Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Natural Language Processing

Lexical Semantics
Word Sense Disambiguation and Word Similarity

Potsdam, 31 May 2012

Saeedeh Momtazi
Information Systems Group

based on the slides of the course book

Outline

1. Lexical Semantics
   - WordNet
2. Word Sense Disambiguation
3. Word Similarity

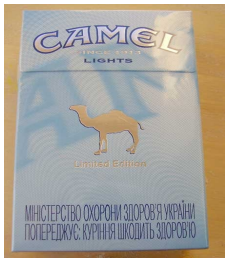
Outline

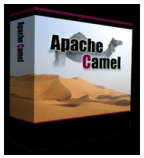
1. Lexical Semantics
   - WordNet
2. Word Sense Disambiguation
3. Word Similarity

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

Word

- Lexeme
  - An entry in a lexicon consisting of a pair:
    a form with a single meaning representation
    - Camel (animal)
    - Camel (music band)

- Lemma
  - The grammatical form that is used to represent a lexeme
    - Camel

Homonymy

- Words which have similar form but different meanings
  - Camel (animal)
  - Camel (music band)

Homographs

  - Write
  - Right

Homophone

Semantics Relations

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

Relations are between senses rather than words

Outline

1. Lexical Semantics
   - WordNet
2. Word Sense Disambiguation
3. 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
  - `http://wordnetweb.princeton.edu/perl/webwn`

WordNet

Number of words in WordNet 3.0

Category	Entry
Noun	117,097
Verb	11,488
Adjective	22,141
Adverb	4,061

Average number of senses in WordNet 3.0

Category	Sense
Noun	1.23
Verb	2.16

Word Sense

Synset (synonym set)

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: band Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

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*

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*
  - *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))
      - S: (n) adornment (a decoration of color or interest that is added to relieve plainness)
        - S: (n) decoration, ornament, ornamentation (something used to beautify)
          - S: (n) artifact, artefact (a man-made object taken as a whole)
            - 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))

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)
      - 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) organized crime, gangland, gangdom (underworld organizations)
      - S: (n) subculture (a social group within a national culture that has distinctive patterns of behavior and beliefs)

Outline

Applications

- Information retrieval
- Machine translation
- Speech synthesis

Information retrieval

camel

About 144,000,000 results (0.14 seconds)

Camel
www.**camel**.com/
R.J. Reynolds Tobacco Company only markets its tobacco products to tobacco consumers who are 21 years of age or older. In order to be eligible to receive ...
You've visited this page 2 times. Last visit: 4/16/12

Camel - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Camel**
A **camel** is an even-toed ungulate within the genus Camelus, bearing distinctive fatty deposits known as humps on its back. There are two species of **camels**: the ...
↳ Bactrian camel - Dromedary - Australian feral camel - Camel (disambiguation)

Camel (band) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Camel**_(band)
Camel are an English progressive rock band formed in 1971. Whilst they didn't achieve the large-scale fame of some of their '70s contemporaries (Pink Floyd, ...

Apache **Camel**: Index
camel.apache.org/
Apache **Camel** provides support for Bean Binding and seamless integration with popular frameworks such as Spring, Blueprint and Guice. **Camel** also has ...

Welcome to the Official **Camel** Website
www.**camel**productions.com/
Official site with news, tour information, timeline, merchandise and jukebox. Home site of founder Andy Latimer.

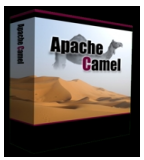
Camel Pictures and Facts
fohn.net/**camel**-pictures-facts/
A comprehensive look at **camels** and their vital role in history. Take a fun quiz, and see how much you learned! Many of the **camel** pictures are also desktop ...

San Diego Zoo's Animal Bytes: **Camel**
www.sandiegozoo.org/animalbytes/t-**camel**.html
Get accurate animal information about **camels** in an easy-to-read style from the San Diego Zoo's Animal Bytes. Buy tickets online and plan a visit to the Zoo or ...

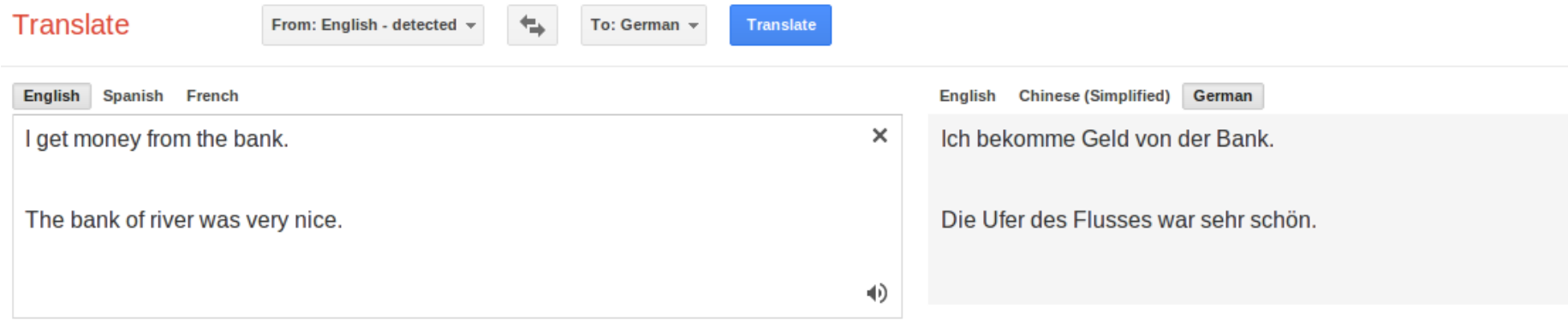
Camel – Free listening, videos, concerts, stats, & pictures at Last.fm
www.last.fm/music/**Camel**
Watch videos & listen free to **Camel**: Freefall, Supertwister & more, plus 58 pictures. **CAMEL** is a progressive rock group from Guildford, Surrey, England.

CAMEL - NU RMX UP!!! | Free Music, Tour Dates, Photos, Videos
www.myspace.com/**camel**blanco
CAMEL - NU RMX UP!!!'s official profile including the latest music, albums, songs, music videos and more updates.

Programming Perl, 3rd Edition - O'Reilly Media
shop.oreilly.com/product/9780596000271.do
Camels are large ruminant mammals, weighing between 1000 and 1600 ... All this having been said, the **Camel** Book is getting a new edition, slated (as of this ...



Machine translation



Example

Sense: band 532736 Music N

The band made copious recordings now regarded as classic from 1941 to 1950. These were to have a tremendous influence on the worldwide jazz revival to come During the war Lu led a 20 piece navy band in Hawaii.

Example

Sense: band 532838 Rubber-band N

He had assumed that so famous and distinguished a professor would have been given the best possible medical attention it was the sort of assumption young men make. Here suspended from Lewis's person were pieces of tubing held on by rubber bands an old wooden peg a bit of cork.

Example

Sense: band 532734 Range N

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.

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

Band

From Wikipedia, the free encyclopedia

Band may refer to:

Clothing, jewelry, and accessories

- 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
- Belt (clothing)
- Wedding ring or wedding band
- Strap, an elongated flap or ribbon, usually of fabric or leather

Science and technology

- Band (radio), a range of frequencies or wavelengths used in radio transmission and radar
- Rubber band, a short length of rubber and latex formed in the shape of a loop
- Möbius strip or Möbius band, an artifact with interesting topological features
- Band (mathematics), an idempotent semigroup
- Spectral bands, part of the optical spectra of polyatomic systems
- Metals and semiconductors
  - Valence band
  - Conduction band
  - Band gap

Medicine and biology

- Bird ringing, or bird banding, placing a numbered bands of metal on birds' legs for identification
- A group of animals, such as gorillas or coyotes
  - Herd
  - Flocking (behavior)
- Band cell, a type of white blood cell
- Protein band, see Coomassie
- Gastric band, a weight-control measure

Organizations

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

Society and government

- Band society, a small group of humans in a simple form of society
- The primary unit of Native Americans in the United States
- Band (First Nations Canada), the primary unit of First Nations Government in Canada

People

- Band (surname)

Places

- Band, Mureş in Romania
- Bánd, a village in Hungary

Music

- Band, a company of musicians—see Musical ensemble
  - Rock band

Rubber band

From Wikipedia, the free encyclopedia

This article is about the common household item. For other meanings, see Rubber band (disambiguation).

"Elastic band" redirects here. For the band and orchestra, see The Elastic Band. For the first aid bandage, see elastic bandage.

A **rubber band** (in some regions known as a **binder**, an **elastic** or **elastic band**, a **lackey band**, **laggy band**, **lacka band** or **gumband**) is a short length of rubber and latex formed in the shape of a loop and is commonly used to hold multiple objects together. The rubber band was patented in England on March 17, 1845 by Stephen Perry.[1][2][3]

Rubber bands in different colors and sizes.

Contents [hide]

The Lesk Algorithm

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

Simplified Algorithm [Kilgarriff and Rosenzweig, 2000]

```
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)
```

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*-dimention vector
    - Finding the frequency of the selected words in the context of the current observation

Collocation

Sense: band 532734 Range N

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-word

Sense: band 532734 Range N

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 $\hat{s}$ out of all possible senses s_i for a feature vector $\vec{f}$ of the word w

$$\hat{s} = \text{argmax}_{s_i} P(s_i|\vec{f})$$

$$\hat{s} = \text{argmax}_{s_i} \frac{P(\vec{f}|s_i) \cdot P(s_i)}{P(\vec{f})}$$

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

$$\hat{s} = \text{argmax}_{s_i} P(\vec{f}|s_i) \cdot P(s_i)$$

Naïve Bayes Classification

$$\hat{s} = \text{argmax}_{s_i} P(s_i) \cdot P(\vec{f}|s_i)$$

Prior Probability

Likelihood Probability

$$\hat{s} = \text{argmax}_{s_i} P(s_i) \cdot \prod_{j=1}^{m} P(f_j|s_i)$$

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

$\#(s_i)$: 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

$$\hat{s} = \text{argmax}_{s_i} P(s_i) \cdot P(\vec{f}|s_i)$$

Prior Probability

Likelihood Probability

$$\hat{s} = \text{argmax}_{s_i} P(s_i) \cdot \prod_{j=1}^{m} P(f_j|s_i)$$

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

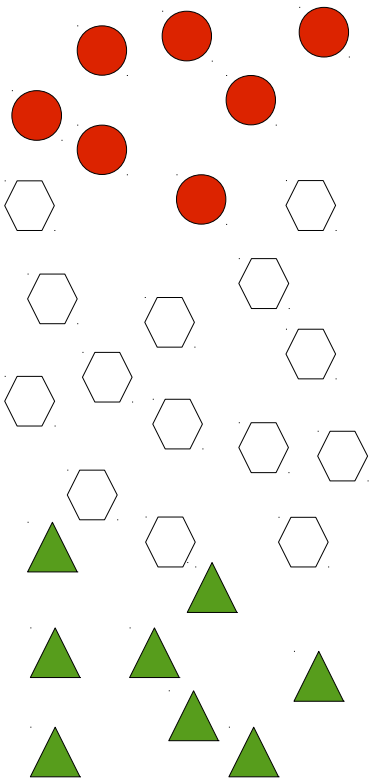
$\#(f_j, s_i)$: the number of times the feature f_j occurred for the sense s_i of word w

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

Semi-supervised Learning

- What is the best approach when we do not have enough data to train a model?

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

- What is the best approach when we do not have enough data to train a model?
  - 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
  - Example for *Band*

sense	*selected word*
Music	play
Rubber	elastic
Range	spectrum

Outline

Word Similarity

- Task
  - Finding the similarity between two words
  - Covering somewhat a wider range of relations in the meaning (different with synonymy)
  - Being defined with a score (degree of similarity)

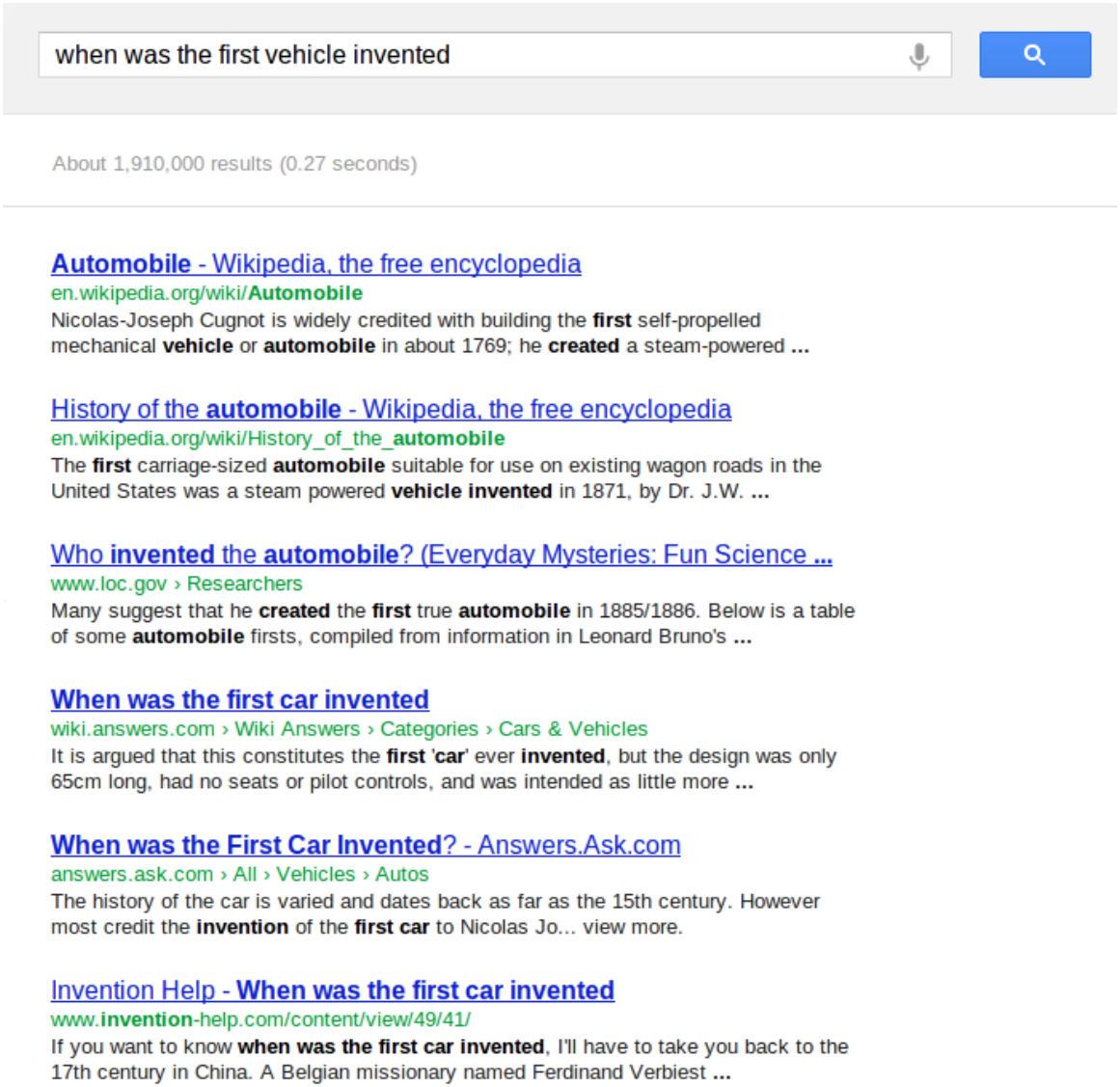
Example

Bank (financial institute) & *fund*
car & *bicycle*

Applications

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

Information retrieval & Question Answering



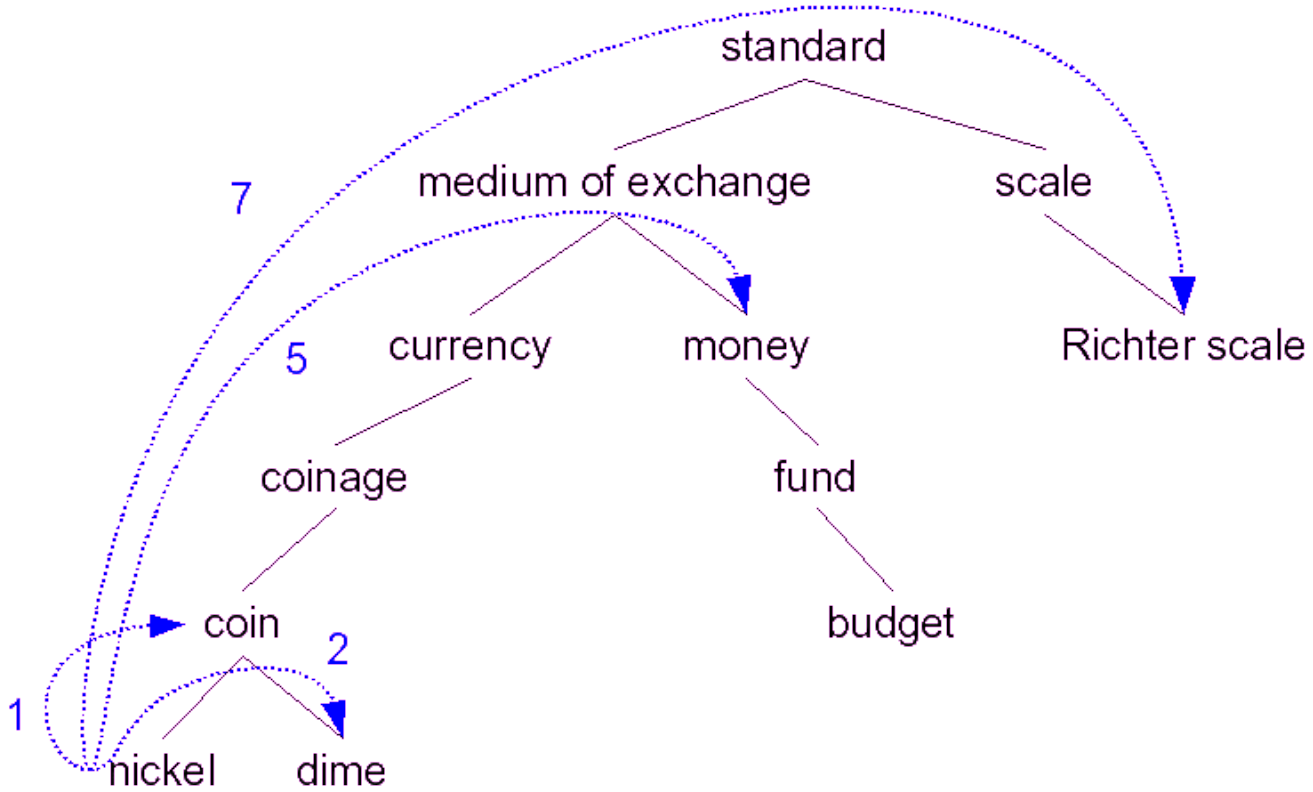
Approaches

- Thesaurus-based
  - Based on their distance in thesaurus
  - Based on their definition in 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

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

- $sim_{path}(c_1, c_2) = -\log pathlen(c_1, c_2)$

- $wordsim(w_1, w_2) = \max_{\substack{c_1 \in senses(w_1) \\ c_2 \in senses(w_2)}} sim(c_1, c_2)$

  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 to *standard*
  - Solution
    - Using a metric which represents the cost of each edge independently
      $\Rightarrow$ Words connected only through abstract nodes are less similar

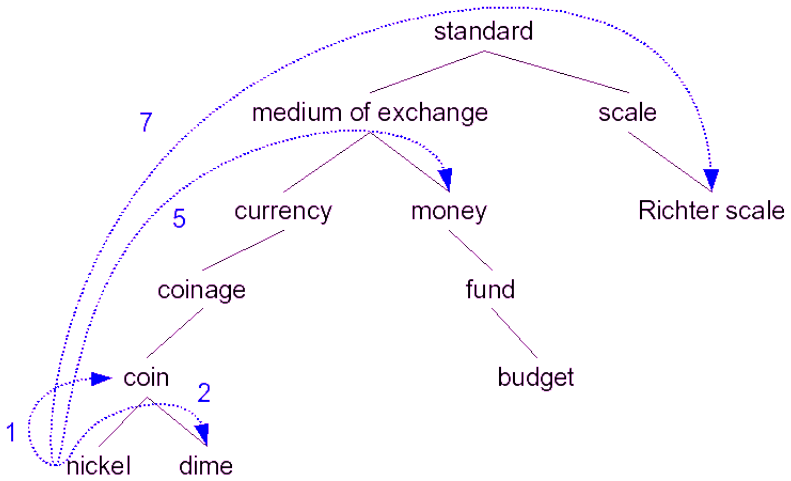
Information Content Similarity

- 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
    $\Rightarrow$ $P(root) = 1$, 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

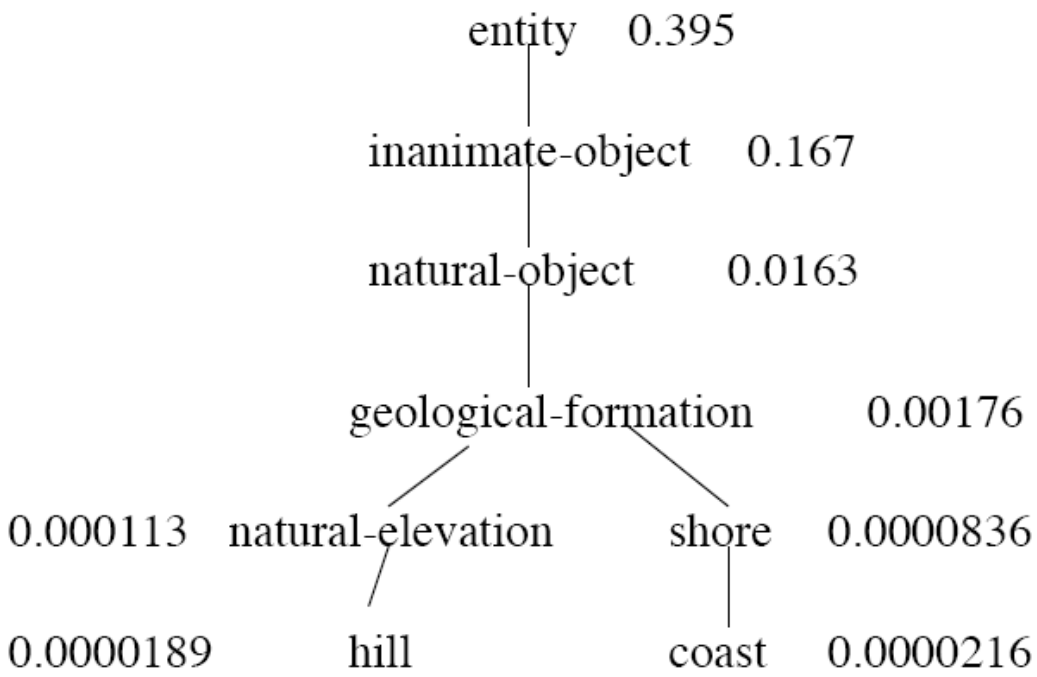


words(coin) = {nickel, dime}
words(coinage) = {nickel, dime, coin}
words(money) = {budget, fund}

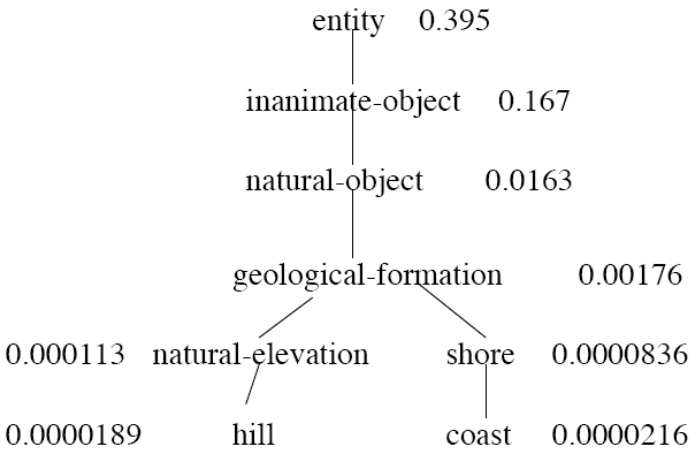
words(medium of exchange) = {nickel, dime, coin, coinage, currency, budget, fund, money}

Information Content Similarity

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



Information Content Similarity



- Information Content:
  $IC(c) = -\log P(c)$
- Lowest common subsumer:
  $LCS(c1, c2)$ = 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

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

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

Information Content Similarity

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

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

$$sim_{Lin}(\text{hill,coast}) = \frac{2 \log P(\text{geological-formation})}{\log P(\text{hill}) + P(\text{coast})}$$

Information Content Similarity

- Jiang-Conrath similarity

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

$$sim_{JC}(\text{hill,coast}) = \frac{1}{\log P(\text{hill}) + P(\text{coast}) - 2 \log P(\text{geological-formation})}$$

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)

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

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

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

Thesaurus-based Similarities

- Overview

$$\begin{aligned}
\text{sim}_{\text{path}}(c_1, c_2) &= -\log \text{pathlen}(c_1, c_2) \\
\text{sim}_{\text{Resnik}}(c_1, c_2) &= -\log P(\text{LCS}(c_1, c_2)) \\
\text{sim}_{\text{Lin}}(c_1, c_2) &= \frac{2 \times \log P(LCS(c_1, c_2))}{\log P(c_1) + \log P(c_2)} \\
\text{sim}_{\text{jc}}(c_1, c_2) &= \frac{1}{2 \times \log P(\text{LCS}(c_1, c_2)) - (\log P(c_1) + \log P(c_2))} \\
\text{sim}_{\text{eLesk}}(c_1, c_2) &= \sum_{r,q \in \text{RELS}} \text{overlap}(\text{gloss}(r(c_1)), \text{gloss}(q(c_2)))
\end{aligned}$$

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?

tezgüino?

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

$\Rightarrow$ An alcoholic beverage

Context Representations

- Considering a target term t
- Building a vocabulary of M words ($\{w_1, w_2, w_3, ..., w_M\}$)
- Creating a vector for t with M features ($t = \{f_1, f_2, f_3, ..., f_M\}$)
  - 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* in 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 **hight** similar
    - *digital* and *information* to be **hight** 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? (what is a neighbor?)
  - Widow 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

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

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

$P(\textit{digital}, \textit{summarize}) = 0.035$
$P(\textit{information}, \textit{function}) = 0.035$

$P(\textit{digital}, \textit{summarize}) = P(\textit{information}, \textit{function})$

$PMI(\textit{digital}, \textit{summarize}) =?$
$PMI(\textit{information}, \textit{function}) =?$

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

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

$P(digital) = 0.212$ $\quad$ $P(summarize) = 0.106$
$P(information) = 0.462$ $\quad$ $P(function) = 0.142$

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

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

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

Distributional similarity

- How terms are weighted?
  - Binary
  - Frequency
  - Pointwise Mutual information
    - $PMI(t,c) = log \frac{P(t,c)}{P(t) \cdot P(c)}$
  - t-test
    - $t - test(t,c) = \frac{P(t,c) - P(t) \cdot P(c)}{\sqrt{P(t) \cdot P(c)}}$

Distributional similarity

- What vector distance metric should be used?
  - Cosine
    - $Sim_{cosine}(\vec{v}, \vec{w}) = \frac{\sum_i v_i \times w_i}{\sqrt{\sum_i v_i^2}\sqrt{\sum_i w_i^2}}$
  - Jaccard
    - $Sim_{jaccard}(\vec{v}, \vec{w}) = \frac{\sum_i min(v_i, w_i)}{\sum_i max(v_i, w_i)}$
  - Dice
    - $Sim_{dice}(\vec{v}, \vec{w}) = \frac{2 \times \sum_i min(v_i, w_i)}{\sum_i (v_i + w_i)}$

Further Reading

- Speech and Language Processing
  - Chapters 19, 20