs:
    - Berlin (Germany's capital)
    - Berlin (music band)

- Homophones
  - write
  - right

Semantics Relations

- Realizing lexical relations among words (senses)
  - Hyponymy (is a) {parent: hypernym, child: hyponym}
    - dog & animal
  - Meronymy (part of)
    - arm & body
  - Synonymy
    - fall & autumn
  - Antonymy
    - tall & short

Outline

- Lexical Semantics
  - WordNet
- Word Sense Disambiguation
- Word Similarity

WordNet

- A hierarchical database of lexical relations
- Three Separate sub-databases
  - Nouns
  - Verbs
  - Adjectives and Adverbs
- Closed class words are not included
- Each word is annotated with a set of senses
- Available online or for download
  - http://wordnetweb.princeton.edu/perl/webwn

WordNet 3.0

Number of words, synsets, and senses

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

PRINCETON UNIVERSITY

WordNet
A lexical database for English

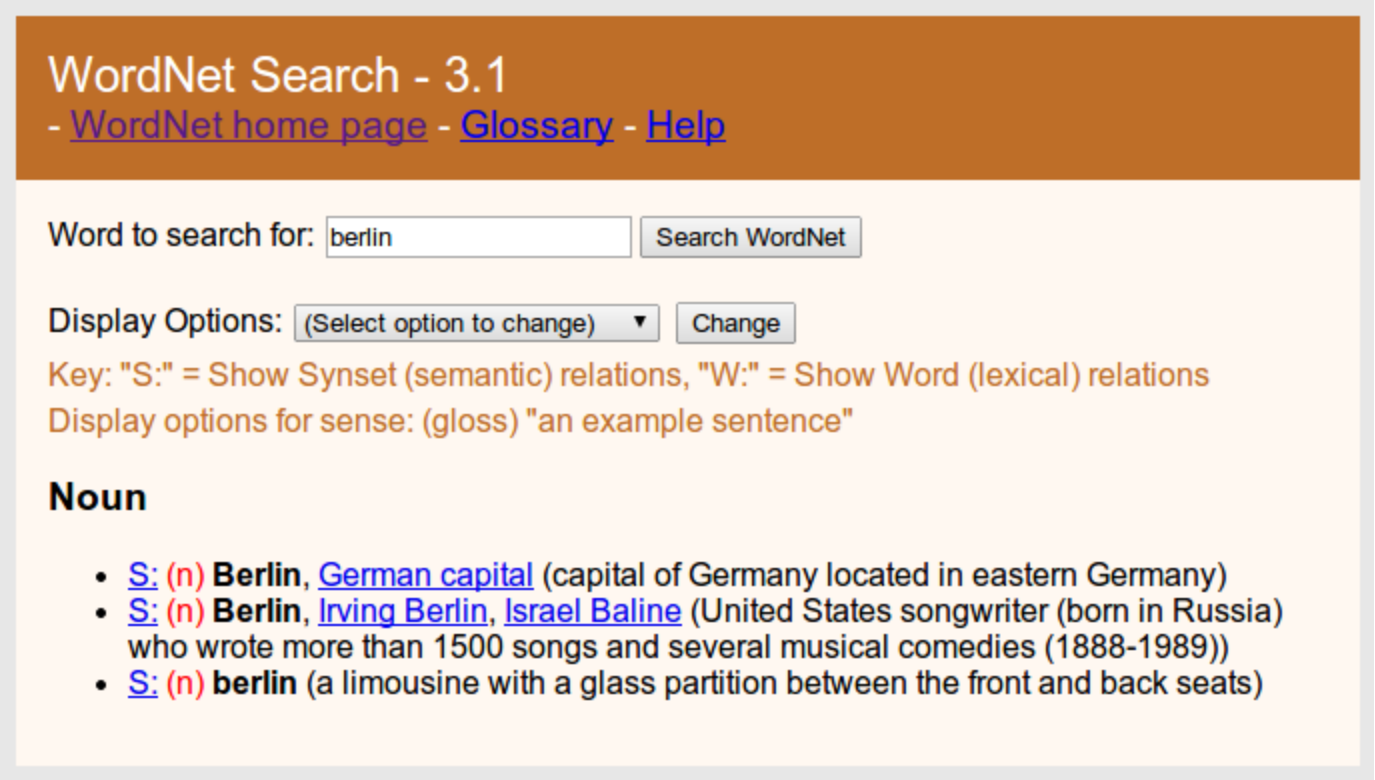
Polysemy information

POS	Average Polysemy Including Monosemous Words	Average Polysemy Excluding Monosemous Words
Noun	1.24	2.79
Verb	2.17	3.57
Adjective	1.40	2.71
Adverb	1.25	2.50

http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

Word sense

- Synset (synonym set)



Word sense

Noun

- S: (n) set, circle, **band**, lot (an unofficial association of people or groups) *"the smart set goes there"; "they were an angry lot"*
- S: (n) **band** (instrumentalists not including string players)
- S: (n) **band**, banding, stria, striation (a stripe or stripes of contrasting color) *"chromosomes exhibit characteristic bands"; "the black and yellow banding of bees and wasps"*
- S: (n) **band**, banding, stripe (an adornment consisting of a strip of a contrasting color or material)
- S: (n) dance band, **band**, dance orchestra (a group of musicians playing popular music for dancing)
- S: (n) **band** (a range of frequencies between two limits)
- S: (n) **band** (a thin flat strip of flexible material that is worn around the body or one of the limbs (especially to decorate the body))
- S: (n) isthmus, **band** (a cord-like tissue connecting two larger parts of an anatomical structure)
- S: (n) ring, **band** (jewelry consisting of a circlet of precious metal (often set with jewels) worn on the finger) *"she had rings on every finger"; "he noted that she wore a wedding band"*
- S: (n) **band** (a driving belt in machinery)
- S: (n) **band** (a thin flat strip or loop of flexible material that goes around or over something else, typically to hold it together or as a decoration)
- S: (n) **band**, ring (a strip of material attached to the leg of a bird to identify it (as in studies of bird migration))
- S: (n) **band** (a restraint put around something to hold it together)

Verb

- S: (v) **band** (bind or tie together, as with a band)
- S: (v) ring, **band** (attach a ring to the foot of, in order to identify) *"ring birds"; "band the geese to observe their migratory patterns"*

Word Relations (Hypernym)

- S: (n) ring, **band** (jewelry consisting of a circlet of precious metal (often set with jewels) worn on the finger) *"she had rings on every finger"; "he noted that she wore a wedding band"*
  - *direct hyponym* / *full hyponym*
    - S: (n) engagement ring (a ring given and worn as a sign of betrothal)
    - S: (n) mourning ring (a ring worn as a memorial to a dead person)
    - S: (n) ringlet (a small ring)
    - S: (n) signet ring, seal ring (a ring bearing a signet)
    - S: (n) wedding ring, wedding band (a ring (usually plain gold) given to the bride (and sometimes one is also given to the groom) at the wedding)
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (n) jewelry, jewellery (an adornment (as a bracelet or ring or necklace) made of precious metals and set with gems (or imitation gems))
  - ***derivationally related form***
    - W: (v) ring [Related to: ring] (attach a ring to the foot of, in order to identify) *"ring birds"; "band the geese to observe their migratory patterns"*

Word Relations (Sister)

- S: (n) set, circle, **band**, lot (an unofficial association of people or groups) *"the smart set goes there"; "they were an angry lot"*
  - *direct hyponym* / *full hyponym*
  - *direct hypernym* / *inherited hypernym* / ***sister term***
    - S: (n) social group (people sharing some social relation)
      - S: (n) body (a group of persons associated by some common tie or occupation and regarded as an entity) *"the whole body filed out of the auditorium"; "the student body"; "administrative body"*
      - S: (n) society (an extended social group having a distinctive cultural and economic organization)
      - S: (n) minority (a group of people who differ racially or politically from a larger group of which it is a part)
      - S: (n) sector (a social group that forms part of the society or the economy) *"the public sector"*
      - S: (n) interest, interest group ((usually plural) a social group whose members control some field of activity and who have common aims) *"the iron interests stepped up production"*
      - S: (n) kin, kin group, kinship group, kindred, clan, tribe (group of people related by blood or marriage)
      - S: (n) kith (your friends and acquaintances) *"all his kith and kin"*
      - S: (n) fringe (a social group holding marginal or extreme views) *"members of the fringe believe we should be armed with guns at all times"*
      - S: (n) gathering, assemblage (a group of persons together in one place)
      - S: (n) congregation, fold, faithful (a group of people who adhere to a common faith and habitually attend a given church)
      - S: (n) organization, organisation (a group of people who work together)
      - S: (n) phylum ((linguistics) a large group of languages that are historically related)
      - S: (n) force (a group of people having the power of effective action) *"he joined forces with a band of adventurers"*
      - S: (n) platoon (a group of persons who are engaged in a common activity) *"platoons of tourists poured out of the busses"; "the defensive platoon of the football team"*
      - S: (n) revolving door (an organization or institution with a high rate of turnover of personnel or membership)

Outline

- Lexical Semantics
  - WordNet
- Word Sense Disambiguation
- Word Similarity

Applications

- Information retrieval
- Machine translation
- Speech synthesis

Information retrieval

Berlin is the capital of Germany.

Berlin may also refer to:

Individuals [edit]

- Berlin (surname)
- Berlin Ndebe-Nlome (born 1987), Cameroonian football player
- Berlin, former stage name for professional wrestler Alex Wright

Places [edit]

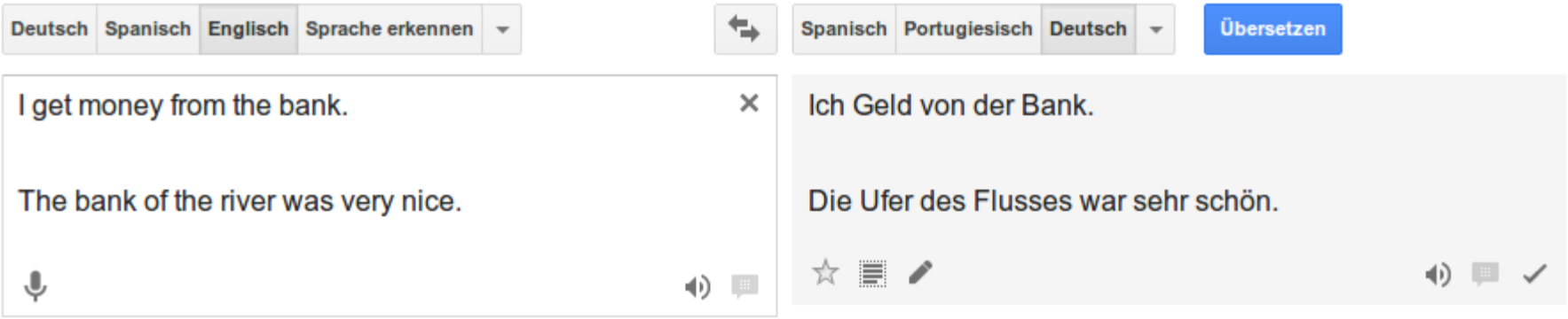
Canada [edit]

- Berlin, former name of Kitchener, Ontario
  - Berlin to Kitchener name change

United States [edit]

- Berlin, California, the former name of Genevra, California
- Berlin, Connecticut
  - Berlin (Amtrak station), rail station in Berlin, Connecticut
- Berlin, Georgia
- Berlin, Illinois
- Berlin, Indiana, extinct town
- Berlin, Kentucky
- Berlin, Maryland

Machine translation



Word Sense Disambiguation

- Input
  - A word
  - The context of the word
  - Set of potential senses for the word

- Output
  - The best sense of the word for this context

Approaches

- Thesaurus-based
- Supervised learning
- Semi-supervised learning

Thesaurus-based

- Extracting sense definitions from existing sources
  - Dictionaries
  - Thesauri
  - Wikipedia

Thesaurus-based

WIKIPEDIA
The Free Encyclopedia

Science and technology [edit]

- BAND (application), a private space for groups
- Band (mathematics), an idempotent semigroup
- Band (radio), a range of frequencies or wavelengths used in radio transmission and radar
- Band cell, a type of white blood cell
- Gastric band, a weight-control measure
- Bird banding, placing numbered bands of metal on birds' legs for identification

Organizations [edit]

- Band (channel), nickname of Brazilian broadcast television network Rede Bandeirantes
- Bands (Italian Army irregulars), military units once in the service of the Italian Regio Esercito
- The Band (professional wrestling), the Total Nonstop Wrestling name for the professional wrestling stable New World Order

Music [edit]

- Band (music), a group of people who perform instrumental or vocal music
  - Concert band, an ensemble of woodwind, brass, and percussion instruments
  - School band, a group of student musicians who rehearse and perform instrumental music together
  - Marching band, a group of instrumental musicians who generally perform outdoors incorporating some type of marching
  - Jazz band, a musical ensemble that plays jazz music
- The Band, a Canadian-American rock and roll group
  - *The Band* (album), its eponymous album released in 1969

Clothing, jewelry, and accessories [edit]

- Bands (neckwear), two pieces of cloth fitted around the neck as part of formal clothing for clergy, academics, and lawyers
- Bandolier or bandoleer, an ammunition belt
- Wedding band, a metal ring indicating the wearer is married
- Belt (clothing), a flexible band or strap, typically made of leather or heavy cloth, and worn around the waist
- Strap, an elongated flap or ribbon, usually of fabric or leather

Results: 1 to 20 of 7561

<< First < Prev Page 1 of 379 Next > Last >>

Name/Gene ID	Description	Location	Aliases	MIM
p53 ID: 2768677	CG33336 gene product from transcript CG33336-RB [*Drosophila melanogaster* (fruit fly)]	Chromosome 3R, NT_033777.2 (18875379..18879804, complement)	Dmel_CG33336, CG10873, CG31325, CG33336, D-p53, DMP53, Dm-P53, DmP53, Dmel\CG33336, Dmp53, Dp53, dmp53, dp53, prac	
TP53 ID: 7157	tumor protein p53 [*Homo sapiens* (human)]	Chromosome 17, NC_000017.11 (7668402..7687550, complement)	BCC7, LFS1, P53, TRP53, TP53	191170
Trp53 ID: 22059	transformation related protein 53 [*Mus musculus* (house mouse)]	Chromosome 11, NC_000077.6 (69580359..69591873)	RP23-56I20.1, Tp53, bbl, bfy, bhy, p44, p53, Trp53	
Tp53 ID: 24842	tumor protein p53 [*Rattus norvegicus* (Norway rat)]	Chromosome 10, NC_005109.3 (55932658..55944087)	Trp53, p53, Tp53	
p53 ID: 100384887	p53 tumor suppressor homolog [*Bombyx mori* (domestic silkworm)]	NW_004581688.1 (1491354..1507674, complement)		
p53 ID: 42722	p53 gene product [*Drosophila melanogaster* (fruit fly)] ⚠ **discontinued**	Chromosome 3R, NT_033777 (18866029..18869867, complement)	CG10873, Dmp53, dp53	
p53 ID: 100702587	cellular tumor antigen p53-like [*Oreochromis niloticus* (Nile tilapia)]	NT_167699.1 (309812..316551)		
P53 ID: 7020953	hypothetical protein [*Bacteriophage APSE-2*]	NC_011551.1 (38386..39303, complement)	APSE242	

http://www.ncbi.nlm.nih.gov/gene/?term=p53

The Lesk Algorithm

- Selecting the sense whose definition shares the most words with the word's context

```
function SIMPLIFIED LESK(word,sentence) returns best sense of word
    best-sense <- most frequent sense for word
    max-overlap <- 0
    context <- set of words in sentence
    for each sense in senses of word do
        signature <- set of words in the gloss and examples of sense
        overlap <- COMPUTEOVERLAP (signature,context)
        if overlap > max-overlap then
                max-overlap <- overlap
                best-sense <- sense
end return (best-sense)
```

http://en.wikipedia.org/wiki/Lesk_algorithm

The Lesk Algorithm

- Simple to implement
- No training data needed
- Relatively bad results

Supervised Learning

- Training data:
  - A corpus in which each occurrence of the ambiguous word w is annotated by its correct sense
    - SemCor : 234,000 sense-tagged from Brown corpus
    - SENSEVAL-1: 34 target words
    - SENSEVAL-2: 73 target words
    - SENSEVAL-3: 57 target words (2081 sense-tagged)

Feature Selection

- Using the words in the context with a specific window size
  - Collocation
    - Considering all words in a window (as well as their POS) and their position
  - Bag-of-word
    - Considering the frequent words regardless their position
    - Deriving a set of k most frequent words in the window from the training corpus
    - Representing each word in the data as a k -dimension vector
    - Finding the frequency of the selected words in the context of the current observation

Collocation

- band
  - There would be equal access to all currencies financial instruments and financial services dash and no major constitutional change. As realignments become more rare and exchange rates waver in narrower bands the system could evolve into one of fixed exchange rates.
- Window size: +/- 3
- Context: waver in narrower bands the system could
- $\{W_{n-3}, P_{n-3}, W_{n-2}, P_{n-2}, W_{n-1}, P_{n-1}, W_{n+1}, P_{n+1}, W_{n+2}, P_{n+2}, W_{n+3}, P_{n+3}\}$
- {waver, NN, in , IN , narrower, JJ, the, DT, system, NN , could, MD}

Bag-of-words

- band
  - There would be equal access to all currencies financial instruments and financial services dash and no major constitutional change. As realignments become more rare and exchange rates waver in narrower bands the system could evolve into one of fixed exchange rates.
- Window size: +/- 3
- Context: waver in narrower bands the system could
- k frequent words for band:
  - {circle, dance, group, jewelery, music, narrow, ring, rubber, wave}
  - { 0 , 0 , 0 , 0 , 0 , 1 , 0 , 0 , 1 }

Naïve Bayes Classification

- Choosing the best sense ŝ out of all possible senses s_i for a feature vector $\vec{f}$ of the word w

$$\hat{s} = argmax_{s_i} P(s_i | \vec{f})$$

$$\hat{s} = argmax_{s_i} \frac{P(\vec{f} | s_i) P(s_i)}{P(\vec{f})}$$

$$P(\vec{f}) \, has \, no \, effect$$

$$\hat{s} = argmax_{s_i} P(\vec{f} | s_i) P(s_i)$$

Naïve Bayes Classification

$$\hat{s} = argmax_{s_i} P(\vec{f} | s_i) P(s_i)$$

Likelihood probability

Prior probability

$$\hat{s} = argmax_{s_i} P(s_i) \prod_{j=1}^{m} P(f_j | s_i)$$

$$P(s_i) = \frac{\#(s_i)}{\#(w)}$$

#(s_i): number of times the sense s_i is used for the word w in the training data

#(w): the total number of samples for the word w

Naïve Bayes Classification

$$\hat{s} = argmax_{s_i} P(\vec{f}|s_i) P(s_i)$$

Likelihood probability

Prior probability

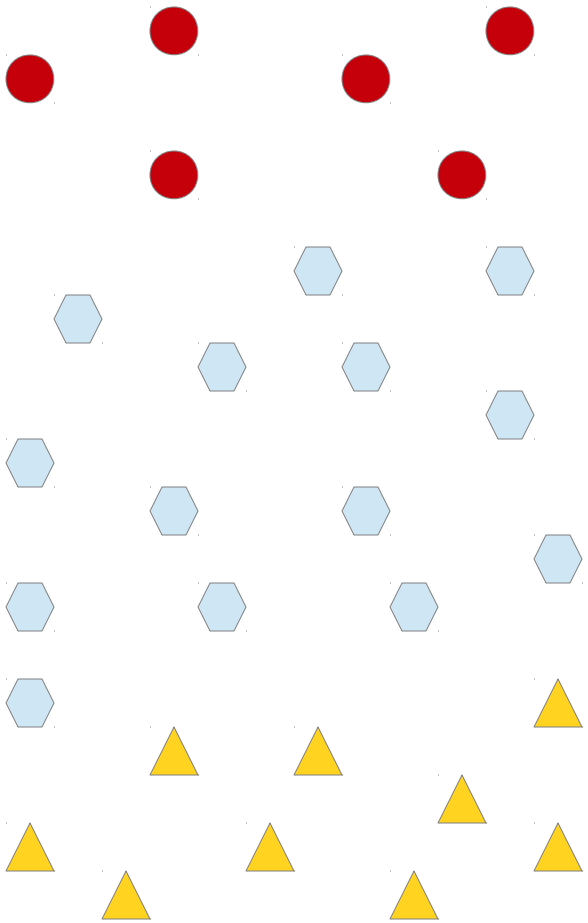
$$\hat{s} = argmax_{s_i} P(s_i) \prod_{j=1}^{m} P(f_j|s_i)$$

$$P(f_j|s_i) = \frac{\#(f_j, s_i)}{\# s_i}$$

$\#(f_j,s_i)$: the number of times the feature f_j occurred for the sense s_i of word w

$\#(s_i)$: the total number of samples of w with the sense s_i in the training data

Semi-supervised Learning



- A small amount of labeled data
- A large amount of unlabeled data

- Solution
- Finding the similarity between the labeled and unlabeled data
- Predicting the labels of the unlabeled data

Semi-supervised Learning

- For each sense,
  - Select the most important word which frequently co-occurs with the target word only for this particular sense
  - Find the sentences from unlabeled data which contain the target word and the selected word
  - Label the sentence with the corresponding sense
  - Add the new labeled sentences to the training data

Semi-supervised Learning

- Example for „band"
  - „play" (music)
  - „elastic" (rubber)
  - „spectrum" (range)

Outline

- Lexical Semantics
  - WordNet
- Word Sense Disambiguation
- Word Similarity

Word similarity

- Task
  - Finding the similarity between two words
  - Covering somewhat a wider range of relations in the meaning (e.g., relatedness)
  - Different with synonymy
  - Being defined with a score (degree of similarity)
- Example
  - Bank (financial institute) & fund
  - car & bicycle; car & gasoline

Applications

- Information retrieval
- Question answering
- Document categorization
- Machine translation
- Language modeling
- Word clustering

Information retrieval & Question Answering

Google

when was the first vehicle invented

Web Images Shopping News Videos More Search tools

About 8,370,000 results (0.33 seconds)

Who invented the automobile? (Everyday Mysteries: Fun ...
www.loc.gov › Researchers
If we had to give credit to one **inventor**, it would probably be Karl Benz from Germany. Many suggest that he **created** the **first** true **automobile** in 1885/1886.

History of the automobile - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**History**_of_the_**automobile**
The **first** carriage-sized **automobile** suitable for use on existing wagon roads in the United States was a steam powered **vehicle invented** in 1871, by Dr. J.W. ...
François Isaac de Rivaz - Timeline of motor vehicle brands

Automobile - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Automobile**
Jump to **History** - The **first** working steam-powered **vehicle** was designed — and most likely ... Coincidentally, in 1807 the Swiss **inventor** François Isaac de ...
History of the automobile - Car (disambiguation) - Ferdinand Verbiest - Motor vehicle

Who invented the world's very first car? - io9
io9.com/5816040/who-**invented**-the-worlds-very-**first-car**
Jun 28, 2011 - Who **invented** the **first car**? If we're talking about the **first** modern **automobile**, then it's Karl Benz in 1886. But long before him, there were ...

What Year Was The First Car Made? - TheOS.IN
theos.in/technology/what-**year-was-the-first-car**-made/
May 28, 2007 - This is considered as the **year** when the **first car** was made. .... No ford didn't make the **first car** he **invented** the **first** assembly line. Reply.

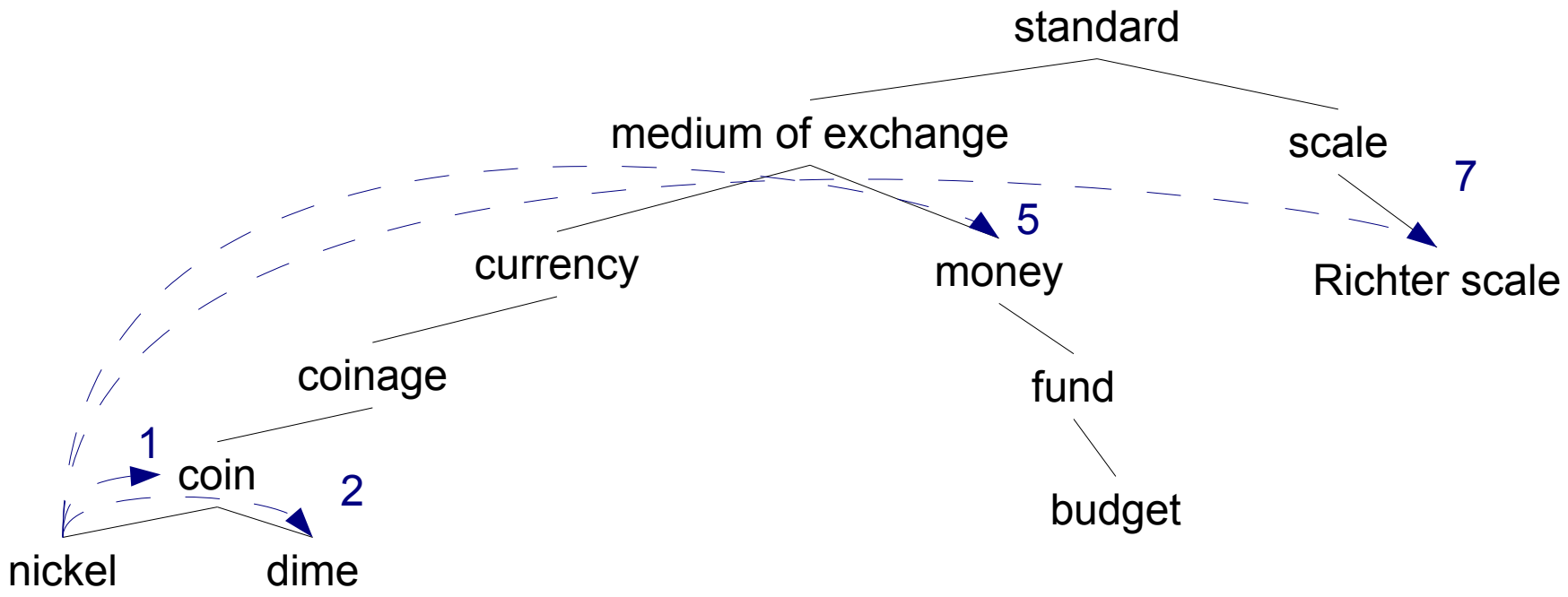
Approaches

- Thesaurus-based
  - Based on their distance in a thesaurus
  - Based on their definition in a thesaurus (gloss)

- Distributional
  - Based on the similarity between their contexts

Thesaurus-based Methods

- Two concepts (sense) are similar if they are "nearby" (if there is a short path between them in the hypernym hierarchy)



Path-base Similarity

- $\text{pathlen}(c_1,c_2) = 1 + \text{number of edges in the shortest path between the sense nodes } c_1 \text{ and } c_2$

- $\text{sim}_{path}(c_1,c_2) = -\log \text{pathlen}(c_1,c_2)$

- $\text{wordsim}(w_1,wc_2) = \max_{c1 \in \text{senses}(w1),\ c2 \in \text{senses}(w2)} \text{sim}(c_1,c_2)$

  when we have no knowledge about the exact sense

  (which is the case when processing general text)

Path-base Similarity

- Shortcoming
  - Assumes that each link represents a uniform distance
    - „nickel" to „money" seems closer than „nickel" to „standard"

- Solution
  - Using a metric which represents the cost of each edge independently

    ⇒ Words connected only through abstract nodes are less similar

Information Content Similarity

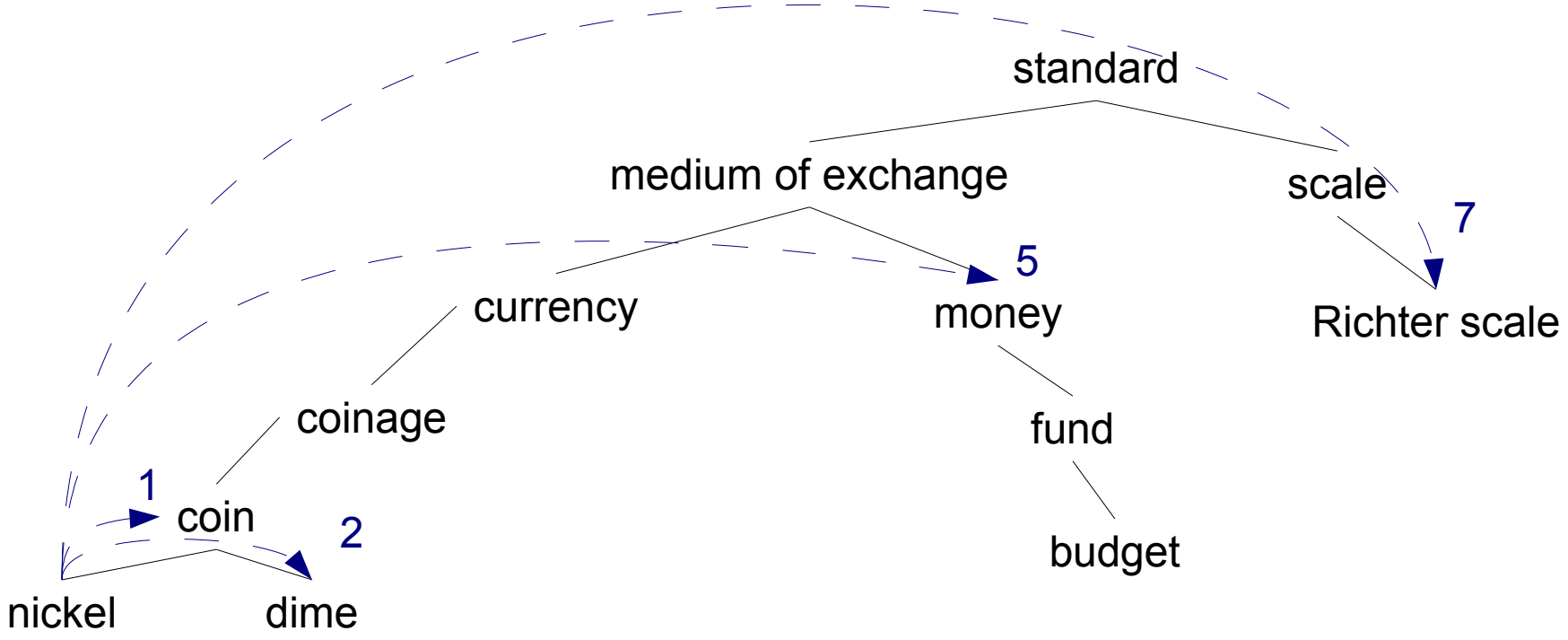
- Assigning a probability P(c) to each node of thesaurus
  - P(c) is the probability that a randomly selected word in a corpus is an instance of concept c

    ⇒ P(root) = 1, since all words are subsumed by the root concept
  - The probability is trained by counting the words in a corpus
  - The lower a concept in the hierarchy, the lower its probability

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

- words(c) is the set of words subsumed by concept c
- N is the total number of words in the corpus that are available in thesaurus

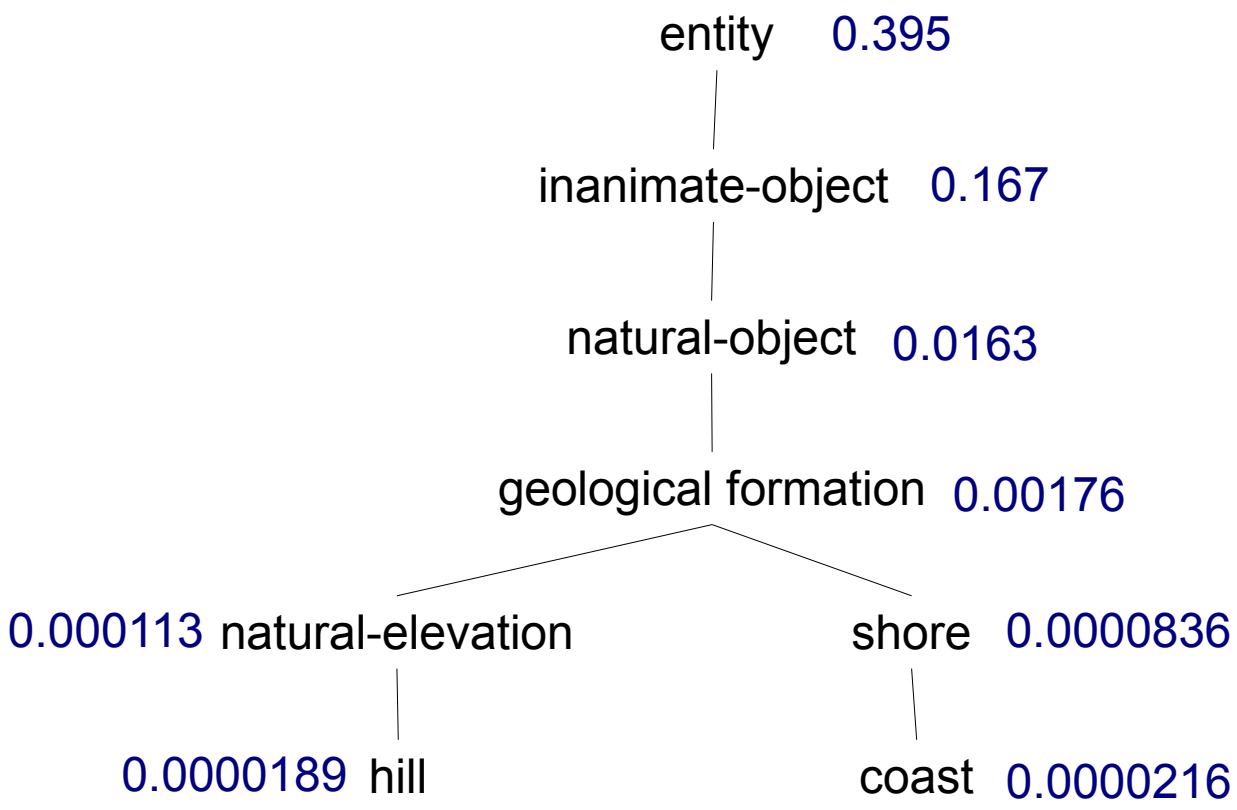
Information Content Similarity



words(coin) = {nickel, dime}
words(coinage) = {nickel, dime, coin}
words(money) = {budget, fund}
words(medium of exchange) = {nickel, dime, coin, coinage, currency, budget, fund, money}

Information Content Similarity

- Augmenting each concept in the hierarchy with a probability P(c)



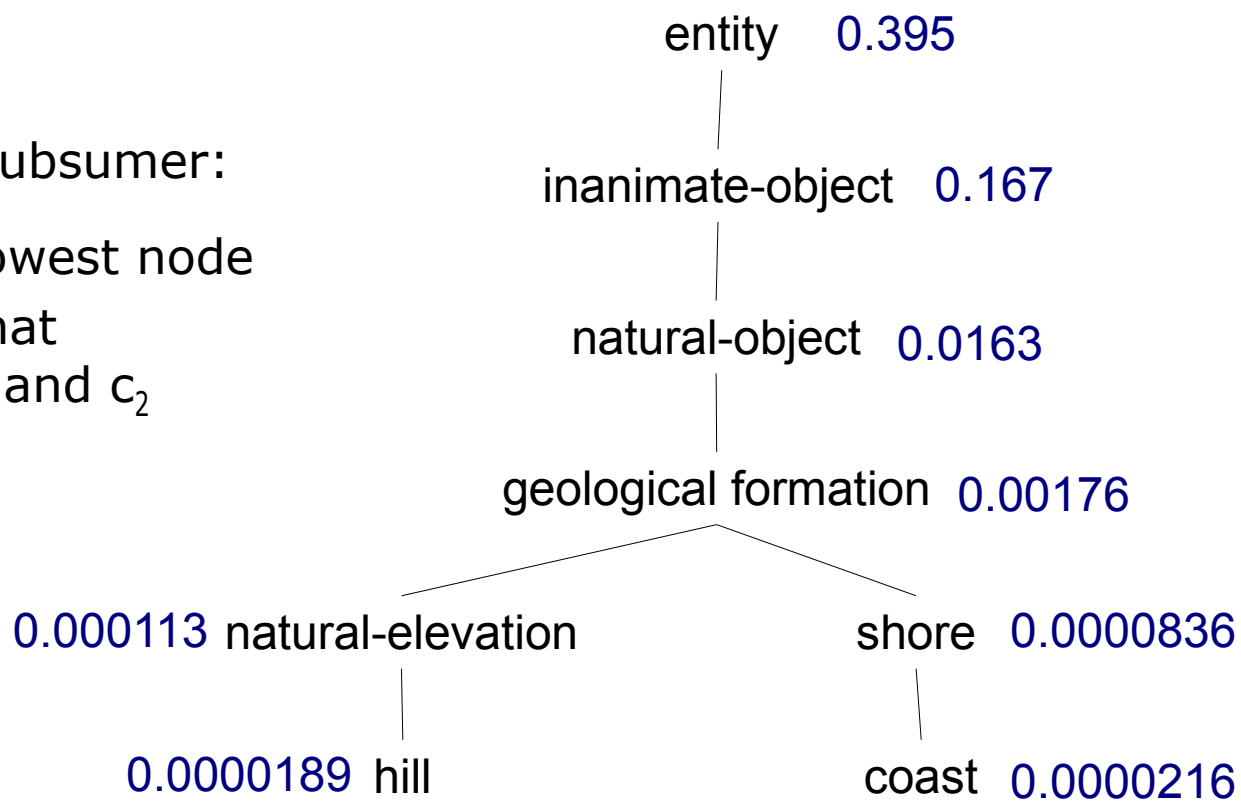
Information Content Similarity

- Information Content:

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

- Lowest common subsumer:

  $LCS(c_1,c_2)$ = the lowest node in the hierarchy that subsumes both c_1 and c_2



Information Content Similarity

- Resnik similarity
  - Measuring the common amount of information by the information content of the lowest common subsumer of the two concepts

$$sim_{resnik}(c_1,c_2) = -\log P(LCS(c_1,c_2))$$

$$sim_{resnik}(hill,coast) = -\log P(geological\text{-}formation)$$

Information Content Similarity

- Lin similarity
  - Measuring the difference between two concepts in addition to their commonality

$$similarity_{LIN}(c_1, c_2) = \frac{2 \log P(LCS(c_1, c_2))}{\log P(c_1) \log P(c_2)}$$

$$similarity_{LIN}(hill, coast) = \frac{2 \log P(geological - formation)}{\log P(hill) \log P(coast)}$$

Information Content Similarity

- Jiang-Conrath similarity

$$similarity_{JC}(c_1, c_2) = \frac{1}{\log P(c_1) + \log P(c_2) - 2 \log P(LCS(c_1, c_2))}$$

$$similarity_{LIN}(hill, coast) = \frac{2 \log P(geological - formation)}{\log P(hill) \log P(coast)}$$

Extended Lesk

- Looking at word definitions in thesaurus (gloss)
- Measuring the similarity base on the number of common words in their definition
- Adding a score of n^2 for each n-word phrase that occurs in both glosses
- Computing overlap for other relations as well (gloss of hypernyms and hyponyms)

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

Extended Lesk

- 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
- common phrases: specially prepared and paper

$$similarity_{eLesk} = 1^2 + 2^2 = 1 + 4 = 5$$

Available Libraries

- WordNet::Similarity
  - Source:
    - http://wn-similarity.sourceforge.net/
  - Web-based interface:
    - http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi

Thesaurus-based Methods

- Shortcomings
  - Many words are missing in thesaurus
  - Only use hyponym info
    - Might useful for nouns, but weak for adjectives, adverbs, and verbs
  - Many languages have no thesaurus

- Alternative
  - Using distributional methods for word similarity

Distributional Methods

- Using context information to find the similarity between words
- Guessing the meaning of a word based on its context

- tezgüino?
  - A bottle of tezgüino is on the table
  - Everybody likes tezgüino
  - Tezgüino makes you drunk
  - We make tezgüino out of corn

  ⇒ An alcoholic beverage

Context Representations

- Considering a target term t
- Building a vocabulary of M words ($\{w_1,w_2,w_3,\ldots,w_M\}$)
- Creating a vector for t with M features ($t = \{f_1,f_2,f_3,\ldots,f_M\}$)
- f_i means the number of times the word w_i occurs in the context of t
- tezgüino?
  - A bottle of tezgüino is on the table
  - Everybody likes tezgüino
  - Tezgüino makes you drunk
  - We make tezgüino out of corn
- t = tezgüino

  vocab = {book, bottle, city, drunk, like, water,...}

  t = { 0, 1 , 0 , 1 , 1 , 0 ,...}

Context Representations

- Term-term matrix
  - The number of times the context word c appear close to the term t within a window

	art	boil	data	function	large	sugar	summarize	water
apricot	0	1	0	0	1	2	0	1
pineapple	0	1	0	0	1	1	0	1
digital	0	0	1	3	1	0	1	0
information	0	0	9	1	1	0	2	0

Goal: finding a good metric that based on the vectors of these four words shows

- apricot and pineapple to be highly similar
- digital and information to be highly similar
- the other four pairing (apricot & digital, apricot & information, pineapple & digital, pineapple & information) to be less similar

Distributional similarity

- Three parameters should be specified
  - How the co-occurrence terms are defined? (what is a neighbor?)
  - How terms are weighted?
  - What vector distance metric should be used?

Distributional similarity

- How the co-occurrence terms are defined?
  - Window of k words
  - Sentence
  - Paragraph
  - Document

Distributional similarity

- How terms are weighted?
  - Binary
    - 1, if two words co-occur (no matter how often)
    - 0, otherwise
  - Frequency
    - Number of times two words co-occur with respect to the total size of the corpus

$$P(t,c)=\frac{\#(t,c)}{N}$$

  - Pointwise Mutual information
    - Number of times two words co-occur, compared with what we would expect if they were independent

$$PMI(t,c)=\log\frac{P(t,c)}{P(t)P(c)}$$

Distributional similarity

(t,c)

	art	boil	data	function	large	sugar	summarize	water
apricot	0	1	0	0	1	2	0	1
pineapple	0	1	0	0	1	1	0	1
digital	0	0	1	3	1	0	1	0
information	0	0	9	1	1	0	2	0

P(t, c) {N = 28}

	art	boil	data	function	large	sugar	summarize	water
apricot	0	0.035	0	0	0.035	0.071	0	0.035
pineapple	0	0.035	0	0	0.035	0.035	0	0.035
digital	0	0	0.035	0.107	0.035	0	0.035	0
information	0	0	0.321	0.035	0.035	0	0.071	0

Pointwise Mutual Information

	art	boil	data	function	large	sugar	summarize	water
apricot	0	0.035	0	0	0.035	0.071	0	0.035
pineapple	0	0.035	0	0	0.035	0.035	0	0.035
digital	0	0	0.035	0.107	0.035	0	0.035	0
information	0	0	0.321	0.035	0.035	0	0.071	0

P(digital, summarize) = 0.035
P(information, function) = 0.035

P(digital, summarize) = P(information, function)

PMI(digital, summarize) =?
PMI(information, function) =?

Pointwise Mutual Information

	art	boil	data	function	large	sugar	summarize	water
apricot	0	0.035	0	0	0.035	0.071	0	0.035
pineapple	0	0.035	0	0	0.035	0.035	0	0.035
digital	0	0	0.035	0.107	0.035	0	0.035	0
information	0	0	0.321	0.035	0.035	0	0.071	0

P(digital, summarize) = 0.035
P(information, function) = 0.035

P(digital) = 0.212
P(function) = 0.142

P(summarize) = 0.106
P(information) = 0.462

$$PMI(digital, summarize) = \frac{P(digital, summarize)}{P(digital) \cdot P(summarize)} = \frac{0.035}{0.212 . 0.106} = 1.557$$

$$PMI(information, function) = \frac{P(information, function)}{P(information) \cdot P(function)} = \frac{0.035}{0.462 . 0.142} = 0.533$$

P(digital, summarize) > P(information, function)

Distributional similarity

- How terms are weighted?
  - Binary
  - Frequency
  - Pointwise Mutual information

$$PMI(t,c) = \log \frac{P(t,c)}{P(t)P(c)}$$

  - t-test

$$t-test(t,c) = \frac{P(t,c) - P(t)P(c)}{\sqrt{P(t)P(c)}}$$

Distributional similarity

- What vector distance metric should be used?
  - Cosine

$$similarity_{cosine}(\vec{v}, \vec{w}) = \frac{\sum_i v_i \times w_i}{\sqrt{\sum_i v_i^2}\sqrt{\sum_i w_i^2}}$$

  - Jaccard

$$similarity_{jaccard}(\vec{v}, \vec{w}) = \frac{\sum_i min(v_i, w_i)}{\sum_i max(v_i, w_i)}$$

  - Dice

$$similarity_{dice}(\vec{v}, \vec{w}) = \frac{2 \cdot \sum_i min(v_i, w_i)}{\sum_i (v_i + w_i)}$$

Further Reading

- Speech and Language Processing
  - Chapters 19, 20