Introduction to Computational Lexical Semantics

Bill MacCartney
CS224U, Lecture 2
Stanford University
12 January 2012

[slides adapted from Dan Jurafsky]

Outline

1) Words, senses, & lexical semantic relations
2) WordNet & other resources
3) Word similarity: thesaurus-based measures
4) Word similarity: distributional measures

Three levels of meaning

1. Lexical Semantics
   - The meanings of individual words
2. Sentential / Compositional / Formal Semantics
   - How those meanings combine to make meanings for individual sentences or utterances
3. Discourse or Pragmatics
   - How those meanings combine with each other and with other facts about various kinds of context to make meanings for a text or discourse

(+ Dialog or Conversational Semantics)

The unit of meaning is a *sense*

- One word can have multiple meanings:
  - *Instead, a **bank** can hold the investments in a custodial account in the client's name.*
  - *But as agriculture burgeons on the east **bank**, the river will shrink even more.*
- We say that a **sense** is a representation of one aspect of the meaning of a word.
- Thus **bank** here has two senses
  - Bank^1:
  - Bank^2:

Some more terminology

- Lemmas and wordforms
  - A **lexeme** is an abstract pairing of meaning and form
  - A **lemma** or **citation form** is the grammatical form that is used to represent a **lexeme**.
    - ***Carpet*** is the lemma for ***carpets***
    - ***Dormir*** is the lemma for ***duermes***
  - Specific surface forms *carpets, sung, duermes* are called **wordforms**
- The lemma *bank* has two **senses:**
  - *Instead, a* ***bank*** *can hold the investments in a custodial account in the client's name.*
  - *But as agriculture burgeons on the east* ***bank****, the river will shrink even more.*
- A **sense** is a discrete representation of one aspect of the meaning of a word

Relations between word senses

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

Homonymy

- Homonyms are lexemes that share a form
  - Phonological, orthographic or both
- But have unrelated, distinct meanings
- Examples:
  - *bat* (wooden stick thing) vs *bat* (flying scary mammal)
  - *bank* (financial institution) vs *bank* (riverside)
- Can be homophones, homographs, or both:
  - Homophones: *write* and *right*, *piece* and *peace*
  - Homographs: *bass* and *bass*

Homonymy, yikes!

Homonymy causes problems for NLP applications:

- Text-to-Speech
- Information retrieval
- Machine Translation
- Speech recognition

Why might homonymy cause problems in these applications? Examples?

Polysemy

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

- Are those the same sense?
  - We might define sense 1 as: "The building belonging to a financial institution"
  - And sense 2: "A financial institution"
- Or consider the following example
  - *While some banks furnish sperm only to married women, others are less restrictive.*
  - Which sense of bank is this?

Polysemy

- We call **polysemy** the situation when a single word has multiple related meanings (bank the building, bank the financial institution, bank the biological repository)
- Most non-rare words have multiple meanings

Polysemy: A systematic relationship between senses

- Lots of types of polysemy are systematic
  - School, university, hospital, church, supermarket
  - Can all be used to mean the institution or the building
- We might say there is a relationship:
  - Building <–> Organization
- Other such kinds of systematic polysemy:

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*)

How do we know when a word has more than one sense?

- Consider examples of the word *serve*:
  - *Which flights serve breakfast?*
  - *Does America West serve Philadelphia?*
- The “zeugma” test:
  - *?Does United serve breakfast and San Jose?*
- Since this sounds weird, we say that these are **two different senses of *serve***

Synonyms

- Word that have the same meaning in some or all contexts.
  - filbert / hazelnut
  - couch / sofa
  - big / large
  - automobile / car
  - vomit / throw up
  - water / H_20
- Two lexemes are synonyms if they can be successfully substituted for each other in all situations
  - If so they have the same **propositional meaning**

Synonyms

- But there are few (or no) examples of perfect synonymy.
  - Why should that be?
  - Even if many aspects of meaning are identical
  - Still may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.
- Example:
  - Water and H_20
  - Big/large
  - Brave/courageous

Synonymy is a relation between senses rather than words

- Consider the words *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

Antonyms

- Senses that are opposites with respect to one feature of their meaning
- Otherwise, they are very 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, up/down*

Hyponymy

- One sense is a **hyponym** of another if the first is more specific, denoting a subclass of the second
  - *car* is a hyponym of *vehicle*
  - *dog* is a hyponym of *animal*
  - *mango* is a hyponym of *fruit*
- Conversely
  - *vehicle* is a hypernym/superordinate of *car*
  - *animal* is a hypernym of *dog*
  - *fruit* is a hypernym of *mango*

superordinate	vehicle	fruit	furniture	mammal
hyponym	car	mango	chair	dog

Hyponymy more formally

- Extensional:
  - The class denoted by the superordinate
  - extensionally includes the class denoted by the hyponym
- Entailment:
  - A sense A is a hyponym of sense B if being an A entails being a B
- Hyponymy is usually transitive
  - (A hypo B and B hypo C entails A hypo C)

II. WordNet

- A hierarchically organized lexical database
- On-line thesaurus + aspects of a dictionary
  - Versions for other languages are under development

Category	Unique Forms
Noun	117,097
Verb	11,488
Adjective	22,141
Adverb	4,601

WordNet

Where to find it:

http://wordnetweb.princeton.edu/perl/webwn

How is "sense" defined in WordNet?

- The set of near-synonyms for a WordNet sense is called a **synset** (**synonym set**); it's their version of a sense or a concept
- 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
- Thus for WordNet, the meaning of this sense of chump *is* this list.

Format of Wordnet Entries

The noun "bass" has 8 senses in WordNet.

1. bass^1 - (the lowest part of the musical range)
2. bass^2, bass part^1 - (the lowest part in polyphonic music)
3. bass^3, basso^1 - (an adult male singer with the lowest voice)
4. sea bass^1, bass^4 - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass^1, bass^5 - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass^6, bass voice^1, basso^2 - (the lowest adult male singing voice)
7. bass^7 - (the member with the lowest range of a family of musical instruments)
8. bass^8 - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

The adjective "bass" has 1 sense in WordNet.

1. bass^1, deep^6 - (having or denoting a low vocal or instrumental range)
   "a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"

WordNet Noun 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	$leader^1 \rightarrow follower^1$

WordNet Verb Relations

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	$increase^1 \Longleftrightarrow decrease^1$

WordNet Hierarchies

```
Sense 3
bass, basso --
(an adult male singer with the lowest voice)
=> singer, vocalist, vocalizer, vocaliser
   => musician, instrumentalist, player
      => performer, performing artist
         => entertainer
            => person, individual, someone...
               => organism, being
                  => living thing, animate thing,
                     => whole, unit
                        => object, physical object
                           => physical entity
                              => entity
               => causal agent, cause, causal agency
                  => physical entity
                     => entity

Sense 7
bass --
(the member with the lowest range of a family of
musical instruments)
=> musical instrument, instrument
   => device
      => instrumentality, instrumentation
         => artifact, artefact
            => whole, unit
               => object, physical object
                  => physical entity
                     => entity
```

Thesaurus Examples: MeSH

- **MeSH (Medical Subject Headings)**
  - organized by terms (~250,000) that correspond to medical subjects
  - for each term syntactic, morphological or semantic variants are given

MeSH Heading	Databases, Genetic
Entry Term	Genetic Databases
Entry Term	Genetic Sequence Databases
Entry Term	OMIM
Entry Term	Online Mendelian Inheritance in Man
Entry Term	Genetic Data Banks
Entry Term	Genetic Data Bases
Entry Term	Genetic Databanks
Entry Term	Genetic Information Databases
See Also	Genetic Screening

Slide from Paul Buitelaar

MeSH (Medical Subject Headings) Thesaurus

MeSH Descriptor

Definition

Neoplasms

Links

New abnormal growth of tissue. Malignant neoplasms show a greater degree of anaplasia and have the properties of invasion and metastasis, compared to benign neoplasms.

Year introduced: /diagnosis was NEOPLASM DIAGNOSIS 1964-1965

Entry Terms:

- Neoplasm
- Tumors
- Tumor
- Benign Neoplasms
- Neoplasms, Benign
- Benign Neoplasm
- Neoplasm, Benign
- Cancer
- Cancers

Synonym set

Slide from Illhoi Yoo, Xiaohua (Tony) Hu,and Il-Yeol Song

MeSH Tree

- **MeSH Ontology**
  - Hierarchically arranged from most general to most specific.
  - Actually a graph rather than a tree
    - normally appear in more than one place in the tree

MeSH Tree

- All MeSH Categories
  - Diseases Category
    - Neoplasms
      - Neoplasms by Site
        - **Digestive System Neoplasms**
          - Biliary Tract Neoplasms
            - Bile Duct Neoplasms +
            - Gallbladder Neoplasms
          - Gastrointestinal Neoplasms
            - Esophageal Neoplasms
            - Gastrointestinal Stromal Tumors
            - Intestinal Neoplasms +
            - Stomach Neoplasms
          - Liver Neoplasms
            - Adenoma, Liver Cell
            - Carcinoma, Hepatocellular
            - Liver Neoplasms, Experimental
          - Pancreatic Neoplasms
            - Adenoma, Islet Cell +
            - Carcinoma, Islet Cell +
            - Carcinoma, Pancreatic Ductal
          - Peritoneal Neoplasms

Slide from Illhoi Yoo, Xiaohua (Tony) Hu,and Il-Yeol Song

MeSH Ontology

- Solving traditional synonym/hypernym/hyponym problems in information retrieval and text mining
- Synonym problems <= Entry terms
  - E.g., Cancer and tumor are synonyms
- Hypernym/hyponym problems <= MeSH Tree
  - E.g., Melatonin is a hormone

Slide from Illhoi Yoo, Xiaohua (Tony) Hu,and Il-Yeol Song

MeSH Ontology for MEDLINE indexing

- In addition to its ontology role
- MeSH Descriptors have been used to index MEDLINE articles.
    - MEDLINE is NLM's bibliographic database
        - Over 18 million articles
        - Refs to journal articles in the life sciences with a concentration on biomedicine
- About 10 to 20 MeSH terms are manually assigned to each article (after reading full papers) by trained curators.
    - 3 to 5 MeSH terms are “MajorTopics” that primarily represent an article.

Slide from Illhoi Yoo, Xiaohua (Tony) Hu, and Il-Yeol Song

Word Similarity

- Synonymy is a binary relation
  - Two words are either synonymous or not
- We want a looser metric: word *similarity* (or 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'll compute them over both words and senses

Why word similarity?

- Information retrieval
- Question answering
- Machine translation
- Natural language generation
- Language modeling
- Automatic essay grading
- Document clustering

Two classes of algorithms

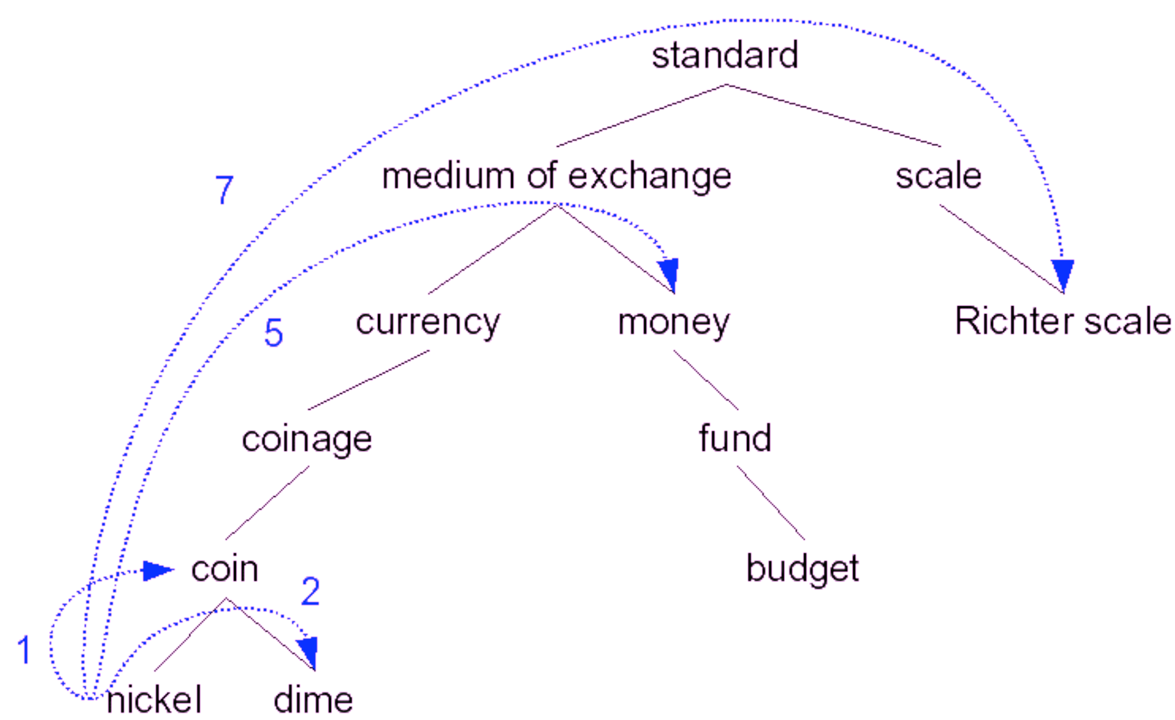
- Thesaurus-based algorithms
  - Based on whether words are "nearby" in Wordnet or MeSH
- Distributional algorithms
  - By comparing words based on their distributional context in corpora

Thesaurus-based word similarity

- We could use anything in the thesaurus:
  - Meronymy, hyponymy, troponymy
  - Glosses and example sentences
  - Derivational relations and sentence frames
- In practice, “thesaurus-based” methods usually use:
  - the is-a/subsumption/hypernym 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 not similar
    - *car*, *bicycle*: similar

Path-based similarity

Idea: two words are similar if they're nearby in the thesaurus hierarchy (i.e., short path between them)



Tweaks to path-based similarity

- $\text{pathlen}(c_1, c_2)$ = number of edges in the shortest path in the thesaurus graph between the sense nodes c_1 and c_2
- $\text{sim}_{\text{path}}(c_1, c_2) = -\log \text{pathlen}(c_1, c_2)$
- $\text{wordsim}(w_1, w_2) =$
  $\max_{c_1 \in \text{senses}(w_1),\, c_2 \in \text{senses}(w_2)} \text{sim}(c_1, c_2)$

Problems with path-based similarity

- Assumes each link represents a uniform distance
- *nickel* to *money* seems closer than *nickel* to *standard*
- Seems like we want a metric which lets us assign different "lengths" to different edges — but how?

Assigning probabilities to concepts

- Define P(c) as the probability that a randomly selected word in a corpus is an instance of concept (synset) c
- Formally: there is a distinct random variable, ranging over words, associated with each concept in the hierarchy
- P(ROOT) = 1
- The lower a node in the hierarchy, the lower its probability

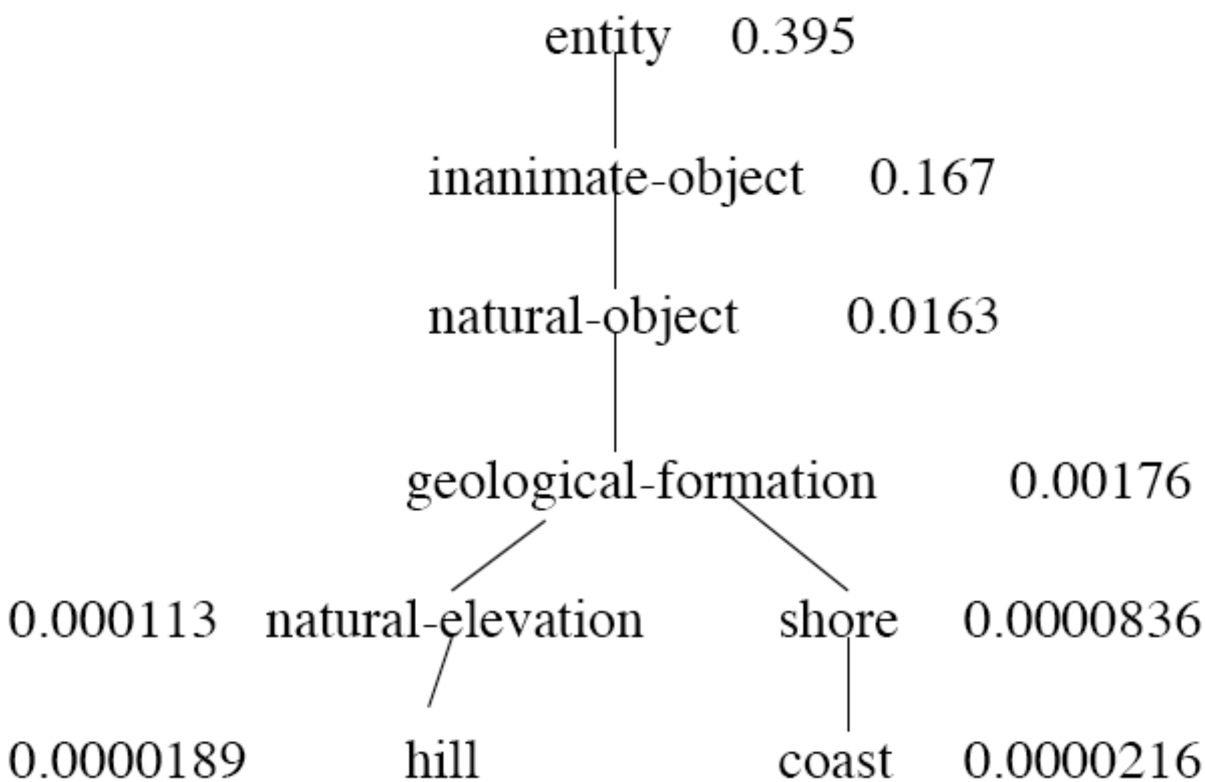
Estimating concept probabilities

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

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

Concept probability examples

WordNet hierarchy augmented with probabilities P(*c*):

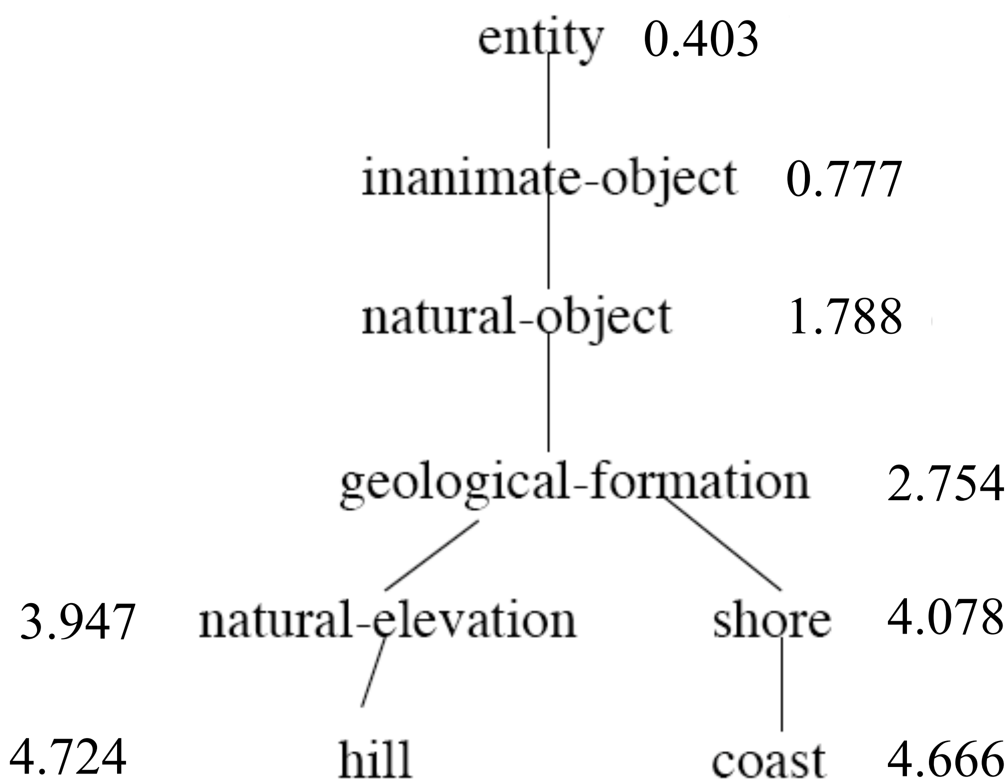


Information content: definitions

- Information content:
  - $IC(c) = -\log P(c)$
- Lowest common subsumer
  - $LCS(c_1, c_2)$ = the lowest common subsumer
    I.e., 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

Information content examples

WordNet hierarchy augmented with information contents IC(c):

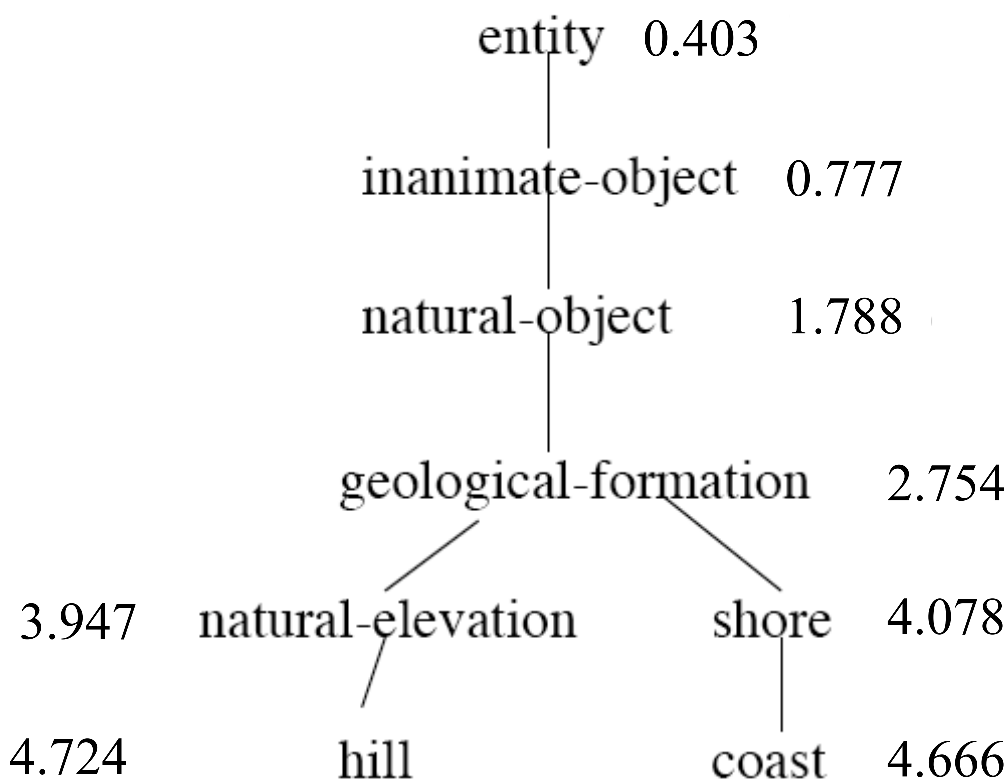


Resnik method

- The similarity between two words is related to their common information
- The more two words have in common, the more similar they are
- Resnik: measure the common information as:
  - The information content of the lowest common subsumer of the two nodes
  - $\mathrm{sim}_{\mathrm{resnik}}(c_1, c_2) = -\log P(\mathrm{LCS}(c_1, c_2))$

Resnik example

$sim_{resnik}(hill, coast) = ?$



Dekang Lin method

- Similarity between A and B needs to do more than measure common information
- The more **differences** between A and B, the less similar they are:
  - Commonality: the more info A and B have in common, the more similar they are
  - Difference: the more differences between the info in A and B, the less similar
- Commonality: IC(common(A, B))
- Difference: IC(description(A, B)) – IC(common(A, B))

Dekang Lin method

- Similarity theorem: The similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are
- $$\mathrm{sim}_{Lin}(A, B) = \frac{\log P(\mathrm{common}(A, B))}{\log P(\mathrm{description}(A, B))}$$
- Lin furthermore shows (modifying Resnik) that info in common is twice the info content of the LCS

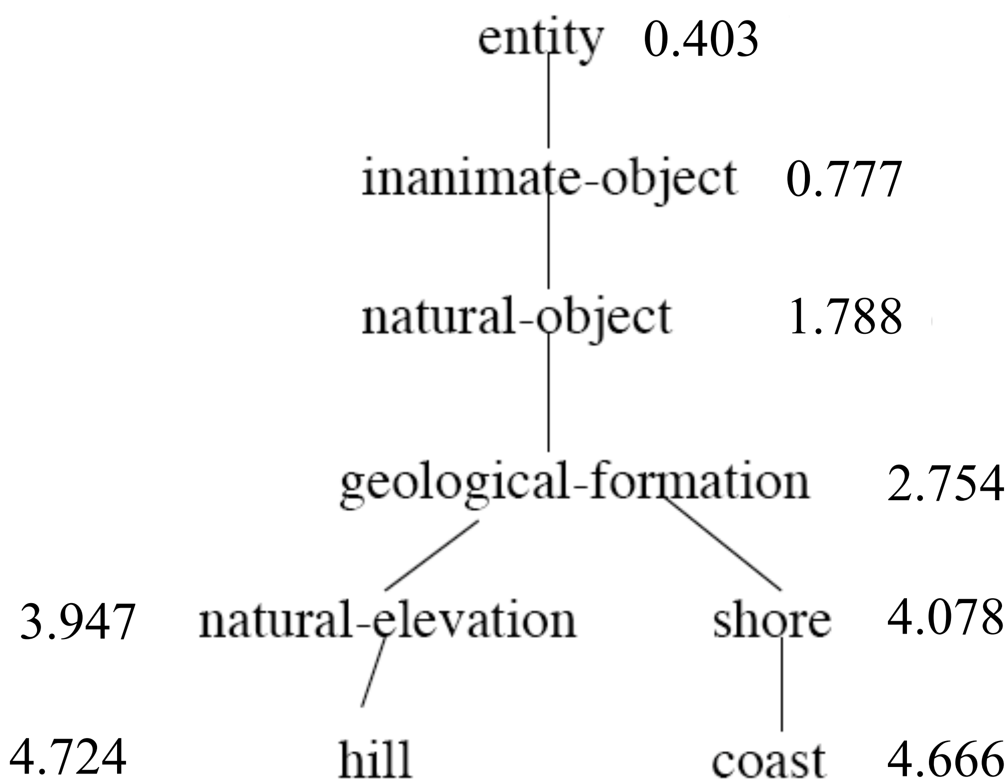
Lin similarity function

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

Or: the information content of LCS(c_1, c_2), *normalized* (divided) by the *average* information content of c_1 and c_2

Lin example

$sim_{Lin}(hill, coast) = ?$



Jiang-Conrath distance

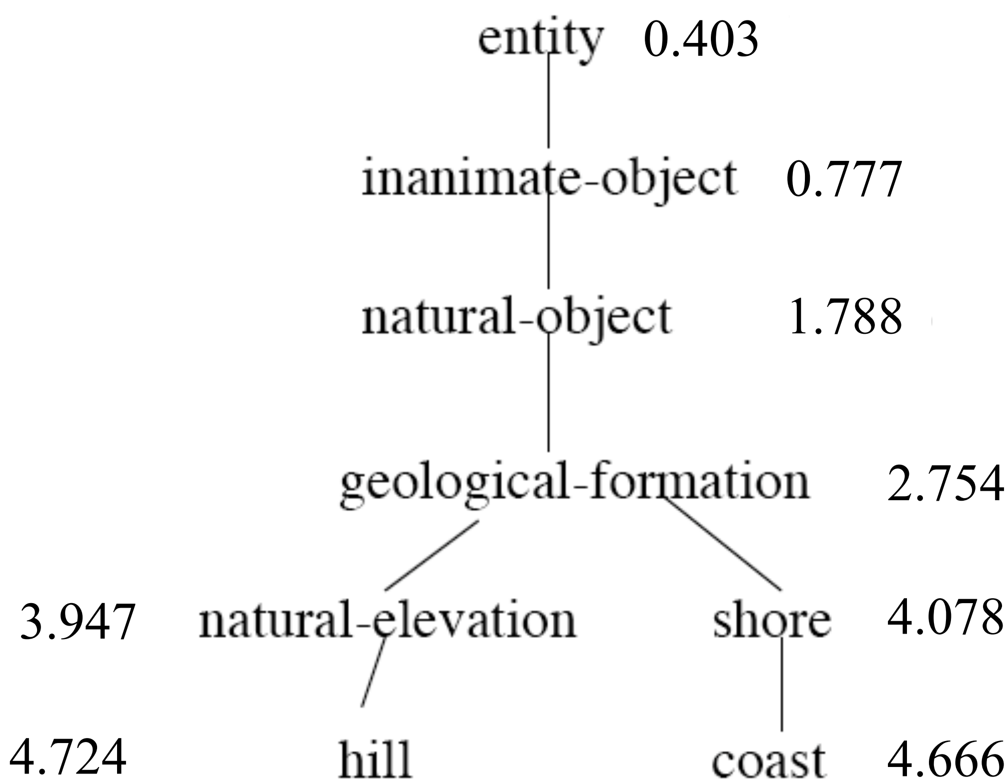
The Jiang-Conrath approach uses information content to assign lengths to graph edges

$$\text{dist}_{JC}(c, \text{hypernym}(c)) = \text{IC}(c) - \text{IC}(\text{hypernym}(c))$$

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

Jiang-Conrath example

$sim_{JC}(hill, coast) = ?$



More examples

Let's examine how the various measures compute the similarity between `gun` and a selection of other words:

w2	IC(w2)	lso	IC(lso)	Resnik	Lin	JiangC
gun	10.9828	gun	10.9828	10.9828	1.0000	0.0000
weapon	8.6121	weapon	8.6121	8.6121	0.8790	2.3708
animal	5.8775	object	1.2161	1.2161	0.1443	14.4281
cat	12.5305	object	1.2161	1.2161	0.1034	21.0812
water	11.2821	entity	0.9447	0.9447	0.0849	20.3756
evaporation	13.2252	[ROOT]	0.0000	0.0000	0.0000	24.2081

IC(w2): information content (negative log prob) of (the first synset for) word w2
lso: least superordinate (most specific hypernym) for "gun" and word w2.
IC(lso): information content for the lso.

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* for $1 + 4 = 5$

Recap: thesaurus-based similarity

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

Problems with thesaurus-based methods

- We don't have a thesaurus for every language
- Even if we do, many words are missing
  - Neologisms: *retweet*, *iPad*, *blog*, *unfriend*, …
  - Jargon: *poset*, *LIBOR*, *hypervisor*, …
- They rely on hyponym hierarchy
  - Strong for nouns
  - But lacking for adjectives and even verbs
- Alternative: distributional methods

Distributional methods

- Firth (1957)
  "You shall know a word by the company it keeps!"
- Example from Nida (1975) noted by Lin:

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

- Intuition:
  - Just from these contexts, a human could guess meaning of *tezgüino*
  - So we should look at the surrounding contexts, see what other words have similar context

Fill-in-the-blank on Google

You can get a quick & dirty impression of what words show up in a given context by putting a * in your Google query:

“drank a bottle of *”

Hi I'm Noreen and I once drank a bottle of **wine** in under 4 minutes
SHE DRANK A BOTTLE OF **JACK**?! harleyabshireblondie.
he drank a bottle of **beer** like any man
I topped off some salted peanuts and drank a bottle of **water**
The partygoers drank a bottle of **champagne**.
MR WEST IS DEAD AS A HAMMER HE DRANK A BOTTLE OF **ROGAINE**
aug 29th 2010 i drank a bottle of **Odwalla Pomegranate Juice** and got **...**
The 3 of us drank a bottle of **Naga Viper Sauce ...**
We drank a bottle of **Lemelson pinot noir** from Oregon ($52)
she drank a bottle of **bleach** nearly killing herself, "to clean herself from her wedding"

Context vector

- Consider a target word *w*
- Suppose we had one binary feature f_i for each of the *N* words in the lexicon v_i
- Which means "word v_i occurs in the neighborhood of *w*"
- $w = (f_1, f_2, f_3, \ldots, f_N)$
- If w = *tezgüino*, v_1 = *bottle*, v_2 = *drunk*, v_3 = *matrix*:
- $w = (1, 1, 0, \ldots)$

Intuition

- Define two words by these sparse feature vectors
- Apply a vector distance metric
- Call two words similar if their vectors are similar

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	0	0	1	1	1	0	1	0
information	0	0	1	1	1	0	1	0

Distributional similarity

So we just need to specify 3 things:

1. How the co-occurrence terms are defined
2. How terms are weighted
   - (Boolean? Frequency? Logs? Mutual information?)
3. What vector similarity metric should we use?
   - Euclidean distance? Cosine? Jaccard? Dice?

1. Defining co-occurrence vectors

- We could have windows of neighboring words
  - Bag-of-words
  - We generally remove **stopwords**
- But the vectors are still very sparse
- So instead of using ALL the words in the neighborhood
- Let's just use the words occurring in particular grammatical relations

Defining co-occurrence vectors

“The meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entitites relative to other entities.”
Zellig Harris (1968)

Idea: parse the sentence, extract grammatical dependencies

I discovered dried tangerines:

discover (subject I)	I (subj-of discover)
tangerine (obj-of discover)	tangerine (adj-mod dried)
dried (adj-mod-of tangerine)	

Co-occurrence vectors based on grammatical dependencies

For the word *cell*: vector of $N \times R$ features

(*R* is the number of dependency relations)

	subj-of, absorb	subj-of, adapt	subj-of, behave	...	pobj-of, inside	pobj-of, into	...	nmod-of, abnormality	nmod-of, anemia	nmod-of, architecture	...	obj-of, attack	obj-of, call	obj-of, come from	obj-of, decorate	...	nmod, bacteria	nmod, body	nmod, bone marrow
cell	1	1	1		16	30		3	8	1		6	11	3	2		3	2	2

2. Weighting the counts

("Measures of association with context")

- We have been using the frequency count of some feature as its weight or value
- But we could use any function of this frequency
- Let's consider one feature
- f = (r, w') = (obj-of, *attack*)
- $P(f|w) = \text{count}(f, w) / \text{count}(w)$

- $\text{Assoc}_{\text{prob}}(w, f) = p(f|w)$

Intuition: why not frequency

Objects of the verb *drink*:

Object	Count	PMI assoc	Object	Count	PMI assoc
bunch beer	2	12.34	wine	2	9.34
tea	2	11.75	water	7	7.65
Pepsi	2	11.75	anything	3	5.15
champagne	4	11.75	much	3	5.15
liquid	2	10.53	it	3	1.25
beer	5	10.20	<SOME AMOUNT>	2	1.22

- "drink it" is more common than "drink wine"
- But "wine" is a better "drinkable" thing than "it"
- We need to control for *expected* frequency
- We do this by normalizing by the expected frequency we would get assuming independence

Weighting: Mutual Information

- **Mutual information** between random variables X and Y

$$I(X,Y) = \sum_x \sum_y P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- **Pointwise mutual information**: measure of how often two events x and y occur, compared with what we would expect if they were independent:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Weighting: Mutual Information

- **Pointwise mutual information**: measure of how often two events x and y occur, compared with what we would expect if they were independent:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- PMI between a target word *w* and a feature *f* :

$$\text{assoc}_{\text{PMI}}(w,f) = \log_2 \frac{P(w,f)}{P(w)P(f)}$$

Mutual information intuition

Objects of the verb *drink*

Object	Count	PMI assoc	Object	Count	PMI assoc
bunch beer	2	12.34	wine	2	9.34
tea	2	11.75	water	7	7.65
Pepsi	2	11.75	anything	3	5.15
champagne	4	11.75	much	3	5.15
liquid	2	10.53	it	3	1.25
beer	5	10.20	<SOME AMOUNT>	2	1.22

Lin is a variant on PMI

- PMI between a target word *w* and a feature *f* :

$$\text{assoc}_{\text{PMI}}(w,f) = \log_2 \frac{P(w,f)}{P(w)P(f)}$$

- Lin measure: breaks down expected value for P(*f*) differently:

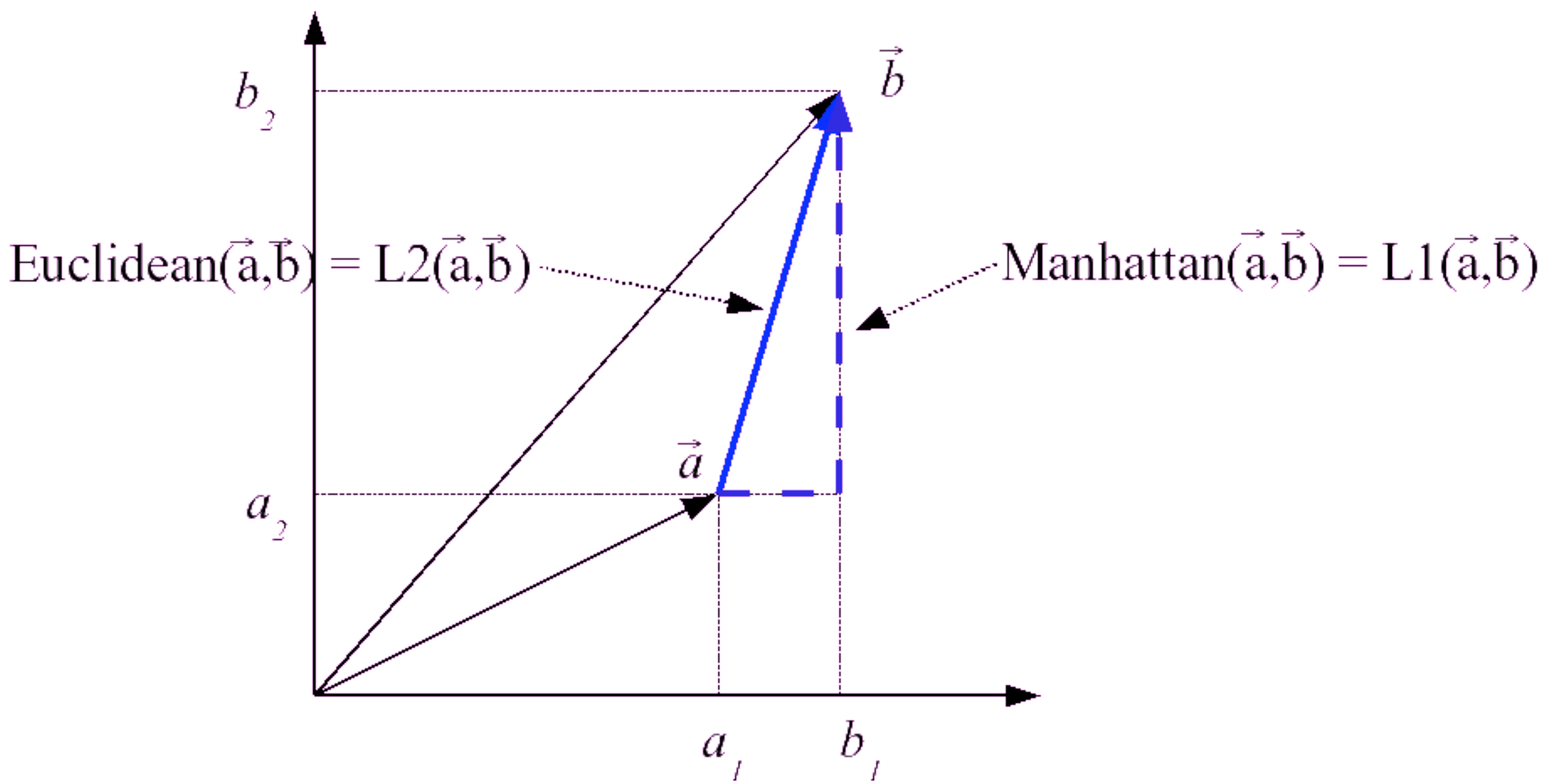
$$\text{assoc}_{\text{Lin}}(w,f) = \log_2 \frac{P(w,f)}{P(w)P(r|w)P(w'|w)}$$

Summary: weightings

- See Manning and Schuetze (1999) for more

$$
\begin{aligned}
\text{assoc}_{\text{prob}}(w, f) &= P(f|w) \\
\text{assoc}_{\text{PMI}}(w, f) &= \log_2 \frac{P(w,f)}{P(w)P(f)} \\
\text{assoc}_{\text{Lin}}(w, f) &= \log_2 \frac{P(w,f)}{P(w)P(r|w)P(w'|w)} \\
\text{assoc}_{\text{t-test}}(w, f) &= \frac{P(w,f) - P(w)P(f)}{\sqrt{P(f)P(w)}}
\end{aligned}
$$

3. Defining vector similarity



Summary of similarity measures

$$\mathrm{sim}_{\mathrm{cosine}}(\vec{v},\vec{w}) = \frac{\vec{v}\cdot\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

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

$$\mathrm{sim}_{\mathrm{Dice}}(\vec{v},\vec{w}) = \frac{2\times\sum_{i=1}^{N} \min(v_i,w_i)}{\sum_{i=1}^{N} (v_i+w_i)}$$

$$\mathrm{sim}_{\mathrm{JS}}(\vec{v}||\vec{w}) = D(\vec{v}|\frac{\vec{v}+\vec{w}}{2}) + D(\vec{w}|\frac{\vec{v}+\vec{w}}{2})$$

Evaluating similarity measures

- Intrinsic evaluation
  - Correlation with word similarity ratings from humans
- Extrinsic (task-based, end-to-end) evaluation
  - Malapropism (spelling error) detection
  - WSD
  - Essay grading
  - Plagiarism detection
  - Taking TOEFL multiple-choice vocabulary tests
  - Language modeling in some application

An example of detected plagiarism

MAINFRAMES

Mainframes are primarily referred to large computers with rapid, advanced processing capabilities that can execute and perform tasks equivalent to many Personal Computers (PCs) machines networked together. It is characterized with high quantity Random Access Memory (RAM), very large secondary storage devices, and high-speed processors to cater for the needs of the computers under its service.

Consisting of advanced components, mainframes have the capability of running multiple large applications required by many and most enterprises and organizations. This is one of its advantages. Mainframes are also suitable to cater for those applications (programs) or files that are of very high demand by its users (clients). Examples of such organizations and enterprises using mainframes are online shopping websites such as Ebay, Amazon, and computing-giant

MAINFRAMES

Mainframes usually are referred those computers with fast, advanced processing capabilities that could perform by itself tasks that may require a lot of Personal Computers (PC) Machines. Usually mainframes would have lots of RAMs, very large secondary storage devices, and very fast processors to cater for the needs of those computers under its service.

Due to the advanced components mainframes have, these computers have the capability of running multiple large applications required by most enterprises, which is one of its advantage. Mainframes are also suitable to cater for those applications or files that are of very large demand by its users (clients). Examples of these include the large online shopping websites -i.e. : Ebay, Amazon, Microsoft, etc.

What to do for the data assignments

- Some things people did last year on the WordNet assignment
- Notice interesting inconsistencies or incompleteness in Wordnet
  - There is no link in the WordNet synset between "kitten" or "kitty" and "cat".
    - But the entry for "puppy" lists "dog" as a direct hypernym but does not list "young mammal" as one.
  - "Sister term" relation is nontransitive and nonsymmetric
  - "entailment" relation incomplete; "Snore" entails "sleep," but "die"doesn't entail "live."
  - antonymy is not a reflexive relation in WordNet
- Notice potential problems in wordnet
  - Lots of rare senses
  - Lots of senses are very very similar, hard to distinguish
  - Lack of rich detail about each entry (focus only on rich relational info)

- Notice interesting things
  - It appears that WordNet verbs do not follow as strict a hierarchy as the nouns.
  - What percentage of words have one sense?

POS	Monosemous Words and Senses	Polysemous Words	Polysemous Senses
Noun	101321	15776	43783
Verb	6261	5227	18629
Adjective	16889	5252	14413
Adverb	3850	751	1870
Totals	128321	27006	78695